RailNorrköping 2019

8th International Conference on Railway Operations Modelling and Analysis (ICROMA)

Norrköping, Sweden

June 17th – 20th, 2019

EDITORS

Anders Peterson, Linköping University
Martin Joborn, Linköping University and RISE Research Institutes of Sweden
Markus Bohlin, KTH Royal Institute of Technology and RISE Research Institutes of Sweden

PRODUCTION AND LAYOUT

Jennifer Warg, KTH Royal Institute of Technology
Nils Breyer, Linköping University

PHOTOS

Thor Balkhed, Linköping University
Marcel Köppe (photo on front page)

Linköping Electronic Conference Proceedings No. 69
ISSN: 1650-3686, eISSN: 1650-3740
ISBN: 978-91-7929-992-7


Norrköping C


050
61208
61208

PREFACE

RailNorrköping 2019, the 8th International Conference on Railway Operations Modelling and Analysis (ICROMA) was organized by the International Association of Railway Operations Research (IAROR) in co-operation with Linköping University and held at Linköping University in Norrköping, Sweden, on June 17th – 20th, 2019.

Rail researchers and industry from all over the world came together for an international transfer of knowledge and co-operation in line with IAROR´s vision. In total, 122 papers were presented. The review of abstracts and full papers was supported by 91 skilled researchers, including the members of the IAROR board and the conference scientific advisory committee. The submissions are categorized into scientific and professional papers. The conference program also included three keynote addresses, a short course on different related topics given by invited speakers as well as a poster session. Rail researchers and industry from all over the world came together for an international transfer of knowledge and co-operation in line with IAROR´s vision.

The conference was organized by Norio Tomii, Nihon University (President of IAROR), Ingo A. Hansen, Delft University of Technology (Vice-President of IAROR), Anders Peterson, Linköping University (Conference Chair), Markus Bohlin, KTH Royal Institute of Technology and RISE Research Institutes of Sweden (Program Chair), Martin Joborn, Linköping University and RISE Research Institutes of Sweden (Conference organization Chair), Emma Solinen, Swedish Transport Administration and Linköping University (Industrial Representant) and Johan Högdahl, KTH Royal Institute of Technology (Program co-chair) supported by Nils Breyer, Carl-Henrik Häll, Tomas Lidén, Christiane Schmidt and Martin Waldemarsson, Linköping University as well as Jennifer Warg, KTH Royal Institute of Technology. We are also gratefully to all sponsors and additional supporters who helped in making RailNorrköping 2019 possible.

The present publication is a selection of the proceedings that comprises all papers, where the respective authors agreed on electronic publishing.

Norrköping, September 2019
The editors

Contents

ORGANIZATION COMMITTEE

Norio Tomii, Nihon University (President of IAROR)
Ingo A. Hansen, Delft University of Technology (Vice-President of IAROR)
Anders Peterson, Linköping University (Conference Chair)
Markus Bohlin, KTH Royal Institute of Technology and RISE Research Institutes of Sweden (Program Chair)
Martin Joborn, Linköping University and RISE Research Institutes of Sweden (Conference organization Chair)
Emma Solinen, Swedish Transport Administration (Industrial Representative) and Linköping University
Johan Högdahl, Royal Institute of Technology (Program co-chair)

SCIENTIFIC ADVISORY COMMITTEE

Mishra Abhyuday, Indian Institute of Technology Kharagpur
Hans Boysen, KTH Royal Institute of Technology and Swedish Transport Administration
Tyler Dick, University of Illinois at Urbana-Champaign
Rob Goverde, Delft University of Technology
Alex Landex, Ramboll
Carlo Mannino, SINTEF ICT Oslo
Lingyun Meng, Beijing Jiaotong University
Lei Nie, Beijing Jiaotong University
Nils Nießen, RWTH Aachen University
Andreas Oetting, Technische Universität Darmstadt
Dario Pacciarelli, University Roma Tre
John Preston, University of Southampton
Stefano Ricci, University of Rome La Sapienza
Joaquin Rodriguez, IFSTTAR
Thomas Schlechte, LBW Optimization GmbH
Alex Wardrop, Independent Railway Operations Research Consultant, Sydney, Australia
Thomas White, VTD Rail Consulting

LOCAL ORGANIZATION COMMITTEE

Nils Breyer, Linköping University
Charlotte Eriksson, Linköping University
Sara Gestrelius, RISE Research Institutes of Sweden and Linköping University
Carl-Henrik Häll, Linköping University
Olivia Jansson, Linköping University
Martin Joborn, Linköping University and RISE Research Institutes of Sweden
Tomas Lidén, VTI, the Swedish National Road and Transport Research Institute and Linköping University
Alice Lindberg, Linköping University
Gustav Nåfält, Linköping University
Anders Peterson, Linköping University
Christiane Schmidt, Linköping University
Emma Solinen, Swedish Transport Administration and Linköping University

Martin Waldemarsson, Linköping University
Jennifer Warg, KTH Royal Institute of Technology

REVIEWERS

Mishra Abhyuday, Indian Institute of Technology Kharagpur
Abderrahman Ait-Ali, VTI, the Swedish National Road and Transport Research Institute and Linköping University
Thomas Albrecht, DXC Technology
Lukas Bach, SINTEF
William Barbour, Vanderbilt University
Ivan Belosevic, University of Belgrade
Nikola Besinovic, Delft University of Technology
Markus Bohlin, KTH Royal Institute of Technology and RISE Research Institutes of Sweden
Ralf Borndoerfer, Zuse-Institute Berlin (ZIB)
Hans Boysen, KTH Royal Institute of Technology and Swedish Transport Administration
Emanuel Broman, VTI, the Swedish National Road and Transport Research Institute and Linköping University
Thorsten Büker, VIA Consulting & Development GmbH
Valentina Cacchiani, University of Bologna
Francesco Corman, ETH Zurich
Nicola Coviello, Politecnico di Torino
Andrea D Ariano, Università degli Studi Roma Tre
Tyler Dick, University of Illinois at Urbana-Champaign
Jawad Elomari, RISE Research Institutes of Sweden
Frank Fischer, Johannes Gutenberg Universität Mainz
Huiling Fu, Beijing Jiaotong University
Taku Fujiyama, University College London
Sara Gestrelius, RISE Research Institutes of Sweden and Linköping University
Yuan Gao, Beijing Jiaotong University
Nima Ghaviha, RISE Research Institutes of Sweden
Rob Goverde Delft University of Technology
Ingo A. Hansen, Delft University of Technology
Dennis Huisman, Erasmus University Rotterdam
Carl Henrik Häll, Linköping University
Johan Högdahl, KTH Royal Institute of Technology
Erik Jenelius, KTH Royal Institute of Technology
Martin Joborn, Linköping University and RISE Research Institutes of Sweden
Pär Johansson, Swedish Transport Administration
Pavle Kecman, Allianz Data Office
Fahimeh Khoshniyat, Linköping University
Ida Kristoffersson, VTI, the Swedish National Road and Transport Research Institute
Alexander Kuckelberg, VIA Consulting & Development GmbH
Yung-Cheng Lai, National Taiwan University
Alex Landex, Ramboll
Tomas Lidén, VTI, the Swedish National Road and Transport Research Institute and Linköping University

Christian Liebchen, Technische Hochschule Wildau
Therese Lindberg, VTI, the Swedish National Road and Transport Research Institute
Per Olov Lindberg, KTH Royal Institute of Technology and VTI, the Swedish National Road and Transport Research Institute
Olov Lindfeldt, MTR Nordic
Sihui Long, Beijing Jiaotong University
Xiaojie Luan, Delft University of Technology
Rémi Lucas, ENSTA ParisTech
Fredrik Lundström, Swedish Transport Administration
Carlo Mannino, SINTEF ICT Oslo
Gabor Maroti, Vrije Universiteit Amsterdam
Ullrich Martin, University of Stuttgart
Giorgio Medeossi, TRENOlab
Lingyun Meng, Beijing Jiaotong University
Matúš Mihalák, Maastricht University
Niloofar Minbashi, KTH Royal Institute of Technology
Gemma Nicholson, University of Birmingham
Nils Nießen, RWTH Aachen Universitiy
Andreas Oetting, Technische Universität Darmstadt
Nils Olsson, Norwegian University of Science and Technology
Yanfeng Ouyang, University of Illinois Urbana-Champaign
Dario Pacciarelli, University Roma Tre
Jörn Pachl, Technische Universität Braunschweig
Carl-William Palmqvist, Lund University
Paola Pellegrini, IFSTTAR
Anders Peterson, Linköping University
Pasqualina Potena, RISE Research Institutes of Sweden AB
John Preston, University of Southampton
Stefano Ricci, University of Rome La Sapienza
Joaquin Rodriguez, IFSTTAR
Clas Rydergren, Linköping University
Mahnam Saeednia, HaCon
Keisuke Sato, Kanagawa University
Stanley Schade, Zuse Institute Berlin
Thomas Schlechte, LBW Optimization GmbH
Christiane Schmidt, Linköping University
Marie Schmidt, Erasmus University Rotterdam
Tilo Schumann, Berlin Senate for Environment, Transport and Climate Protection
Hans Sipilä, Sweco
Emma Solinen, Swedish Transport Administration and Linköping University
Birgitta Thorslund, VTI, the Swedish National Road and Transport Research Institute
Ambra Toletti, SBB CFF FFS
Norio Tomii, Nihon University
Pengling Wang, Delft University of Technology
Pieter Vansteenwegen, KU Leuven Mobility Research Centre
Alex Wardrop, Independent Railway Operations Research Consultant, Sydney, Australia
Jennifer Warg, KTH Royal Institute of Technology

Norman Weik, RWTH Aachen University
Thomas White, VTD Rail Consulting
Jing Xun, Beijing Jiaotong University
Jiateng Yin, Beijing Jiaotong University
Wuyang Yuan, Beijing Jiaotong University
Jun Zhang, The Key Laboratory of Road and Traffic Engineering Tongji University
Jinchuan Zhang, Beijing Jiaotong University
Jun Zhao, Southwest Jiaotong University

KEYNOTE SPEAKERS

Gunnar Alexandersson, Senior researcher (Stockholm School of Economics, Sweden) and senior adviser, regulations and international affairs (SJ AB, Sweden). *Railway Market Opening and Organisational Reforms in Sweden*
Rob Goverde, Professor of Railway Traffic Management & Operations (Department of Transport & Planning, TU Delft, The Netherlands). *Railway operations research and the development of digital railway traffic systems*
Jonas Eliasson, Professor Transport systems (Linköping University, Sweden). *What is the social value of railway capacity? Connecting transport economics and railway operations research*

SPONSORS

A Study on Train Travel Time Simulation focused on detailed Dwell Time Structure and On-Site Inspections

Shigeaki Adachi [a,1], Masahito Koresawa [b], Giancarlos Troncoso Parady [a], Kiyoshi Takami [a], Noboru Harata [a]
[a] Department of Urban Engineering, The University of Tokyo
Engineering building 14, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656 Japan
[1] E-mail: adachi@ut.t.u-tokyo.ac.jp, Phone: +81 (3) 3837-7318
[b] Department of Operation, Tokyo Metro Co., Ltd.
3-19-6 Higashi-ueno, Taito-ku, Tokyo 110-8614 Japan

Abstract
In order to reduce passenger congestion during morning rush hour, railway companies in the Tokyo metropolitan area have increased the number of trains. On the other hand, once a train exceeds a dwell time due to sudden events such as passengers rushing onto a train, passengers agglomerating in specific cars and doors, objects getting caught in doors etc., delays propagate to subsequent trains quickly. To evaluate daily train transport stability and countermeasures against train delays, a train travel time simulation model is needed. However, it has been difficult so far to replicate the occurrence of sudden events and the fluctuations in passenger demand. In this paper, we use detailed data based on dwell time structure and on-site inspections to construct a train travel time simulator. In addition, we evaluate several case-studies of timetable adjustments and passenger demand variations.

Keywords
Train delay, Train travel time simulation, Ticket gate ingress and egress record data, Smirnoff-Grubbs test, Dwell time

1 Introduction

Railway companies in Tokyo metropolitan area of Japan have increased the number of trains to alleviate passenger congestion and improve train delays during morning rush hour. However, train headways are limited by the capacity of the signalling system. Under such circumstances, train delays propagate to subsequent trains because of short headways. Furthermore, during boarding and alighting, when small sudden events such as passengers rushing onto a train, passengers agglomerating in specific cars and doors, or objects getting caught in doors occur, dwell times are extended.

Train travel time simulation models have been constructed so far. Railway simulation using traffic record data has been studied by Carey, M. (1999), Hürlimann, D. (2004), Van der Meer, D. (2010), Graffagnino, T (2012). Furthermore, Hansen, I. et al (2014). have studied various kinds of train simulators focusing on railway system functions. Janecek, D. (2010) studied simulations focusing on changes in the infrastructure and timetable. Ushida, K. et al. (2011) developed a chromatic diagram visualized reflecting train delays as colours. In terms of train delay measures, Yamamura, A. (2013 & 2014) and Adachi, S. (2016) have studied various kind of measures against train delays on one of the most congested lines in Tokyo and evaluate those of effects on operation stability.

However, these studies have mainly focused on railway system, simulator functions and train delay measures. . So far, it has been difficult to consider daily passenger demand and the effect of small sudden events which occur frequently. Thus, consideration of these detailed elements is important to improve daily train operations. We have focused on composition of dwell time, and the relationship between passenger demand and dwell time including such sudden events that has not been well studied so far.

In this paper, we construct a detailed train travel time simulation focusing on the Tokyo Metro Tozai Line, which is one of the most congested lines in Tokyo.

2 Train diagram composition

Train head way is constructed by dwell time and minimum headway and buffer time. In a dense timetable such as lines running in the city center, buffer times are set at almost minimal, therefore once dwell time extends, buffer time becomes negative. This means that train delays propagate to subsequent trains.

Dwell time is segmented into 4 parts: passenger alighting time (A), passenger boarding time (B), door closing confirmation time (C), and safety confirmation time (S). In terms of door closing confirmation time (C), station staff judge timing of door closing at the end of passenger boarding. After passenger board, the staff give a signal to close doors to the conductor, and the conductor close the doors. After door close, station staff confirm the safety along cars and give a signal for departure to the conductor. This operation time is defined as safety confirmation time (D). The most time-consuming door to alight and board affects sum of passenger alighting time (A) and passenger boarding time (B).

Furthermore, it takes 2 seconds for doors to open after arriving at a station. According to these definitions, dwell time at station i of train j is defined as (1). All times are given in seconds.

$$D_{i,j} = 2(sec) + \max_{k,l}(A_{i,j,k,l} + B_{i,j,k,l}) + C_{i,j} + S_{i,j} \quad (1)$$

$A_{i,j,k,l}$: Alighting time at station i of train j, car No. k, door No. l
$B_{i,j,k,l}$: Boarding time at station i of train j, car No. k, door No. l
$C_{i,j}$: Closing confirmation time at station i of train j
$D_{i,j}$: Dwell time at station i of train j
$S_{i,j}$: Safety confirmation time at station i of train j

3 Factors influencing each time to construct dwell time

To build a detailed train travel time simulation, it is necessary to know what kind of factors influence each time to construct dwell time. Factor affecting composed time are illustrated in Figure 1. Alighting and boarding times are influenced by the number of passengers and by passenger congestion degree in a car. In Tozai line, some trains have wider door than usual cars. This width also affects alighting and boarding times.

In terms of door closing operations, when staff judge the timing in some station, multiple station staff members cooperate due to curved nature of some platform and depending on the congestion levels in the platform. Door closing confirmation time fluctuates depending on these characteristics.

After door close, staff confirm safety along cars in the same way as during door closing confirmation operations. Safety confirmation time also fluctuates depending on these

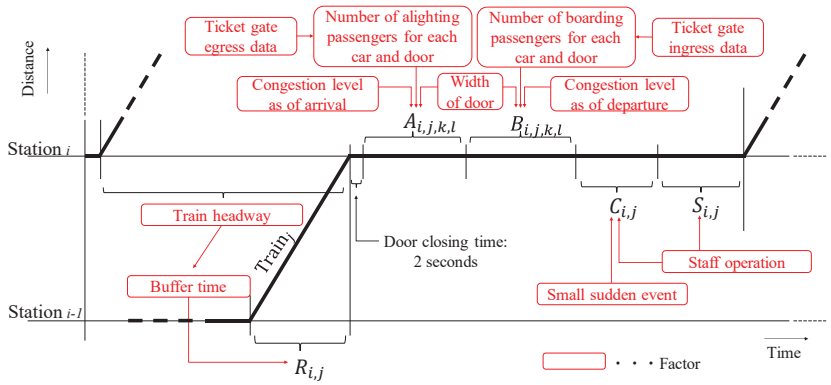


Figure 1: Factor related chart on train diagram

characteristics.

We estimated each time model in dwell time considering the abovementioned causes.

4 Train Travel Time Simulation Outline

The outline of the simulation is illustrated in Figure 2. In the initial condition, the simulation starts with 78 trains running on the Tozai line in direction of the city center between 6:30 to 10:00 distributed along stations.

First, departure times and number of passengers for each car at each starting station are input. Departure time data is acquired from train traffic record data which is obtained from electric circuit on a track. at each station. Passenger number data is acquired from a five-day on-site inspection conducted on November 2015. Then, each time that makes up dwell time is estimated for each train.

In terms of the number of passengers alighting and boarding, ticket gate ingress and egress count record data aggregated in 30-minute intervals is utilized. Using these data, the number of alighting passengers is allocated to each train and car based on passenger congestion degree. The number of boarding passengers is allocated based on train headways. Furthermore, the calculated number of passengers is allocated to each door based on rate of door utilization observed during the on-site inspections. We model alighting and boarding times using linear regression analysis.

Door closing confirmation time and safety confirmation time are estimated based on on-site inspection results. Especially during door closing confirmation time, there are some small sudden events such as passengers rushing onto a train, passengers agglomerating in specific cars and doors, objects getting caught in doors etc. These events must be considered to build a more detailed simulation. In this study, these events are applied by Smirnoff-Grabs test.

Running time is calculated depending on whether the buffer time is negative or positive. Minimum headways are determined by the signalling system, so excess of planned running times are influenced by the negative buffer time at each station.

4.1 Estimation of alighting and boarding times

To estimate alighting and boarding times, the number of passengers should be calculated. Ticket gate egress and ingress data is utilized to estimate them. The cumulative distribution

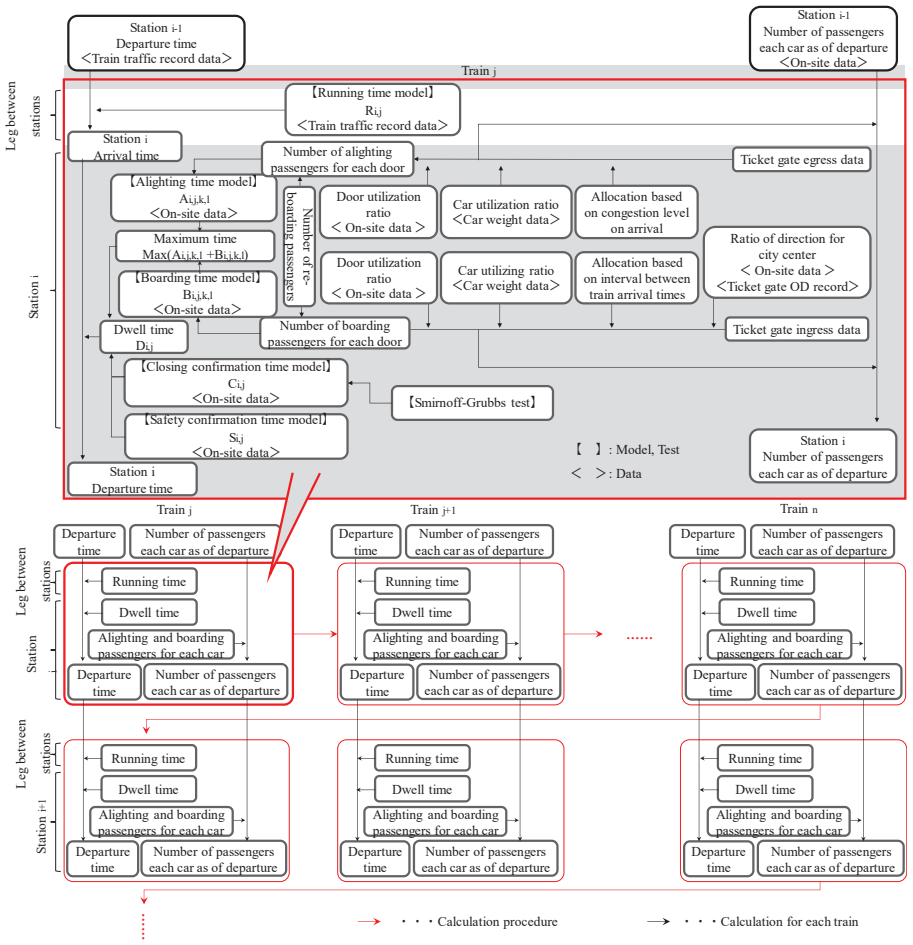


Figure 2: Simulation flow

is approximated by Gompertz curve (Figure 3), and the number of egress and ingress passengers in second-scale are derived. In a precise sense, time differences between ticket gate and train door should be considered. In this simulation, the time difference between ticket gate and the most time-consuming door to alight and board is considered, and the time difference is adjusted on the curve.

In terms of the ticket gate egress, the data has OD record for each 30-minute time interval, and boarding direction of egress passenger is observed. To distribute egress passengers to each train, the total number of egress passengers are calculated as following (2) to (4). Furthermore, train direction to the city center is defined as *A* and train direction to the suburbs is defined as *B*.

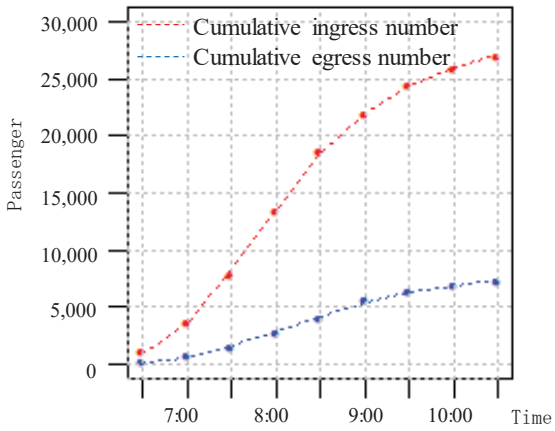


Figure 3: Example of cumulative distribution of ticket gate ingress and egress number

$$Eg_{d,i,t,A}=Eg_{d,i,t,A\,\&\,B} \cdot DRa_{d,i,t,A} \tag{2}$$

$$X_{d,i}(\mathrm{s}) = X'_{d,i}(\mathrm{s} + Mov_{i,A}) \tag{3}$$

$$NA_{d,i} = X_{d,i}(s_{d,i,78}) - X_{d,i}(s_{d,i,1}) \tag{4}$$

$Eg_{d,i,t}$: The number of egress passengers on day d at station i on time zone t
$DRa_{d,i,t,A}$: Rate of direction A on day d at station i on time zone t
$X'_{d,i}(\mathrm{s})$: Function of cumulative distribution approximated $Eg_{d,i,t,A}$
$X_{d,i}(\mathrm{s})$: Function of cumulative distribution adjusted the time difference on $X'_{d,i}(\mathrm{s})$
$s_{d,i,j}$: Arrival time on day d at station i of train j
$Mov_{i,A}$: Time difference between ticket gate and the most time-consuming door to alight and board at station i for direction A
$NA_{d,i}$: Total number of alighting passengers on day d for direction A

In general, the number of alighting passengers for each train is influenced by those of passenger congestion degree on arrival. Given that $NA_{i,t,A}$ is the number of alighting passengers at station i on time zone t for direction A, the equations are expressed as (5) and (6).

$$NA_{d,i,j} = NA_{d,i} \cdot (ArrCon_{d,i,j}/ArrCon_{d,i,J}) \tag{5}$$

$$ArrCon_{d,i,J} = \sum_{j \in J} ArrCon_{d,i,j} \tag{6}$$

$NA_{d,i,j}$: The number of alighting passengers on day d at station i on train j
$ArrCon_{d,i,j}$: Passenger congestion degree on arrival on day d at station i on train j
J: Set of train j

On the other hand, ticket gate ingress data doesn't have OD record. Thus, the number of boarding passengers and rate of direction *A* are calculated using on-site inspection data and $Eg_{d,i,t,A}$. The number of passengers at the time of departure for direction *A* is calculated that the number of passengers on arrival plus alighting passengers minus boarding passengers.

The equations are expressed as (7) and (8). To simulate on the day which is not inspection days, $DRb_{d,i,A}$ is adopted as average rate.

$$Ing_{d,i,t,A} = (\sum_{j \in J_{i,t}} DepCon_{d,i,j} - \sum_{j \in J_{i,t}} ArrCon_{d,i,j}) + Eg_{d,i,t,A} \quad (7)$$

$$DRb_{d,i,A} = \sum_{t} Ing_{d,i,t,A} / \sum_{j \in J_{i,t}} Ing_{d,i,t,A\&B} \quad (8)$$

$Ing_{d,i,t}$: The number of ingress passengers on day *d* at station *i* on time zone *t*
$DepCon_{d,i,j}$: The number of passengers at the time of departure on day *d* at station *i* of train *j* for direction *A*
$DRb_{d,i,A}$: Average rate of direction *A* on day *d* at station *i* on time zone *t*
$J_{i,t}$: Set of train *j* at station *i* on time zone *t*

Using DRb_{iA} and ticket gate ingress data, the number of boarding passengers each train is calculated as following (9) to (11). Since it is difficult to grasp how long it takes for passengers to get on the train during dwell time, then the number of boarding passengers each train is defined as the cumulative numbers between subsequent train's arrival time and following train's arrival time.

$$Ing_{d,i,t,A} = Ing_{d,i,t,A\,\&\,B} \cdot DRb_{iA} \quad (9)$$

$$Y_{d,i}(\mathrm{s}) = Y'_{d,i}(\mathrm{s} - Mov_{i,A}) \quad (10)$$

$$NB_{d,i,j} = Y_{d,i,j}(s_{d,i,j}) - Y_{d,i,j}(s_{d,i,j-1}) \quad (11)$$

$DRb_{i,A}$: Rate of direction *A* on day *d* at station *i* on time zone *t*
$Y'_{d,i}(\mathrm{s})$: Function of cumulative distribution approximated $Ing_{d,i,t,A}$
$Y_{d,i}(\mathrm{s})$: Function of cumulative distribution adjusted the time difference on $Y'_{d,i}(\mathrm{s})$
$NB_{d,i,j}$: The number of boarding passengers on day *d* at station *i* on train *j* for direction *A*

To distribute alighting and boarding passengers to each car and door, utilization rate of cars and doors must be estimated. Utilization rate of car each station is estimated from car weight data acquired between October 2015 and December 2015. And utilization rate of each door is grasped from the on-site inspection results. Both rates are implemented as fixed average value on the simulator.

4.2 Alighting time model

In terms of alighting time, two significant parameters are adopted, one is the number of alighting passengers and second is wider doors described earlier. To create the model, we

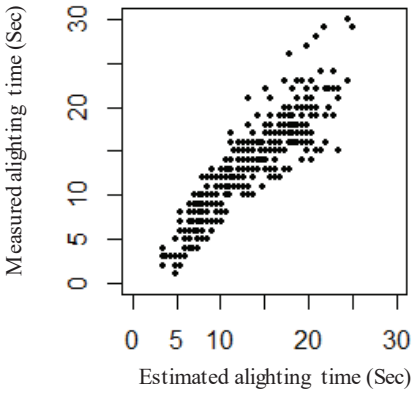


Figure 4: Relationship between measured value and estimated value in alighting time

utilize the video recording data which records passenger alighting and boarding on the platform at each station. Alighting number of passengers counting data which is each 391 samples is given by video data. Passenger congestion degree and wider door are determined from on-site inspections. In fact, passenger flow on platform affects dwell time. However, it is assumed that the model expresses the effects due to on-site inspection results including the flow.

In the alighting time regression model, explanatory variables are the number of alighting passengers and the presence or absence of wider door. The equation is expressed as (12). Figure 4 and Table 1 show the results.

$$A_{i,j,k,l} = \alpha + \beta_1 \cdot NA_{i,j,k,l} + \beta_2 \cdot Wide + \varepsilon \quad (12)$$

$A_{i,j,k,l}$: Alighting time at station i of train j, car No. k, door No. l
$NA_{i,j,k,l}$: The number of alighting passengers at station i of train j, car No. k, door No. l
Wide: Wider door dummy
α, β_1, β_2: Parameter
ε: Error term

Table 1: Result of alighting time model

Parameter	**Coefficient**	**t value**	**p value**
Intercept	4.89	19.97	1.56E-61
Number of alighting passengers	0.52	43.58	1.5E-151
Wider door dummy	-1.52	-6.19	1.5E-09
R^2: 0.83	Sample: 391 trains		

The result obtains good fit by $R^2$0.83, however there is variability between measured value and estimated value due to uncertain passenger flow. Therefore, the estimated value of alighting time is given by adding the normal random value of estimation error..

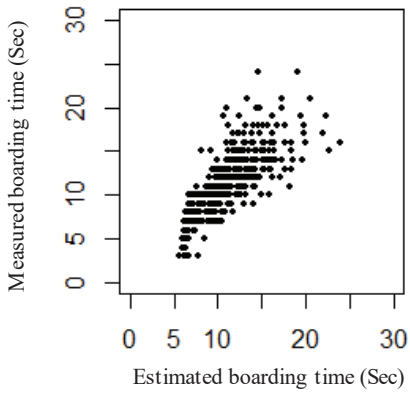


Figure 5: Relationship between measured value and estimated value in boarding time

4.3 Boarding time model

In terms of boarding time, two parameters are estimated, the number of boarding passengers and passenger congestion degree in the car as of departure. To create the model, we utilize the video recording data as is the case with alighting model. Boarding number of passengers counting data has also 391 samples.

In the boarding time model, since boarding time tends to extend due to congestion, and this distribution increase towards one side the dependent variable is log-transformed. Explanatory variables are the number of boarding passengers and passenger congestion degree in the car as of departure. The equation is expressed as (13).

$$logB_{i,j,k,l} = \alpha + \beta_1 \cdot NB_{i,j,k,l} + \beta_2 \cdot DepCon_{i,j,k} + \varepsilon \quad (13)$$

$B_{i,j,k,l}$: Boarding time at station i of train j, car No. k, door No. l
$NB_{i,j,k,l}$: The number of boarding passengers at station i of train j, car No. k, door No. l
$DepCon_{i,j,k}$: Passenger congestion degree at departure time at station i of train j, car No. k
α, β_1, β_2: Parameter
ε: Error term

Figure 5 and Table 2 show the estimation results. The result obtains good fit from $R^2$0.67, however there is variability between measured value and estimated value due to uncertainly passenger flow. Therefore, the estimated value of boarding time is given by adding the normal random value of estimation error.

Table 2: Result of boarding time model

Parameter	Coefficient	t value	p value
Intercept	0.63	17.56	3.37E-51
Number of boarding passengers	0.030	26.96	6.1E-91
Passenger congestion degree at departure	0.00051	2.36	0.019
R^2: 0.67 Sample: 391 trains			

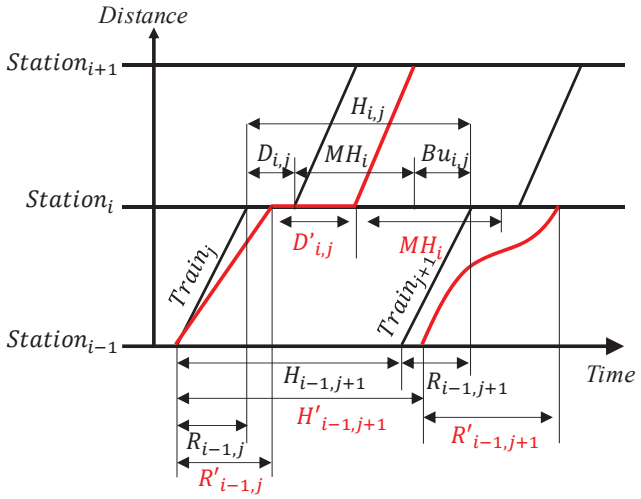


Figure 6: Mechanism of train delay propagation

4.4 Estimation of door closing confirmation time

Door closing confirmation time depends on station staff operations. To estimate door closing confirmation time, normal random numbers were simulated based on the distribution observed during the on-site inspections at each station. Moreover, detailed analysis of the time should consider small sudden events that happen frequently. The events are considered that a dwell time excess degree is discerned by Smirnoff-Grabs test based on long term dwell time records.

Regarding train j, the test statistics is defined as T_j, the logarithmic value of dwell time is defined as X_j, the average of logarithmic value of dwell time is defined as $\overline{X_j}$, the standard deviation is defined as s_j, the equation is expressed as (14). This judgement is focused on excess dwell time, so one sided-testing is adopted.

$$T_j = (X_j - \overline{X_j})/s_j \tag{14}$$

4.5 Estimation of safety confirmation time

Safety confirmation time also depends on station staff operations. As such, similar to door closing confirmation time. normal random numbers were simulated based on the distribution observed during the on-site inspections at each station.

4.6 Estimation of running time

To estimate running time, buffer time is considered. If the buffer is positive, the train would run following the planned running time. However, if the buffer time is negative, subsequent trains slow down or stop between stations because they are too close to the preceding train. The buffer time is determined by the signalling system design at each station. The phenomenon is illustrated in Figure 6, 7 and following (15) and (16). In figure 6, train headway (H) is segmented into 3 parts: dwell time (D), minimum headway which is determined by signaling system each station (MH), buffer time (Bu), running time (R). The red lines are expressed actual train behavior, and red letters with dash are actual time.

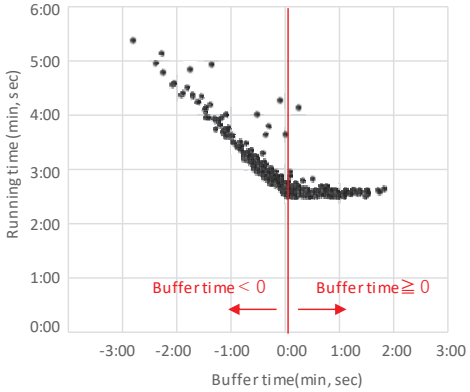


Figure 7: Relationship between Buffer time and Running time

$$Bu'_{i,j+1}=H'_{i-1,j+1}-(R'_{i-1,j}-R_{i-1,j})-D'_{i,j}-MH_{i,j+1} <0 \quad (15)$$

$$Bu'_{i,j+1}=H'_{i-1,j+1}-(R'_{i-1,j}-R_{i-1,j})-D'_{i,j}-MH_{i,j+1} \geq 0 \quad (16)$$

In terms of the relationship between buffer time and running time, with increasing negative buffer time, running time increases linearly (See Figure 7). Utilizing this linearity property, running time between stations is calculated.. When buffer time is positive, the train driver can adjust to recover lost time, but train driver operation is different with each driver, therefore, in the simulation, when buffer time is positive, trains run according to the planned running time.

4.7 Adjustment of train headway

In daily operations, if there is change in train headways, the control center operator adjusts the headways to prevent agglomerate of passenger congestion. If the train interval is longer than 1 minute 30 seconds and less than 2 minutes compared to the planned headway at the time of the departure, the preceding train is adjusted by a planned dwell time + 1 minute after the departure time. In the same way, the train interval is longer than 2 minutes and less than 2 minutes and 30 seconds, the adjustment time of preceding train is planned dwell time + 1 minute and 30 seconds.

In usual situations, the number of boarding passengers is calculated between arrival times. However, in the case of headway adjustment, the number of boarding passengers is calculated between arrival time of subsequent train and the time which subtract departure time of following train considered adjustment from the door closing confirmation time and the safety confirmation time.

5 Assessment of simulation reproducibility

To confirm that the simulation reproducibility and its accuracy is maintained, we put into the departure time and congestion data at starting station which is the 5 days data based on the construction of the simulation, then simulate 100 times for each day. Residual error RMS (Root mean square) is adopted as the performance index.

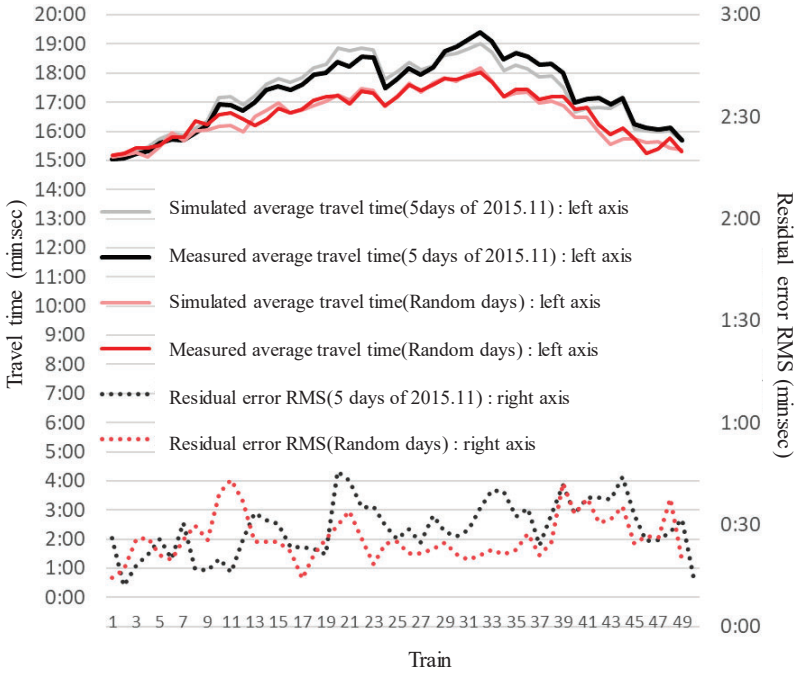


Figure 8: Simulation results of reproducibility

Further, we simulated 100 times for 10 days at random excluding the 5 days. There's no way to get some data on random days, we estimate them as follows.

Congestion degree of those random days at starting station is figured out based on proportion of the 5 days average degree to those of degree on random days.

The number of alighting passengers each station in random days is figure out based on equation (2) to (4). Ticket gate ingress and egress data replace the 5 days data with random days data, and rate of direction is adopted average rate of direction A on the 5 days. The number of boarding passengers each station in random days is figure out based on equation (9) to (11). Ticket gate ingress and egress data replace the 5 days data with random days data, and rate of direction is adopted average rate of direction A on the 5 days. In addition, wider door is set at random.

Figure 8 shows the results of the reproducibility test. The actual average of travel time is 17 minutes and 13 seconds and standard deviation is 1 minute and 22 seconds, and simulated that time is 17 minutes and 16 seconds and standard deviation is 1 minute and 31 seconds. High accuracy is maintained compared to references. Also, in the case of the data selected at random, those of simulated travel time is confirmed high accuracy that error between travel time and standard deviation are few seconds.

5.1 Case study for improvement of train delay

Railway companies have taken measures to improve train delay and train congestion. There are two types of measures, one is improvement of train timetable, second is distribution of passenger congestion. The former measure aims at avoiding delay propagation to subsequent trains. Important point to avoid propagation is to expand buffer times. This is

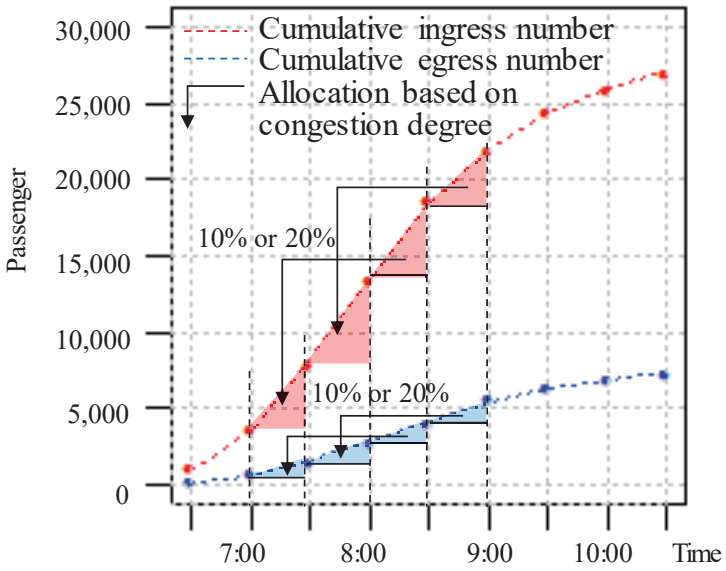


Figure 9: Allocation of passenger demand

also conducted by daily operation at control center.

The latter measure aims at distributing congestion agglomeration of specific cars and doors. Station staff encourage passengers to use more empty cars or use earlier trains. In 2017 summer, Tokyo metropolitan government implemented "Jisa Biz" staggered commuting campaign and many companies addressed changes in work start time during the campaign term. In 2020, the Tokyo Olympic and Paralympic Games will be held. Especially, congestion of peak-hour adding spectators would over the limit of train transportation capacity in Tokyo. The government would like to build staggered commuting as routine by 2020. Furthermore, for legacy, staggered commuting would be conductive to smooth transports and flexible lifestyles.

Utilizing the proposed simulation, we estimate the effect of staggered commuting on Tozai line focusing on one day. Passengers demand on starting station during 7:30 to 8:29 reduce 10%, and the 10% passengers are allocated to each train running on time zone 6:30 to 7:29 based on each train congestion degree. And boarding passengers during 8:00 to 8:29 and 8:30 to 8:59 reduce 10%, and the 10% passengers are allocated to each train running on time zone 7:00 to 7:29 and 7:30 to 7:59 based on each train passenger congestion degree. The number of alighting passengers is calculated as same way of boarding case. Furthermore, in the case of 20% reduce is calculated as same way (Figure 9).

Figure 10 shows the results. The actual average travel time is 16 minutes and 20 seconds, and passenger 10% moving case is 16 minutes and 14 seconds and that of 20% moving case is 16 minutes and 11 seconds. The average travel time is alleviated due to demand moving. Particularly, before peak hour, travel time increases by 17 seconds in the case of 10% moving case, and 28 seconds in the case of 20% moving case. On the other hand, on peak hour, the maximum improvement time is 24 seconds in the case of 10% moving case, and 41 seconds in the case of 20% moving case. The effects have decent improvement, but further demand moving deal is necessary for legacy.

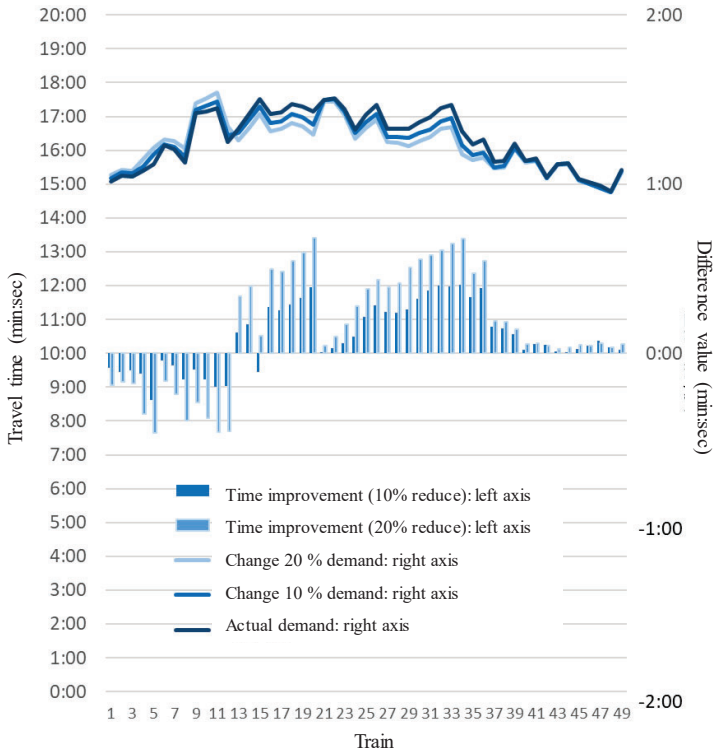


Figure 10: Simulation results of demand change

From this result, it is confirmed that travel time before peak hour increase temporarily, but travel time at peak hour improve well and average travel time is also shortened.

6 Conclusion

We have introduced an innovative method of train travel time simulation model utilizing daily ticket gate ingress and egress data and detailed on-site inspection results. Especially, focussing on each time model in dwell time is new characteristic of the simulation. Also, utilizing past traffic record data to model sudden small events during closing confirmation time is reproduced detailed situation. We obtained high reproducibility and confirm the usefulness of the proposed method. In the case of the staggered commuting campaign, we confirmed the effect of travel time change due to moving passenger demand. In this case, we confirmed certain level of peak hour improvement. However, for flexible commuting, staggered activities should be promoted more.

In order to contribute to the improvement of passenger congestion and train delays, further work should consider the characteristics of different lines and different situation of passenger alighting and boarding situations and simulate more cases reflecting other demand change deal. station situation and actual operations more.

References

Carey, M., 1999. *EX ante heuristic measures of schedule reliability*, Transportation Research Part B, Vol.33, pp473-494.

Hürlimann, D., Nash, A. 2004. "Railway simulation using Opentrack", In: *Proceedings of The 9th International Conference on Railway Engineering Design and Optimization (CompRail2004)*, Dresden, Germany.

Janecek, D., Weymann, F. 2010. "LUKS – Analysis of lines and junctions", In: *Proceedings of The 12th World Conference on Transport Research (WCTR2010)*, Lisbon, Portugal.

Van der Meer, D., Goverde, R., Hansen, I. 2010. "Prediction of train running times and conflicts using track occupation data", In: *Proceedings of The 12th World Conference on Transport Research (WCTR2010)*, Lisbon, Portugal.

Ushida, K., Makino, N., Tomii, N., 2011. "Increasing Robustness of Dense Timetables by Visualization of Train Traffic Record Data and Monte Carlo Simulation", In: *Proceedings of The 9th World Congress on Railway Research (WCRR2011)*, Lille, France.

Graffagnino, T., 2012. "Ensuring timetable stability with train traffic data", In: *Proceedings of The 13th International Conference on Railway Engineering Design and Optimization (CompRail2012)*, New Forest, UK.

Yamamura, A., Koresawa, M., Adachi, S., Tomii, N., 2013. "How we have succeeded in Regaining Punctuality on the Tokyo Metropolitan Railway Network", In: *Proceedings of The 10th World Congress on Railway Research (WCRR2013)*, Sydney, Australia.

Hansen, I., Pachl, J. (eds.), 2014. *Railway Timetabling & Operations: Analysis – Modelling - Performance Evaluation 2nd Edition*, DVV Media Group, Eurailpress.

Yamamura, A., Koresawa, M., Adachi, S., Tomii, N., 2014. "Taking effective reduction measures using delay elements of indices on the Tokyo metropolitan railways", In: *Proceedings of The 14th International Conference on Railway Engineering Design and Optimization (CompRail2014)*, Rome, Italy.

Adachi, S., Hrata, N., Takatori, Y., Koresawa, M., 2016. "Research on train operation stability by installation of platform doors", In: *Proceedings of The 15th International Conference on Railway Engineering Design and Optimization (CompRail2016)*, Madrid, Spain.

Locomotive rotation optimization as basis for efficient rail cargo operation

(B: Professional paper)

Thomas Albrecht[a], Jonatan Gjerdrum[b]
[a] Rail & Transit Solutions, DXC Technology
Bergstraße 2, 01069 Dresden, Germany
[1] E-mail: talbrecht@dxc.com, Phone: +49 (0) 351 47771 45
[b] Planning, Green Cargo
Svetsarvägen 10, 17141 Solna, Sweden

Keywords
Rail cargo, Locomotive Rotation, Mixed-Integer Optimization, Timetable optimization

1 Introduction

To withstand the high pressure of the competition in the rail cargo market, railway undertakings must operate at highest possible efficiency. Optimal resource utilization is an important prerequisite for high efficiency, in particular for locomotives which represent the most expensive resource in rail cargo operations. Most rail cargo operators use computer-aided manual rotation planning, which often does not produce optimal plans because strong variations in traffic demand and irregular traffic in cargo traffic makes planning difficult.

The use of mathematical optimization well integrated into the IT landscape of a rail undertaking can help to reduce the manual planning effort, quantify the number of resources needed to operate a plan and use the available resources in an optimal way.

Green Cargo, the largest Swedish rail cargo undertaking, has been utilizing optimization systems for locomotive planning since the 1990's and introduced a new "Locomotive Optimization System (LOOP)" in 2017. This contribution describes this solution which has been developed by DXC Technology in close cooperation with Green Cargo. It is used for two major problem classes of locomotive planning:

- Tactical (operational) level: This task shall provide a day-to-day plan to produce the actual transport demand. It is performed monthly at Green Cargo.
- Strategic level for yearly planning and strategic scenarios: Here the traffic demand of one template week is assumed to be repeating weekly. Even though this is hardly true on a detailed level for cargo traffic, the assumption can be considered valid for strategic purposes when suitable traffic schemes are chosen, e.g. those of high traffic demand for determining required fleet size. Furthermore, all changes involving adaptations of timetables need to be aligned with the infrastructure manager and are considered as strategic optimization problems.

The article describes the models used for the mathematical optimization of locomotive rotations, how these are applied to practical planning problems and their integration in the planning process and toolchain. Optimization results will be presented in practical case studies based on real-world operating scenarios provided by Green Cargo.

2 Optimization models for tactical planning

2.1 Description of the Optimization problem

Green Cargo uses in the order of 350 locomotives for all its operations in Sweden, Norway and Denmark. These locomotives are optimized to provide the maximum efficiency to the current traffic program. There is no home station to the locomotives which might restrict their potential of usage but some locomotives have special features such as remote control (most), or ability to run on ETCS controlled tracks (few).

The purpose of locomotive planning in general and optimization with respect to scheduling in particular is to satisfy the traffic program requirements with the lowest possible asset number while maintaining robustness and customer satisfaction. Thus, the timetable planning (the most important part of the traffic program in this context) and locomotive scheduling problems are somewhat integrated. Green Cargo applies for timetable slots provided by infrastructure managers (mainly Trafikverket) but maintains a database for many more timetable scenarios which are evaluated in terms of locomotive utilization (as well as other resources such as crew). It is of key importance to apply for the right timetables at the right time with the right pulling power requirements i.e. occasional multiple locomotives. The problem is further complicated by the fact that Green Cargo competes for track capacity with other freight and passenger operators both long haul and commuter traffic and that much of the Swedish network is single-track only. Green Cargo therefore needs to interact closely with infrastructure managers to clarify its requirements both on a long-term and short-term horizon.

The starting point for locomotive optimization is a specific version of the timetable, be it the yearly plan or a monthly update of it. This timetable contains all Green Cargo operated trains and the requirements on locomotives based on so-called task classes which incorporate the special features of the locomotives. A task class comprises one or several locomotive types of similar driving characteristics. Using task classes allows the infrastructure manager to compute train running times with sufficient reliability whereas the railway undertaking still maintains a minimum level of freedom in the use of the actual locomotive types. Beside the task class, the minimum number of locomotives of each task class is also defined in the timetable (i.e. one or two or in rare instances three locomotives).

The required task classes for a train run can change along the journey. Therefore, a train run is split into so-called train legs at operational locations, where a change of task classes is required according to the timetable. To increase planning flexibility, train runs are also split at operational locations, where locomotive changes are allowed and sufficient time is available in the timetable, thereby creating more train legs.

The task of the monthly planning cycle is to identify which locomotive types to use in which number on each train leg (assignment problem) and to find rotations for each locomotive, i.e. the sequence of train legs a locomotive shall run on during the planning period. The monthly planning problem is solved as so-called dated planning problem, i.e. each train run is considered individually for the planning, even if it repeats several times on different days during a month or week. In the process it shall be possible to consider multiple locomotive types at the same time. This optimization problem is decomposed into a three-stage process, see Figure 1:

First, the possible combinations of locomotive types are computed for each train leg (stage 1), then an optimal assignment is searched for in two steps (stage 2.1 and 2.2). The rotation plan is handled as separate decision problem in stage 3.

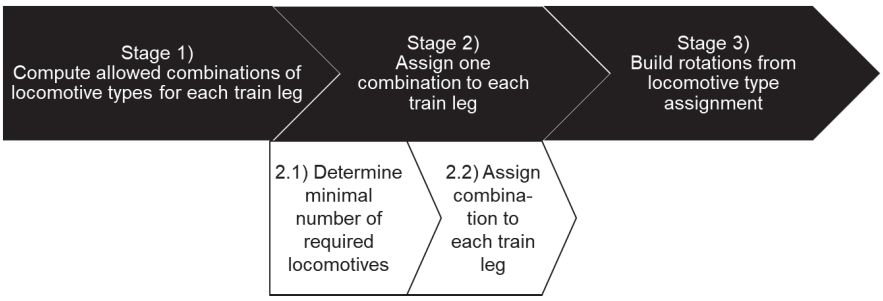


Figure 1: Optimization process architecture

2.2 Finding allowed combinations of locomotive types

For each train leg, the minimal number of required locomotives and their task classes is given. The maximum number of active locomotives depends on the infrastructure used and can be taken from the infrastructure model (typically two or three active locomotives).

For a specific train leg, the use of a specific locomotive type is possible under the following conditions:

1. All power supply systems installed along the train leg must be available on all locomotive types of the active locomotives.
2. One train protection system installed on the track must be available on the locomotive type of the leading locomotive.
3. The use of a locomotive type must be permitted along the entire train leg (The use of some locomotive types might be restricted to certain regions only)

From these constraints, the permitted locomotive types Θ_l for a train leg l are derived. From these permitted locomotive types, the possible combinations $c \in C_l$ for this train leg can be computed. Combinations are only allowed when all locomotive types in a combination are compatible to run in multiple traction with each other. Each combination c consists of a number $\lambda_{c,\theta}$ of locomotives of type θ.

2.3 Modelling approaches for the assignment problem

In a valid solution of locomotive assignment each train leg in the timetable must be served by exactly one of the allowed combinations of locomotive types. Therefore, decision variables $x_{l,c}$ are introduced which take the value 1, if train leg $l \in L$ is served by the combination c of locomotive types and 0 otherwise. The number of locomotives $\lambda_{\theta,l}$ of type θ on leg l can be expressed as

$$\lambda_{\theta,l} = x_{l,c} \cdot \lambda_{c,\theta}$$

For any valid assignment the following constraint must be fulfilled (one combination on each leg):

$$\forall l: \sum_{c \in C_l} x_{l,c} = 1$$

Furthermore, it must be guaranteed, that the same locomotive is being used only on train legs, which do not overlap in time. In the literature (see e.g. (Aronsson, M. et al., 2006) and (Giacco, G.L. et al., 2011) for an overview) this constraint is modelled as a multi-commodity network flow problem in a graph using two different approaches (see also Figure 2):

1. *Connection edges*: Each train leg is modelled as a vertex v in the graph. The potential connections of locomotives between train legs are modelled by directed edges e in the graph, i.e. edges are created between each arriving and departing train leg at the same operational location, where the departure time is later than the arrival time (See e.g. Aronsson, M. et al., 2006). Supplementary decision variables $s_{e,\theta}$ are introduced for the number of locomotives of type θ on edge e.
2. *Waiting edges*: Each departure and arrival of a train leg is modelled as a vertex v in a graph. Train legs are modelled as directed edges between these vertices. Furthermore, waiting edges are introduced between consecutive (departure or arrival) vertices at the same operational location o (See e.g. BMWi project, 2005). Supplementary decision variables $s_{e,\theta}$ are introduced for the number of locomotives of type θ on each waiting edge e. When two different trains arrive or depart at the same time, a waiting edge of length 0 is introduced between the vertices, therefore there exists always exactly one incoming waiting edge and one outgoing waiting edge and either a departing train leg edge or an arriving train leg edge for each vertex.

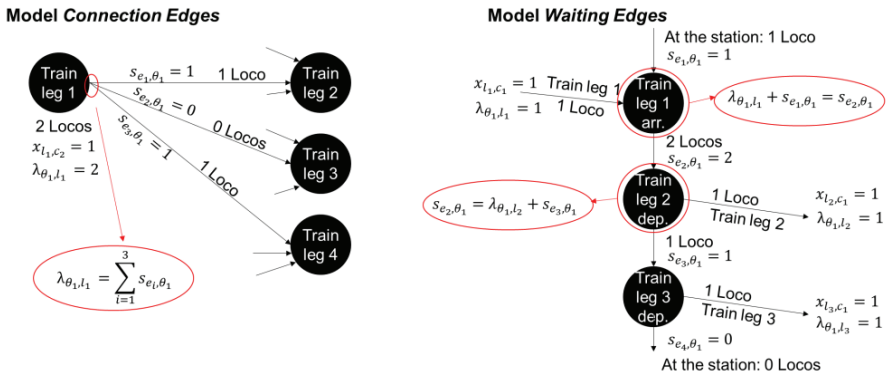


Figure 2: Comparison of the two different graph modelling approaches (assumption: one locomotive type θ_1 only, the combination c_1 has one locomotive, c_2 two locomotives)

For a valid solution, the flow constraint must be fulfilled on each vertex in the graph, where for each locomotive type the number of locomotives on the incoming edges e_{in} must be equal to the number on the outgoing edges e_{out}.

In the *connection edges* model both numbers are equal to the number of locomotives used on train leg l and can be expressed by:

$$\forall \theta \in \Theta, v \in V: \sum_{e_{\text{in},v}} s_{e,\theta} = \lambda_{\theta,l} = \sum_{e_{\text{out},v}} s_{e,\theta}$$

In the *waiting edges* model, for each departure or arrival event of any train leg at an operational location o (modelled as vertex v) the following flow constraint must be fulfilled:

$$\forall \theta \in \Theta, o \in O: \forall v \in V_o: \lambda_{\theta,l_{\text{arr}}} + s_{e_{\text{in},v},\theta} = \lambda_{\theta,l_{\text{dep}}} + s_{e_{\text{out},v},\theta}$$

where either $\lambda_{\theta,l_{\text{arr}}} > 0$ or $\lambda_{\theta,l_{\text{dep}}} > 0$.

The number of decision variables per operational location o and locomotive type is

$l_{\mathrm{dep},o} \cdot l_{\mathrm{arr},o}$ for the *connection edges model* and $2 \cdot (l_{\mathrm{dep},o} + l_{\mathrm{arr},o})$ in the *waiting edges model*.

Additional decision variables can be introduced in the *waiting edges* model to represent the number of locomotives of each type at the start and the end of the considered planning period at each operational location. In the *connection edges* model virtual train runs could be introduced as additional vertices to represent the possible start and end conditions. In practice, this constraint is not used.

During experiments it was shown that both models provide advantages and shortcomings: in the model using *connection edges* constraints for locomotive transfers in stations can easily be considered, e.g. whether there is enough time to couple locomotives at a station. The *waiting edge* model on the contrary requires significantly less decision variables and therefore typically computes in shorter time. Furthermore it allows for the explicit consideration of parking capacity at the stations, although this is currently not considered in the implemented model.

In the previous locomotive planning system used at Green Cargo the dated planning was performed manually based on weekly optimization. Dated optimization as provided by LOOP reduced the time required and the manual work to produce a plan as well as the restrictions for optimization based on manual input. Optimization across various locomotive types also improves the solution obtained.

2.4 Objective function and optimization approach

The objectives for locomotive optimization are manifold and partially contradictory as is often the case when optimizing both on cost and quality. The main objective is to produce the lowest number of locomotives that could satisfy all train legs as well as all constraints. However, extra costs are introduced if there are passive moves of locomotives (a.k.a. deadheading), if there are more locomotive changes in a train on-route or if an expensive locomotive is run when a cheaper one could have been used.

Thus, some of the objectives originally specified by the Green Cargo were:

1. Reduce the number of locomotives needed
2. Reduce the overall distance travelled by all locomotives (compute a plan with the minimal effort for re-positioning locomotives)
3. Consider the running cost of different locomotive types
4. Create robust locomotive rotations (avoid short connections between consecutive locomotive runs, in particular when trains serving different business areas are combined)
5. Avoid overlapping of pre- and post-processing times if possible
6. Try to ensure certain connections between train legs
7. Avoid changing compositions of locomotives (multi traction)

During an intensive experimental phase different approaches of multi-objective optimization (Branke et al., 2008) for producing locomotive rotations have been implemented and the obtained results were examined by the Green Cargo planners. As a compromise between solution quality, computation time and controllability of the solution a combination of lexicographical ordering (which leads to a decomposition of stage 2) and weighted sum has been chosen. In order to satisfy the most important goal of optimization (minimal number of locomotives in a plan) this number is used as objective in a first optimization run without considering any other objective (2.1, see also Figure 1). In all consecutive optimization stages, this minimal number of required locomotives is considered as additional constraint.

The second most important objective is the reduction of operating costs including e.g.

running costs per locomotive types which is considered in a second stage (2.2) as a weighted sum of distance dependent cost per locomotive type on each train leg, considering also different cost for active or passive locomotive usage. For stages 2.1 and 2.2 the waiting edges model is used.

In stage 3, rotations are built. The assignment of locomotive types obtained from stage 2.2 is used as additional constraint in this stage. The fulfilment of the flow constraints (in stages 2.1 and 2.2) guarantees that a valid rotation plan can be built. Decisions on this stage must only be made if more than one locomotive is assigned to a train leg or more than one locomotive is waiting at an operational location which results in a significantly smaller number of decision variables. The objective function is a weighted sum of penalties for unwanted connections within a rotation, e.g. too short connection times, breaking desired connections between consecutive legs of the same train run or between pre-specified pairs of train legs, combining legs of different business areas, uncoupling/ coupling of multi-traction when it can be avoided.

For the locomotive assignment problem, the optimum can only be reached by considering all locomotive types at the same time. Because of the practical assignment of locomotive types to task classes however it is possible to decompose the model into so-called subproblems and thereby reduce computation time significantly (e.g. by treating diesel-hauled and electric locomotive types separately). These subproblems are also created where dedicated fleets shall be used to operate special kinds of traffic (e.g. for postal trains which run at higher speeds than other cargo traffic).

The process of building rotations (stage 3) is executed separately per locomotive type.

3 Strategic optimization problems

Even though a timetable considers a yearly time frame it is practical to extract a specific week and make that week representative of a time period. This is called cyclic planning and for this type of strategic optimization problems, a repeating week is assumed. Cyclic planning is further used to analyze (seasonal) traffic patterns and derive potential measures to control dated planning (by so-called locks, see section 4.2).

The assumption of a cyclic planning problem means that the assignment of locomotives at the end of the considered template week (assignment to train legs and stock in stations) must be equal to the state at the beginning of the template week. In the *waiting edges* model these constraints can be introduced by adding a waiting edge for each operational location starting at the last vertex in the planning period and connecting it to the first vertex of the planning period. In the *connection edges* model connection edges are introduced from an arriving edge to all departing edges regardless of their departure time. If the departure time is earlier than the arrival time of the incoming train the connection edge represents a number of locomotives waiting over the end of the template period.

To compute the number of locomotives needed in cyclic planning, the number of locomotives used on all edges at any time point within the planning interval must be summed up. In the *waiting edges* model this is straightforward, for the connection edges model several special cases have to be considered in particular for very long train runs (Aronsson, M. et al., 2006).

In strategic planning, there are opportunities to create timetables with a better fit to locomotive rotations. Trains could be shifted in time so as to create shorter, more efficient standstills in terminals which in turn can reduce the number of locomotives. The approach presented in (Aronsson, M. et al., 2006) based on a *connection edge model* has been extended for multiple locomotive types and integrated in the LOOP solution. The *waiting*

edge model is not suitable for this kind of problem as the network topology changes for different timeshift.

4 Practical application

4.1 Implementation in IT solution

Using the described optimization algorithms in practice requires an IT solution which is fully integrated in the IT landscape and thus allows for a high degree of automation of the planning process. LOOP is based on DXC's Rail Cargo Management Solution RCMS (DXC, 2018), which had been equipped with interfaces to the systems used for timetable planning (of the infrastructure manager), crew planning and railway operations management. Infrastructure data is imported from the infrastructure manager system. Different validity periods of infrastructure data are modelled in LOOP in order to consider (future) changes in network topology for simulation of scenarios. LOOP is the leading system for all data on locomotive types which is used to assure the compatibility between locomotive and tracks and locomotives of different types among each other. These comprehensive network and locomotive models allow a high degree of automation of the planning process.

The rotation creation and optimization process use the standard RCMS scenario technology: The timetable is imported into a so-called timetable scenario, for which different resource scenarios can be created which contain the (interative improvements of the) locomotive rotations. The planning solutions obtained by the optimizer are presented to the planners in different GANTT charts. Here the planners can analyze the results and change the plans interactively. By introducing so-called locks between one or multiple consecutive train legs they can create input for a next optimizer run. There are so-called hard-locks on connections, which must not be broken by the optimizer, and soft-locks, which can be broken at the cost of a penalty only. The planning results are also presented in tabular format which can be exported for further analysis.

If any of the optimization runs does not find a suitable solution (in stage 1 of the optimization process), LOOP provides different views to analyze the root causes of the infeasibility and includes specific optimization problems to identify infeasible legs.

The solution is built in Java and incorporates CPLEX as solver for the different optimization models. The LOOP system is in full productive use at Green Cargo since summer 2017.

4.2 Planning process

In the yearly process, the aim is to provide a template for the coming year and make sure that the locomotive fleet is sufficient to enable the traffic program or to propose changes to the fleet sizes. At the same time, a number of productivity targets are set and changes to timetables are proposed. The locomotive planner is both a stakeholder and a support person in this process. Timing between arriving trains and departing trains in a station is crucial to create locomotive rotation plans that are efficient and robust. Therefore, several train planning related strategic optimization methods are tested to reduce the number of locomotives required whereas maintaining low effects on the traffic program: timetable shift, passive moves and remove trains in a pre-specified set. These strategic optimization methods based on a tight integration between timetable, traffic program and locomotive optimization are foreseen to have positive implications on the future locomotive plans. They are currently under test at Green Cargo.

The implemented yearly timetable is the basis for the monthly plan. The monthly

process is repetitive in nature and mainly reacts to factors such as (mainly smaller) variations in business volumes, track work and sometimes locomotive maintenance. Typically, the planner imports all relevant trains from the timetable system and iteratively optimizes the plan for the month of interest until a sufficiently good match between trains and locomotive rotations is found. Manual input based on expert knowledge is made both in the timetable system and in LOOP, typically by balancing the number of trains to and from stations, e.g. by empty runs, or by controlling the optimizer through parameter settings or restricting which locomotive turns are permitted.

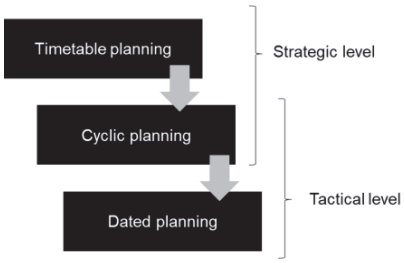


Figure 3: Overview over planning levels

5 Case studies

In order to illustrate the effects of the different models on computation time they have been run in different real-world scenarios. The results of these experiments (one run per scenario) are given in Table 1. The small scenario comprises one part of Green Cargo traffic which is run with a dedicated locomotive fleet. The big scenario contains all electrically-hauled traffic which is not run by dedicated fleets. The connection edges model is only used in cyclic planning. In dated planning the plan for a full month (31 days) is computed with the waiting edges model.

The number of train legs and the obtained number of locomotives gives an indication of the problem size. The small difference between the connection edges model and the waiting edges model in the number of train legs is due to the fact, that some hard-lock constraints are not considered in the connection edge model. The number "non-zeros" refers to the non-zero elements in the problem matrix after pre-processing by CPLEX. It is a good indicator on problem complexity and computational effort (Hill, F. et al., 1984). It can be seen that this number is approximately 20 times higher in the big scenario with the connection edges model compared to the waiting edges model. The computation time is compared for the first step in the locomotive assignment process, i.e. the computation of the minimal number of required locomotives and the proof of optimality by CPLEX. All experiments ran on a 4-core Intel Server with 2.6 GHz processor and 8 GB RAM. It can be seen that the small problems solve immediately regardless of the used model, but there are significant differences in the computation time of the big problem. The connection edges model takes very long to compute a valid solution in the cyclic problem. The waiting edges model performs significantly better in the cyclic problem and even the computation time for the big dated planning problem with three times more train legs is shorter than the one required by the connection edges model for cyclic planning. The table also shows the total computation times of the rotation building process (including all stages 2.1, 2.2 and 3).

The computation times are acceptable from the viewpoint of Green Cargo for this kind

of planning problems. Even in the dated planning they allow for several planning iterations per working day. It has also been shown that the waiting edge model is able to compute a reliable lower bound for the number of needed locomotives within one CPLEX iteration, i.e. within a few seconds. This property is used in practice to speed up dated optimization of the big problem to less than one hour.

Table 1: Computational results (Computation times are given in h:mm:ss)

	Cyclic planning				**Dated planning**	
Scenario	Small	Small	Big	Big	Small	Big
Model used	Conn. edges	Waiting edges	Conn. edges	Waiting edges	Waiting edges	Waiting edges
Train legs	70	70	1 746	1 671	290	6 926
Non-zeros	1 292	466	1 119 454	57 226	1 887	238 159
Locomo-tives	9	9	161	164	9	169
Comp. time assign-ment	<0:00:01	<0:00:01	5:00:14	0:02:17	<0:00:01	0:30:14
Comp. time full	-	<0:00:01	-	0:03:32	0:00:01	1:21:13

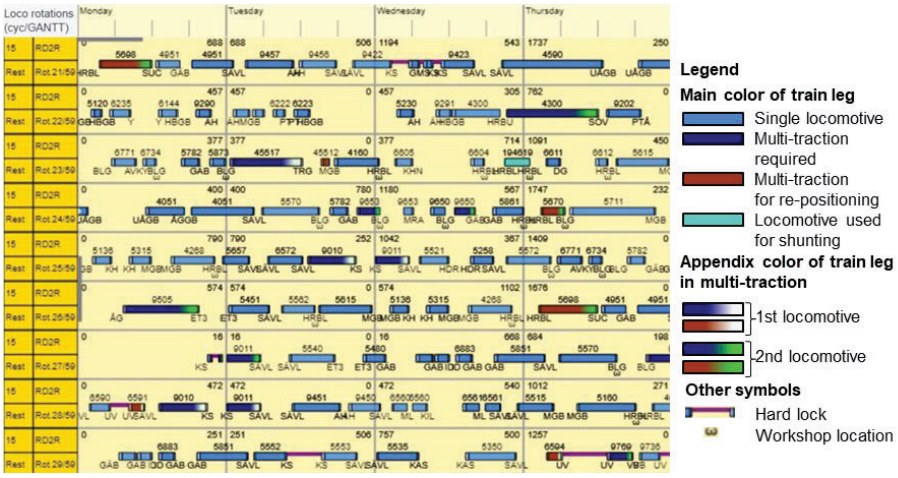


Figure 4: Excerpt of a GANTT view of the locomotive rotations in LOOP for a cyclic plan

Figure 4 shows an excerpt of the main GANTT chart displaying results of the cyclic optimization of the big scenario. One locomotive rotation is displayed per row. In the left column (orange, the rotation header), locomotive type (here: RD2R), and rotation number within a cycle are given (e.g. Rot. 25/59 is week 25 in a rotating cycle of 59 weeks in length). It should be noted that the planning cycle length is not part of the planning

objectives. In the right (yellow shaded part) of the figure, the sequence of train legs planned for the locomotive is displayed. The different colors, shadings and symbols allow for quick check of the plan efficiency and correctness by the planners.

6 Conclusions and further Research

The introduction of the Locomotive Optimization System LOOP allowed Green Cargo to reduce the number of locomotives needed to operate the timetable by increasing the so-called locomotive productivity, i.e. the distance travelled in commercial operation per locomotive. Moreover, process and integration development increased planning speed and improved the solution quality. The perceived key benefits from a Green Cargo planning process perspective are a lower amount of user restrictions allows for better optimization, dated optimization and planning of multiple locomotive types at the same time allow for less manual work and a better integration with the timetable and network data that maintains feasibility of locomotive types on the track network.

A further improvement of locomotive usage is foreseen with a tighter integration of LOOP with the surrounding planning processes and tools:

1. A tighter integration with load and timetable planning would allow to fix the locomotive categories and required minimal number later in the planning process allowing for improved usage of the most recent powerful and multi-purpose locomotives.
2. An integration with crew scheduling is challenging in particular in cargo operation and under the Swedish geography with long travelled distances. But it could be highly beneficial as it would allow reducing the number of manual constraints (locks) that are introduced today by the planners to consider driver constraints during locomotive planning and might thereby reveal new options for driver and locomotive rotations.
3. A tighter coupling with maintenance planning and workshop task scheduling is particularly interesting when workshop capacities are limited and/ or new methods and processes for preventive locomotive maintenance are introduced.
4. The application of LOOP for operational planning and adaptation of locomotive rotations considering real-time information on timetable deviations would make it possible to profit from the optimization capabilities during unplanned events.

References

Aronsson, M., Kreuger, P. and Gjerdrum, J.: *An efficient MIP Model for Locomotive Scheduling with Time Windows*, ATMOS 2006

Giacco, G.L., D'Ariano, A., Pacciarelli, D.: Rolling stock rostering optimization under maintenance constraints, In: *2nd International Conference on Models and Technologies for Intelligent Transportation Systems*, Leuven, Belgium, 2011.

BMWi-Projekt Innovationen für Gleisanschlussverkehre, 2005

Branke, J., Deb, K., Miettinen, K., Slowinski, R. (eds.): *Multiobjective Optimization, Interactive and Evolutionary Approaches*, Springer, LNCS 5252, 2008.

DXC RCMS Product Sheet: http://www.dxc.technology/travel_and_transportation/offerings/43437/81345-rail_cargo_management_system, last access: 01.02.2018

Hill, P.E.., Murray, W., Saunders, M.A. and Wright, M.H.: *Sparse matrix methods in Optimization.* SIAM J. Sci. Stat. Comput., Vol. 5, No. 3, 562-589, 1984

An Assessment of Virtual Integration for Passenger Rail Services in Great Britain

Sultan Alsaedi [a], John M. Preston [a]
[a] Transportation Research Group, Faculty of Engineering and Physical Sciences University of Southampton, Southampton Boldrewood Innovation Campus, University of Southampton, Burgess Road, Southampton SO16 7QF, UK

Abstract

Vertical integration was introduced in the British railway system in the form of a virtual alliance between Network Rail (NR) and South West Trains (SWT). The introduction of this alliance in 2012 was due to the Rail Value for Money Study that was published by McNulty in 2011. However, this alliance was ended in 2015, which was two years earlier than initially agreed by the Department for Transport (DfT). This paper aims to investigate whether the performance quality, in terms of punctuality and reliability, was a reason to end this alliance. The investigation is based on a comparison of the performance quality of SWT with other comparable Train Operating Companies (TOCs), which are Govia Thameslink Railway (GTR) and Southeastern (SE). Furthermore, the measurements of the Public Performance Measures (PPM) and Cancellation and Significant Lateness (CaSL) of these TOCs were used to deliver the comparisons. As a result, the investigation indicated that punctuality and reliability are not influenced by whether the organisation is vertically separated or virtually integrated. Overall, the virtual integration in this case does not seem to have had an impact, on the overall performance quality of passenger rail services provided by SWT.

Keywords
Capacity Utilisation, Punctuality, Reliability, Vertical Separation, Virtual Integration.

Glossary
CaSL: Cancellation and Significant Lateness.
GTR: Govia Thameslink Railway.
IM: Infrastructure Manager.
NR: Network Rail.
ORR: Office of Rail and Road.
PPM: Public Performance Measure.
RU: Railway Undertaking
SE: Southeastern.
SWT: South West Trains.
TOC: Train Operating Company.

1 Introduction

Punctuality and reliability are important factors to measure the performance of the railway system in terms of quality of service and passenger satisfaction (Carey, 1999; Goverde, 2005; Yuan, 2006). These indicators may deteriorate when the railway network is more extensively utilised to accommodate the growth in demand rather than extending or upgrading the track network (Yuan, 2006; Yuan and Hansen, 2007). There are several methods that are used to

optimise the performance by designing a robust timetable or control strategy (Carey, 1998; Parbo et al., 2016). Optimal scheduling techniques, for example, can be implemented in the railway system in order to plan, operate and manage passenger train services, and these techniques, for example, can help to maintain the conflict between trains that operate on a single track (Ferreira and Higgins, 1996). Despite the development of scheduling techniques, there is a lack of these methods to mitigate the impact of delay on the performance (Carey and Carville, 2000).

The British rail system was vertically separated as a result of the 1993 Railways Act, with the key distinction being between the Infrastructure Manager (IM), since 2002 Network Rail (NR), and the Railway Undertakings (RUs), known as Train Operating Companies (TOCs). However, there have been long standing concerns about the weak alignment of incentives between the RUs and the IM caused by this vertical separation (Preston, 2002). This issue was revisited by the Rail Value for Money Study chaired by McNulty (2011). Partly as a result, an experiment was conducted between the Wessex Route of NR and the dominant TOC on the route, South West Trains (SWT), in which a form of virtual integration was introduced. The key features of this deep alliance were a single senior management team responsible for trains and track and the joint operations of the Waterloo control centre. This alliance was approved by the Department for Transport (DfT) and the then Office of Rail Regulation (ORR), as regulatory bodies, for a period of five years starting in 2012. In 2015, the DfT announced the end of the virtual alliance two years earlier than scheduled. Given this background, this paper aims to investigate whether changes in punctuality and reliability were reasons to end the virtual alliance between NR and SWT. Therefore, the paper is structured as follows. Section 2 provides a brief description of punctuality and reliability and their causes. In addition, the change in the railway organisation due to the reform is outlined briefly in this section. Section 3 illustrates the research methods that are used to achieve the aim. Section 4 contains a discussion of the results obtained by the three methods. Section 5 draws the final conclusion and key findings.

2 Literature Review

2.1 Punctuality and Reliability: definitions and causes

Punctuality and reliability have various definitions according to different literature. Punctuality is usually related to the running time with respect to an acceptable deviation from the designed timetable, which means a train is considered as punctual if this train runs within the accepted deviation (Olsson and Haugland, 2004; Preston et al., 2009). Punctuality is often described as the proportion of the trains arriving at, passing or departing from a point with a delay lower than a particular time, usually in minutes (Veiseth et al., 2007; Yuan, 2008). The deviation to determine the punctuality of trains varies between railway systems. In Great Britain, the amount of deviation to determine the punctuality relies on the journey length of the service; a train is described as punctual if it arrives at its final destination within five minutes of the timetabled arrival, but the deviation is increased to ten minutes for long distance services (ORR, 2016; Preston et al., 2009). For Switzerland and the Netherlands, a train can be described as punctual if it arrives within four minutes and three minutes respectively (Yuan, 2008). Reliability, on the other hand, is often implemented to illustrate the ratio of trains that have been cancelled (Preston et al., 2009). By contrast, Barron et al. (2013) define the reliability as the predictability of given travel time being experienced by a passenger and the degree of variation around the average travel time. According to Vromans

(2005), the reliability of the railway system depends on whether the trains are operated according to the scheduled timetable.

There is a substantial correlation between the punctuality and reliability indices with the overall delay in train operations. According to Yuan (2008), train delays are classified into three categories. Firstly, the initial delay is recorded when a train crosses the boundary of the investigated network later than timetabled (Yuan, 2008). Secondly, the original delay is the delay caused in the network due to operating trains at a lower speed compared to the scheduled speed, technical faults in the network, excess passenger boarding time and weather conditions (Yuan, 2008). The third category is knock-on delays, and this term is used to describe the delay that is transmitted between trains in the network (Yuan, 2008). When a train is delayed, the other trains that operate on the same route will be delayed (Parbo et al., 2016; Yaghini et al., 2013; Yuan, 2008). A related classification of delays is primary and secondary delays. The primary delay is the direct impact of several factors on the train itself, while the transmitted delays between trains are called secondary delays (Preston et al., 2009; Veiseth et al., 2007; Yuan and Hansen, 2007). According to Preston et al. (2009), the primary delays contributes to 40% of performance delay in the UK, while the remaining 60% is caused by secondary causes. This indicates that the initial and original delay, as stated above, are considered as primary delays, whereas the knock-on delay is defined as a secondary delay. According to Xia et al. (2013), there is a significant impact of bad weather conditions, such as high levels of wind, temperature, humidity and rainfall, on the rail performance, and this impact can lead to a significant delay in train operations. These reasons could have an impact on the train operation such as running the trains at lower speed or derailment.

To ensure punctuality, some aspects should be taken into consideration to design a railway timetable, as stated by Goverde and Hansen (2013). Infrastructure occupation should be considered with respect to three factors that can have an impact on capacity consumption, which are average train speed, number of trains and heterogeneity. These factors have an impact on the headway between trains. For example, the headway is influenced when trains run on the network at different speed levels; fast trains require larger headway due to longer breaking distance. Other aspects are that the timetable should be feasible and robust. Timetable feasibility means the ability of all planned trains to adhere to their scheduled routes. This can eliminate the conflicts between trains, which allow trains to run smoothly without braking. Conflict-free routes can be achieved when the process time of a train exceeds the scheduled time. On the other hand, a timetable can achieve robustness when it is capable of resisting design errors, parameter variations and changing operational conditions. For example, a process time for a scheduled train is calculated with basic parameters based on an estimation made by experts or determined by different methods. The robustness can absorb the design errors of this estimation when the estimated values are slightly different compared to the real values.

With regards to the railway system in Britain, monitoring the performance quality is highly dependent on two indicators, the Public Performance Measures (PPM) and Cancellation and Significant Lateness (CaSL). PPM is an indicator to measure the performance of train operations for passenger services in order to evaluate both the reliability and punctuality of the service (ORR, 2016). This indicator has two categories based on the journey length of the service to describe the status of the train if it is late or not (ORR, 2016). The first category is designed for regional operators, including London and South East operators; if a train arrives at its destination within five minutes compared to the timetable, this train will be considered as on time (ORR, 2016). The second category is designed for long distance services; a train will be defined as on time when it arrives at its destination within ten minutes compared with the designed timetable (ORR, 2016). CaSL, on the other

hand, is the proportion of passenger trains that have been cancelled or arrived at the last destination more than 30 minutes late compared with the designed timetable (ORR, 2016).

A National Task Force (NTF) sub-group proposed new performance metrics to replace PPM and CaSL (NR and ORR, 2017), and a brief description of these metrics is given here. Firstly, Total Passenger Lateness is an indicator to measure the total of time lost for passengers in million hours. This metric focuses on passenger rail serviced by TOCs. Secondly, 'Reliability – cancellations and severe disruption' is a metric to describe the proportion of planned trains that did not serve the full journey or skipped some planned station stops. Moreover, there is a cancellation weight for each train depending on if a train is cancelled fully or partially. This indicator aims to describe a pure reliability of rail services by excluding significant lateness compared to CaSL. Thirdly, 'On Time and Time to 15' metrics are used to describe planned trains that arrive at all recorded stations less than one minute (within 59 seconds) and 15 minutes (within 14 minutes and 59 seconds) respectively. These metrics aim to provide a better explanation of punctuality of rail services.

2.1.1 Previous Studies on Punctuality and Reliability

There are numerous studies to investigate the factors that influence the punctuality and reliability in the railway performance. Each study attempts to determine the factors that have a significant impact on passenger train services in terms of punctuality and reliability, and various methods and models were used for analyses based on the characteristics of variables and collected data (Vromans, 2005).

The first study to consider is the research led by Harris (1992). The purpose of this study was to study the punctuality of railway performance in the UK and Netherlands by selecting different factors. The factors that were considered are the number of previous stops, the length of the train, distance covered, the age of motive power unit and track occupation. The methodology that was used by Harris for analysing was least-squares multiple linear regression. As a result of Harris's analysis, the factors that influenced the determination of the punctuality were the train length and the covered distance.

The second study to consider is the research led by Olsson and Haugland (2004). The purpose of the study was to determine the factors that influence the punctuality on passenger train services in Norway. The factors considered in this study were passenger number, train capacity rate (passenger per seats), the usage of infrastructure capacity, cancellations, the construction work of the network, a temporary decrease in speed, the punctuality of departure and arrival and operational priority rules. As a result, it was found that the punctuality was influenced significantly by the determination of the usage of infrastructure capacity based on the timetable.

The Swedish National Audit Office led a study to investigate the factors influencing the punctuality and reliability between 1976 and 1986. The study found about 50% of the delayed trains were caused by rainfall, temperature and patronage levels (Olsson and Haugland, 2004; Preston et al., 2009).With respect to the latter, if the level of patronage increased by 10% for a month, the punctuality dropped by about 6% on Sundays and about 10% on weekdays, especially on Fridays up to 14% (Olsson and Haugland, 2004; Preston et al., 2009). With respect to weather, the punctuality declined by around 5% as the average temperature decreased by one centigrade below -5C in one month (Olsson and Haugland, 2004; Preston et al., 2009).

The relationship between capacity utilisation (CUI) and congestion-related reactionary delay (CRRD) has been investigated. Armstrong and Preston (2017), for example, delivered research aimed to assess the relationships between capacity utilisation and rail performance,

particularly at junctions and stations. The key finding is that there some consistency between CUI and CRRD. The amount of delay escalates due to the increase in the level of capacity utilisation.

2.2 Railway Reform

Privatising the railway system was implemented in order to achieve certain aims. Preston (1996) listed the goals and aims of privatising the railway system, which are to maximise the use the of the railway system; to provide a better satisfaction to the rail users; to improve the performance quality of the railway system; and maximise the net economic benefits of the railway system. Another aim of the rail privatisation that was mentioned by Knowles (2013) is to provide a competitive market for the private sector by limiting the role of the governments in order to improve the performance efficiency and to provide better benefits to the rail users. As reliability and punctuality are used to measure the performance quality of the rail services in terms of the customer satisfaction (Goverde and Meng, 2011), privatising the railway system could provide more reliable and punctual rail services to the rail users.

Railway organisation has changed as a result of liberalising reforms. For instance, Amaral and Thiebaud (2015) illustrated the four types of organisation that have emerged in Europe. The first type is a fully vertically separated organisation, which means that the IM is separated fully from the RUs. The second type is vertically separated organisation with a delegation, which means that the IM and RU are separated, but the RU is responsible for at least some of the IM tasks. The third type is a vertically separated organisation within a holding company, which means that the IM and RU are separated, but both are owned by one holding company. The fourth type is a vertically integrated organisation, which means the IM and RU are managed and operated by one company. However, a new form of railway organisation that has been experimented with in Britain is virtual integration, which retains separation of the IM and RU but encourages joint working, particularly at the operational level.

According to Mizutani et al. (2015), the purpose of the variety of organisation in the railway system is to provide a competitive market for all parties that are involved in the rail market. In Europe, for example, the successive legislations originating with Directive 91/440 require at least an accounting separation between the IM and RU in order to provide a competitive rail market. Furthermore, the separated organisations generate two forms of competition in the rail market, which are open access competition for freight services and competitive tendering for domestic passenger services. However, there is a concern about performance efficiency when the organisation of the railway system is vertically separated or integrated. The concern is that the performance efficiency can deteriorate due to the transaction costs between the infrastructure and train operators and reduced incentives for efficiency and for appropriate investment by the IM (Drew and Nash, 2011).

2.2.1 Previous Studies on Railway Organisation

There are several studies that have attempted to investigate the impact of the organisation forms on the railway system. For example, research published by Merkert et al. (2010) shows that the impact of the vertical separation was not significant on the performance measurements. This research was based on the assessment of the performance efficiency measurements for a cross-section of countries, but the vertical separation may not be the main factor to measure performance efficiency. Similarly, Wetzel (2008) concluded that the performance efficiency is not influenced significantly by vertical separation. However, the cost of rail systems does seem to vary between vertically separated or integrated

organisations with regard to train density. Research by Mizutani et al. (2015) concluded that vertically integrated organisations are more beneficial in terms of unit costs at high levels of train density, while the vertically separated organisations are more beneficial at low levels of train density, as shown in Figure 1.

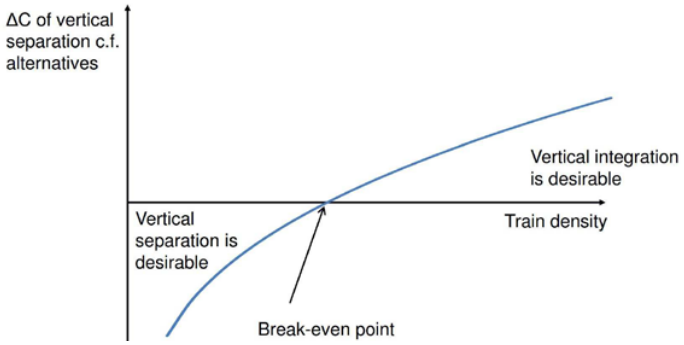


Figure 1: The effect of train density on the cost of different organisations (Source: (Mizutani et al., 2015))

2.3 The British situation

Since the British railway was reformed and privatised, demand levels and rail performance quality have changed dramatically. Figure 2 shows that passenger rail demand as measured by passenger-kilometres has increased substantially since the British railway was reformed (Merkert, 2005). This indicates that the rail performance required more attention to accommodate the increase of the demand, especially for the efficiency of the performance quality.

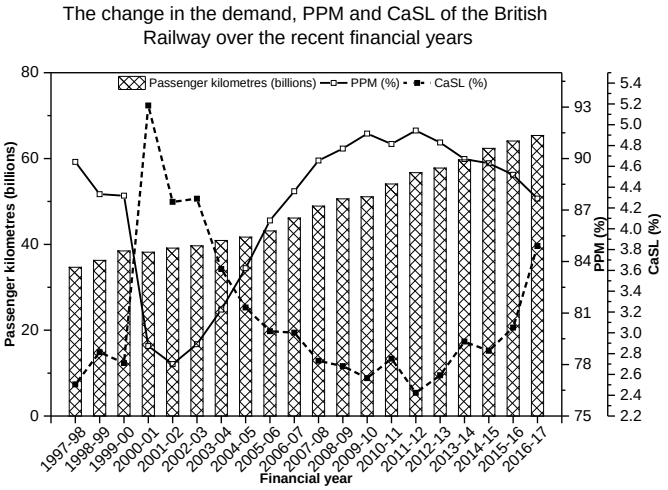


Figure 2: The change in the demand on the British Rail (Source: (ORR, 2017))

With regard to the performance quality, the change in the PPM and CaSL indicators is shown in Figure 2. According to this figure, there was a substantial adverse change in the PPM and CaSL measurements following the Hatfield accident between 2000-01 and 2002-03. This was exacerbated due to the sharp increase of the rail traffic in the rail network and the failure of Railtrack to maintain the track sufficiently (Drew and Ludewig, 2011). As a result, the railway industry in Britain was reformed by replacing Railtrack by NR in 2002 (Drew and Ludewig, 2011) and there was some recovery in performance.

However, the Rail Value for Money Study (McNulty, 2011) remained concerned about the misalignment of incentives. As a result a new scheme was proposed in the form of deep-alliances NR and the TOCs (Thompson, 2013). This would include the joint operations of control centres, with the purpose being to enhance rail performance and minimise the cost, which should improve customer satisfaction.

3 Methodology

The methodology of this research is based on data analysis techniques to investigate the impact of virtual integration on performance quality. The investigation is a comparison between the performance of SWT with similar operators, SE (Southeastern) and GTR (Govia Thameslink Railway). The process of the investigation is divided into three approaches. The first approach is an assessment of the change in the performance quality of SWT to evaluate the changes in the measurements of PPM and CaSL. The second approach is the organisation effect analysis to assess the effect of virtual integration on performance quality. The last approach is a prediction assessment of the performance quality of SWT to deliver a comparison between the actual measurements of PPM and CaSL with the predicted measurements. The data that is used for this investigation is collected from ORR.

3.1 Change Assessment

In this analysis, different approaches of comparisons are considered. The first approach is an individual comparison for SWT, which means that the PPM and CaSL measurements are used to assess the change in the performance quality performance prior, during and post the alliance for the mainline and suburban routes. For example, when the PPM measurements are considered, the proportion of trains arrived on-time of each route during the prior-alliance period is compared with the proportion during the alliance period and with the post-alliance period, and Table 1 outlines the start and end of prior, during and post-alliance periods as financial years and periods. The process is repeated similarly for the CaSL measurements.

Table 1: The allocation of the investigated periods in the financial years and periods.

Prior-alliance period		**Alliance period**		**Post-alliance period**	
Start	**End**	**Start**	**End**	**Start**	**End**
2010/11	2011/12	2012/13	2015/16	2015/16	2018/19
Period 01	Period 13	Period 01	Period 05	Period 06	Period 06

The second approach is ‘the cross comparison’, which means that the change in the performance quality of SWT is compared to the change of GTR and SE for each comparable route, and this comparison consists of two scenarios. Firstly, the change in the measurements of the PPM and CaSL is considered to compare the change of the SWT performance quality with the other TOCs. For example, the proportion of trains that arrived on-time for SWT

during the prior-alliance period is compared with the same proportion of each comparable route of GTR and SE during the same period. The process is repeated similarly for the remaining periods and for the CaSL measurements.

Secondly, the amount of change in the quality performance is assessed between the TOCs through prior, during and post the alliance. Moreover, the average change in the measurements of the PPM and CaSL indicators for SWT from the prior-alliance period to the alliance period and from the alliance period to post-alliance period are examined. In addition, these changes are compared with the average changes in the measurements of GTR and with SE. However, the average value of the PPM and CaSL indicators will be calculated based on the periods in the three investigated periods separately for each TOC. The change of each indicator then is obtained based on the following equations:

$$\mathrm{PI}_{\mathrm{C}_i} = \mathrm{PI}_{\mathrm{P}_i} - \mathrm{PI}_{\mathrm{A}_i} \tag{1}$$

Where:

$\mathrm{PI}_{\mathrm{Ci}}$ = The change of the average value of the performance indicator, whether PPM or CaSL, from the prior to the alliance period.
$\mathrm{PI}_{\mathrm{Pi}}$ = The average value of the performance indicator in the prior-alliance period.
$\mathrm{PI}_{\mathrm{Ai}}$ = The average value of the performance indicator in the alliance period.
i = Period.

$$\mathrm{PI}_{\mathrm{D}_i} = \mathrm{PI}_{\mathrm{A}_i} - \mathrm{PI}_{\mathrm{E}_i} \tag{2}$$

Where:

$\mathrm{PI}_{\mathrm{Di}}$ = The change of the average value of the performance indicator, whether PPM or CaSL, from the alliance to post-alliance period.
$\mathrm{PI}_{\mathrm{Ai}}$ = The average value of the performance indicator in the alliance period.
$\mathrm{PI}_{\mathrm{Ei}}$ = The average value of the performance indicator in the post-alliance period.
i = Period.

The test that is used for these comparisons is the hypothesis test of two sample proportions as the provided data is expressed in proportions (Johnson, 2001).

3.2 Effect Analysis

The effect of organisation forms on performance quality is assessed based on regression analysis. Moreover, different regression models are created for each performance quality indicator. These models contain numerical and categorical variables. The numerical variable is the average rolling stock age (RSA). The categorical variables are the financial years (FY), periods (P), route types (RT) and organisation forms (OF). As these variables cannot be implemented directly into the regression analysis, each group will be recoded as 1 and 0 through dummy coding, which is an approach to recode the categorical data to be applicable to use in the regression model (Fox, 2015, p. 118). The process of the dummy coding involves excluding one variable of each category to be set as a reference of this category and recoding the remaining variables. The dummy coding for the four categories are the following:

$$\text{TOC} = \begin{cases} D_{GTR} = 1, \text{if the TOC is GTR, otherwise } D_{GTR} = 0 \\ D_{SE} = 1, \text{if the TOC is SE, otherwise } D_{SE} = 0 \\ \text{SWT is the reference for this category} \end{cases}$$

$$\text{FY} = \begin{cases} D_{2011-12} = 1, \text{if } 2011-12 \text{ is the financial year, otherwise } D_{2011-12} = 0 \\ D_{2012-13} = 1, \text{if } 2012-13 \text{ is the financial year, otherwise } D_{2012-13} = 0 \\ D_{2013-14} = 1, \text{if } 2013-14 \text{ is the financial year, otherwise } D_{2013-14} = 0 \\ D_{2014-15} = 1, \text{if } 2014-15 \text{ is the financial year, otherwise } D_{2014-15} = 0 \\ D_{2015-16} = 1, \text{if } 2015-16 \text{ is the financial year, otherwise } D_{2015-16} = 0 \\ D_{2016-17} = 1, \text{if } 2016-17 \text{ is the financial year, otherwise } D_{2016-17} = 0 \\ D_{2017-18} = 1, \text{if } 2017-18 \text{ is the financial year, otherwise } D_{2017-18} = 0 \\ 2010-11 \text{ is the reference for this category} \end{cases}$$

$$\text{RT} = \begin{cases} D_{Suburban} = 1, \text{if the route is suburban, otherwise } D_{Suburban} = 0 \\ \text{The mainline route is the reference for this category} \end{cases}$$

$$\text{OF} = \begin{cases} D_{VS} = 1, \text{if the organisation is vertically separated, otherwise } D_{VI} = 0 \\ \text{The virtually integrated organisation (VI) is the reference for this category} \end{cases}$$

$$\text{P} = \begin{cases} D_{P02} = 1, \text{if period is the second period, otherwise } D_{P02} = 0 \\ D_{P03} = 1, \text{if period is the third period, otherwise } D_{P03} = 0 \\ D_{P04} = 1, \text{if period is the fourth period, otherwise } D_{P04} = 0 \\ D_{P05} = 1, \text{if period is the fifth period, otherwise } D_{P05} = 0 \\ D_{P06} = 1, \text{if period is the sixth period, otherwise } D_{P06} = 0 \\ D_{P07} = 1, \text{if period is the seventh period, otherwise } D_{P07} = 0 \\ D_{P08} = 1, \text{if period is the eighth period, otherwise } D_{P08} = 0 \\ D_{P09} = 1, \text{if period is the ninth period, otherwise } D_{P09} = 0 \\ D_{P10} = 1, \text{if period is the tenth period, otherwise } D_{P10} = 0 \\ D_{P11} = 1, \text{if period is the eleventh period, otherwise } D_{P11} = 0 \\ D_{P12} = 1, \text{if period is the twelveth period, otherwise } D_{P12} = 0 \\ D_{P13} = 1, \text{if period is the thirteenthperiod, otherwise } D_{P13} = 0 \\ \text{The first period is the reference for this category} \end{cases}$$

The regression model that will be used to estimate the coefficients for the PPM and CaSL indicators is the following:

$$\begin{aligned} PI = \beta_0 + \beta_1 RSA + \beta_2 D_{GTR} + \beta_3 D_{SE} + \beta_4 D_{2011-12} + \beta_5 D_{2012-13} + \beta_6 D_{2013-14} \\ + \beta_7 D_{2014-15} + \beta_8 D_{2015-16} + \beta_9 D_{2016-17} + \beta_{10} D_{2017-18} + \beta_{11} D_{P02} \\ + \beta_{12} D_{P03} + \beta_{13} D_{P04} + \beta_{14} D_{P05} + \beta_{15} D_{P06} + \beta_{16} D_{P07} + \beta_{17} D_{P08} \\ + \beta_{18} D_{P09} + \beta_{19} D_{P10} + \beta_{20} D_{P11} + \beta_{21} D_{P12} + \beta_{22} D_{P13} \\ + \beta_{23} D_{Suburban} + \beta_{24} D_{VS} + \varepsilon \end{aligned} \quad (3)$$

As the measurements of the PPM and CaSL indicators are limited between 0 and 1, these measurements are required to be transformed on the logit scale because the data is bounded

between 0 and 1 (Fox, 2015, p. 72). The formula that will be used for transforming is the following:

$$Logit(PI_i) = log \frac{PI_i}{1 - PI_i} \tag{4}$$

The regression analysis relies on disaggregated data published by ORR. This means that other explanatory variables are not included in the analysis, such as train length, service frequency and passenger rail demand, which could contribute to a better explanation of changes in performance quality.

3.3 Performance Quality Prediction

The objective of predicting the performance quality is to analyse and compare the actual measurements of the performance quality of the SWT with the predicted measurements, and this is delivered by two stages. Firstly, the actual measurements of PPM and CaSL of the prior-alliance period are used to forecast the measurements during the alliance. Secondly, the actual measurements of PPM and CaSL of the alliance period are used to forecast the measurements post the alliance.

The process of creating the forecast models is based on the Autoregressive Integrated Moving Average (ARIMA) model process. The ARIMA models contain the Autoregressive (AR) and Moving Average (MA) as the parameters of the ARIMA model are ARIMA(p,d,q), where p, d and q are related to the autoregressive order, the moving average order and the required difference of the model to achieve stationarity respectively (Washington et al., 2003, p. 180). The process of the ARIMA models according to Washington et al. (2003, p. 183) is as follows:

1. Plot the Autocorrelation Function (ACF) and the Partial Autocorrelation Function (PACF).
2. Estimate the parameters of the ARIMA model.
3. Check the accuracy of the model.

Plotting Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) is significant in creating the ARIMA model. As the ARIMA model is a combination of Autoregressive (AR) and Moving Average (MA), ACF and PACF are used to determine the order of these combinations. Moreover, Table 2 contains the process of selecting the ARIMA model and estimating the orders of these models. In some cases, ACF plot shows repeated significant lags in the same period. This indicates that there is a seasonal effect on the data, and this requires an upgrade of the ARIMA model to include the seasonal effect, which means that the seasonal ARIMA model (SARIMA) will be more appropriate for forecasting. The order of the SARIMA is SARIMA (p,d,q)x(P,D,Q)t, where p, d and q are stated above, and P, D and Q are the order of the autoregressive order, the moving average order and the required difference of the model to achieve stationarity respectively for the seasonal effect at period t. However, ARIMA models can be developed to accommodate the effect of periods as regressors, and this method is known as ARIMAX models.

After selecting the order of the ARIMA model, the parameters of the ARIMA model will be estimated in order to examine and check the selected model accuracy. This means that even if the order of the ARIMA model was selected based on the plots of the ACF and PACF,

the order can be modified based on the model accuracy process. The parameters will be estimated by using the application R.

Table 2: The process of selecting the ARIMA model and estimating the orders (Source: (Washington et al., 2003, p. 183)).

	AR(p)	**MA(q)**	**ARMA(p,q)**
ACF	Trails off exponentially	Cuts off after lag q	Trails off exponentially
PACF	Cuts off after lag p	Trails off exponentially	Trails off exponentially

The model check is divided into two approaches. The first approach is to assess the errors of the fitted values of the created model. This assessment is based on the equations that are provided by Washington et al. (2003, p. 190) such as root mean square error (RMSE). After calculating the accuracy measures, the interpretation of the results is as the values become closer to 0 so the ARIMA models are more accurate for forecasting. The second approach is to check the diagnostic of the residuals, and the purpose of this process is to check the estimated parameters (Cryer and Chan, 2008, p. 238). The process contains checking the standardized residuals, the ACF of the residuals and the Ljung-Box test. For the standardized residuals, the residuals will be plotted in order to assess the pattern of these residuals. On the other hand, the ACF of the residuals will be plotted in order to examine if there is any residual that is statistically significant. However, Table 3 contains the estimation of the model coefficients that are used to forecast the performance quality.

Table 3: Parameter estimation for ARIMA family.

	Model 1	**Model 2**	**Model 3**	**Model 4**
$AR_1(p_1)$	-0.54891**	0.10893*	0.09447*	0.88883***
$AR_2(p_2)$	-0.62864***	---	0.43099**	---
$MA_1(q_1)$	---	---	---	-0.60429**
$SAR_1(D_1)$	---	---	-0.70447***	-0.72249***
Intercept	---	0.01479***	0.93483***	0.01975***
D_{P02}	0.00630	-0.00203	-0.00654	0.00036
D_{P03}	-0.01262	0.00719	-0.00658	-0.00202
D_{P04}	-0.01039	0.00356	-0.02731***	0.00401
D_{P05}	-0.00538	0.00004	-0.02427***	0.00830**
D_{P06}	0.00587	-0.00516	0.01065	-0.00485
D_{P07}	0.00044	-0.00424	-0.01366	0.00406
D_{P08}	-0.03821***	0.00049	-0.07141***	0.01820***
D_{P09}	-0.06277***	0.02103***	-0.08180***	0.00938**
D_{P10}	-0.06045***	0.01093*	-0.08476***	0.02972***
D_{P11}	-0.01839*	0.00531	-0.07177***	0.01979***
D_{P12}	-0.02614**	0.00225	-0.05445***	0.01542***
D_{P13}	-0.00167	0.00263	-0.01524*	0.00095
RMSE	0.00788	0.00540	0.01282	0.00650

Model 1: ARIMAX (2,1,0). Model 2: ARIMAX (1,0,0). Model 3: SARIMAX (2,0,0)×(1,0,0). Model 4: SARIMAX (1,0,1) ×(1,0,0).
*, **, *** indicates significance at 0.05, 0.01, and 0.001 respectively.

4 Results and Discussion

4.1 Change Assessment

The investigation of the change in the performance quality of SWT was processed through two assessments. The first assessment focussed on SWT itself, through the three periods divided on the basis of performance indicators, PPM and CaSL, and route types, mainline and suburban routes, as provided in Table 4. The results of this assessment indicated that there was a continuous and statistically significant reduction of the performance quality of both routes of SWT through the three periods. Precisely, the proportion of trains arrived on-time reduced continually through the three periods. Similarly, the proportion of trains cancelled or significantly late escalated through the three periods.

Table 4: A summary of the individual and cross comparisons for both routes.

Comparison status	Measurement differences			
	Mainline		Suburban	
	PPM	CaSL	PPM	CaSL
Individual comparison (SWT itself)				
Case (1)	0.0268***	-0.0077***	0.0240***	-0.0084***
Case (2)	0.0606***	-0.0112***	0.0377***	-0.0116***
Case (3)	0.0874***	-0.0189***	0.0617***	-0.0200***
Proportion comparison (SWT with GTR)				
Case (4)	0.0295***	-0.0107***	0.0188***	-0.0090***
Case (5)	0.0372***	-0.0133***	0.0450***	-0.0197***
Case (6)	0.0546***	-0.0333***	0.0741***	-0.0375***
Proportion comparison (SWT with SE)				
Case (4)	0.0348***	-0.0101***	0.0188***	-0.0120***
Case (5)	0.0014**	0.0004	0.0037***	-0.0027***
Case (6)	-0.0239***	0.0034***	-0.0126***	0.0001
Change comparison (SWT with GTR)				
Case (7)	-0.0084	0.0025	-0.0285**	0.0112*
Case (8)	-0.0210**	0.0219*	-0.0281**	0.0178*
Change comparison (SWT with SE)				
Case (7)	0.0334***	0.0112	0.0160*	-0.0098*
Case (8)	0.0256	0.0178	0.0153	-0.0027

Cases: Cases (1): Prior-alliance minus alliance periods. Case (2): Alliance minus post-alliance periods. Case (3): Prior-alliance minus post-alliance periods. Case (4): Prior-alliance period for SWT minus prior-alliance period for other TOCs. Case (5): Alliance period for SWT minus alliance period for other TOCs. Case (6): Post-alliance period for SWT minus post-alliance period for other TOCs. Case (7): Prior- to alliance periods for SWT minus similar change for other TOCs. Case (8): Alliance to post-alliance periods for SWT minus similar change for other TOCs.
*, **, *** indicates significance at 0.05, 0.01, and 0.001 respectively.

The second assessment was based on comparing the change in the performance quality of SWT with GTR and SE through the three periods, and the comparisons were divided similarly to the individual assessment in terms of performance indicators and route types, as shown in Table 4. For SWT and GTR comparison, SWT performed more effectively in terms

of the PPM and CaSL measurements compared to GTR in the three periods. This means that SWT had a statistically significant higher proportion of trains arrived on-time and a lower proportion of trains cancelled or significantly late compared to GTR. However, when the change in the PPM and CaSL measurements is considered, both TOCs had statistically similar changes in the performance quality from the pre-alliance to the alliance periods for mainline services, whilst there was a statistical significant difference for suburban services. Although there are a decrease in the proportion of trains arrived on-time and an increase in the proportion of trains cancelled or significantly late for SWT, these changes are significantly higher for GTR. In contrast, there is a statistical significant difference in the PPM and CaSL measurements between SWT and GTR from the alliance to post-alliance periods for both routes. Although there is a deterioration in the measurements of the performance quality of SWT, the change in these measurements is significantly higher for GTR.

For SWT and SE comparisons, the performance quality of both routes of SWT was statistically better in terms of the PPM measurements compared with SE in the prior- and during the alliance periods, but SE performed better post the alliance period. For the CaSL indicator, SWT performed more effectively for both routes during the prior-alliance period compared to SE. For the alliance period, both TOCs had a similar performance in the mainline route while SWT performed better for the suburban route. For the post-alliance period, SE had a better performance quality in the mainline route while there is a similarity in the performance quality for both TOCs in the suburban route. With regard to the changes in performance quality, both routes of SWT faced significant adverse changes in the PPM measurements from the prior-alliance to alliance periods compared to the changes in SE's performance, whilst the change in CaSL was also adverse for suburban services. In contrast, there are no statistically significant differences in the changes in the measurements of PPM and CaSL for both TOCs in the mainline and suburban routes from the alliance to post-alliance periods.

4.2 The Effect Analysis

The effect analysis aims to assess the change in the PPM and CaSL indicators with respect to several factors. With regards to the PPM indicator, Table 5 contains summary results of the regression model of the PPM indicator. According to these results, the impact of the average age of the rolling stock on the PPM measurements is not statistically significant. This is expected, if the rolling stock is well maintained, its age should not affect the measurements of PPM. For the comparison between TOCs, the difference in the PPM measurements between SWT and SE is not statistically significant while the difference between SWT and GTR is statistically significant. Moreover, the odds ratio of the PPM indicator for GTR is 36% lower than SWT. This could be explained as the rail infrastructure of GTR is not reliable, and the timetable is not suitable for the peak services, and these reasons could affect train operation (Gibb, 2016). For the route type comparison, the difference in the PPM measurements between the mainline and suburban routes is statistically significant. Precisely, the odds ratio of the PPM indicator for the suburban route is higher by 20.70% compared to the mainline route. For the comparison between different organisation forms, the difference in the PPM measurements between the virtually integrated and separated organisation is not statistically significant. This means that there is no change in the performance quality measured by the PPM indicator between the virtually integrated and vertically separated organisation. For the financial year comparison, the difference in the PPM measurements between the financial years is statistically significant, except for 2011-12 and 2012-13 where

the difference is not statistically significant. The results indicates that the odds ratios of 2013-14, 2014-15, 2015-16, 2016-17 and 2017-18 are lower by 22.29%, 30.38%, 40.36%, 54.50% and 48.66% respectively compared to 2010-11. For the comparison between the financial periods, there is a statistical difference between the first period and the other periods, except for the second period where the difference is not statistically significant. The worst financial periods in terms of the PPM measurements are the eighth, ninth, tenth, eleventh, twelfth periods (broadly October to February) where the odds ratios are 54.48%, 62.29%, 56.46% 46.98% and 41.97% lower respectively. This could be explained by the change in weather conditions in autumn (leaf fall) and winter (snow and frost) seasons when bad weather is most likely to affect performance.

Table 5: Regression results for the PPM and CaSL indicators.

Term	PPM model			CaSL model		
	β	**exp(β)**	**Odds ratio**	**β**	**exp(β)**	**Odds ratio**
Constant	2.6984***			-4.0655***		
Rolling stock age	0.0042	1.0043	0.42	-0.0089	0.9911	0.88
TOC						
GTR	-0.4524***	0.6361	36.39	0.6109***	1.8422	84.22
SE	-0.0625	0.9394	6.05	0.1309*	1.1399	13.98
Route type						
Suburban	0.1881***	1.2070	20.69	-0.0185	0.9816	-1.84
Organisation form						
VS	0.0643	1.0665	6.65	-0.1062	0.8992	-10.07
Financial year						
2011-12	0.0102	1.0104	1.03	0.02795	1.0283	2.83
2012-13	-0.0768	0.9260	7.39	0.08692	1.0908	9.08
2013-14	-0.2522***	0.7771	22.29	0.3011***	1.3514	35.13
2014-15	-0.3620***	0.6962	30.37	0.3434***	1.4098	40.97
2015-16	-0.5168***	0.5964	40.36	0.5445***	1.7238	72.38
2016-17	-0.7875***	0.4550	54.50	0.8383***	2.3126	131.26
2017-18	-0.6666***	0.5134	48.65	0.7011***	2.0160	101.59
Period						
P02	-0.0583	0.9433	5.66	0.0224	1.0227	2.27
P03	-0.1600**	0.8521	14.79	0.2303**	1.2590	25.90
P04	-0.2205***	0.8021	19.78	0.1809*	1.1983	19.83
P05	-0.1848**	0.8312	16.88	0.2010*	1.2227	22.27
P06	-0.1267*	0.8810	11.90	0.0917	1.0961	9.60
P07	-0.2565***	0.7737	22.62	0.1444	1.1554	15.53
P08	-0.7869***	0.4552	54.47	0.4311***	1.5391	53.90
P09	-0.9923***	0.3707	62.92	0.7120***	2.0381	103.81
P10	-0.8315***	0.4354	56.46	0.7489***	2.1147	111.47
P11	-0.6344***	0.5302	46.97	0.5236***	1.6882	68.82
P12	-0.5442***	0.5803	41.96	0.5000***	1.6488	64.88
P13	-0.2659***	0.7665	23.35	0.2486**	1.2823	28.22
R-squared		0.7237			0.5673	
Adj. R-squared		0.7126			0.5499	

*, **, *** indicates significance at 0.05, 0.01, and 0.001 respectively.

Besides the analysis of the PPM indicator, Table 5 contains an estimation of the coefficients of the regression model for the CaSL indicator. According to this table, the impact of the average age of rolling stock on the CaSL indicator is also not statistically significant. Regarding the comparison between TOCs, the difference in CaSL measurements between SWT and GTR is statistically significant and the odds ratio of the CaSL indicator for GTR is 84.22% higher than SWT, and this can be linked with the reasons that are stated for the PPM indicator. Similarly, the difference in the CaSL measurements between SWT and SE is statistically significant and SE has a higher odds ratio by almost 14% compared to SWT. This means that SWT performed more efficiently in terms of the CaSL indicator compared to GTR and SE from 2010-11 to 2017-18 financial years. For the route type comparison, there is no statistical difference in the CaSL measurements between the mainline and suburban routes. This means that there is no difference in the proportion of trains cancelled or significantly late between the mainline and suburban routes. With regard to the impact of different organisation forms on the CaSL indicator, the results show no statistical difference between the vertical separation and virtual integration. This means that there is no difference in the CaSL measurements due to the change in the railway organisation. For the financial years, the difference in the CaSL measurements between 2011-12 and 2012-13 with the reference financial year, 2010-11, is not statistically significant. This indicates that the change in the proportion of trains cancelled or significantly late in 2011-12 and 2012-13 is not statistically significant compared to 2010-11. In contrast, the results show that there is a statistical difference in the CaSL measurements in 2013-14, 2014-15, 2015-16, 2016-17 and 2017-18 compared to the reference year. This means that the odds ratios of the CaSL indicator for 2013-14, 2014-15, 2015-16, 2016-17 and 2017-18 are higher by 35.13%, 40.97%, 72.38%, 131.26% and 101.59% respectively than the reference year. In addition, 2016-17 can be observed as the worst financial year in terms of the CaSL indicator. In a similar way to the financial year comparison, the results have two indications for the period comparison. The first indication is that there is no statistical difference in the comparison of the second, sixth and seventh periods with the first periods, which is treated as a reference. The second indication is that the difference in the CaSL measurements in the third, fourth, fifth, seventh, eighth, ninth, tenth eleventh, twelfth and thirteenth periods compared to the first period is statistically significant. This indicates that the odds ratios of the CaSL indicator for the third, fourth, fifth, seventh, eighth, ninth, tenth eleventh, twelfth and thirteenth periods are higher by 25.9%, 19.83%, 22.27%, 53.90%, 103.81% 111.47% 68.82% 64.88% and 28.22% respectively compared to the reference period. Additionally, the tenth period can be considered as the worst period in terms of the CaSL indicator. The explanation of this effect follow the same reasons for bad weathers as described above.

4.3 Performance Quality Prediction

The prediction of the PPM and CaSL indicators of SWT is divided into two approaches. The first approach is predicting the PPM and CaSL measurements during the virtual integration based on the measurements of the prior-integration period. The second approach is that the measurements of the PPM and CaSL indicators during the virtual integration are used to predict the measurements post the integration period. The procedure of predicting the PPM and CaSL measurements is based on ARIMAX models to accommodate the effect of periods on the PPM and CaSL indicators, as discussed in Section 3.3. For the first approach, the models that are used to predict the PPM and CaSL measurements are ARIMAX (2,1,0) and ARIMAX (1,0,0) respectively. Figure 3 and Figure 4 show the observed and predicted measurements of the PPM and CaSL indicators during the virtual integration. According to

these figures, it can be seen that there is a reduction in the performance of the PPM and CaSL indicators in the eighth, ninth and tenth periods compared to the predicted values. In addition, as discussed in Section 4.2, the eighth, ninth and tenth periods can be observed as the worst periods that have significant deterioration in the PPM and CaSL indicators for all TOCs, including SWT. Having said that, the virtual integration did not contribute to mitigating the deterioration in the performance quality during that period.

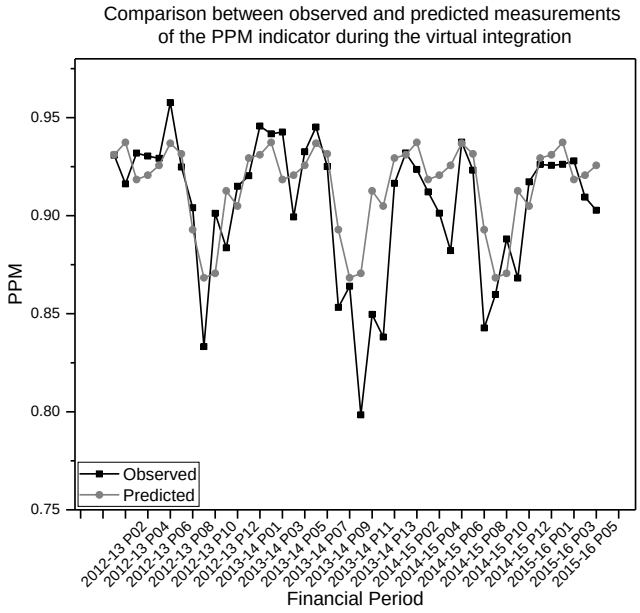


Figure 3: The observed and predicted measurements of the PPM indicator during the virtual integration for SWT.

The second approach is to predict the PPM and CaSL measurements in the post-integration period based on the integration period. The models that are used to predict the PPM and CaSL measurements are SARIMAX (2,0,0)×(1,0,0) and SARIMAX (1,0,1)×(1,0,0) respectively. Figure 5 shows the predicted and observed measurements of the PPM indicator post the integration period. It can be seen that there is a downtrend in the measurements of the PPM indicator in the post-integration period. In addition, several financial periods have a decrease in the proportion of trains arrived on-time compared to the predicted measurements. For the CaSL indicator, the predicted and observed values for this indicator are shown in Figure 6. According to this figure, there is a fluctuation in the observed measurements as predicted, but several periods have a higher proportion of trains cancelled or significantly late than predicted. Having said that, there is a significant deterioration in the performance quality of SWT post the integration period.

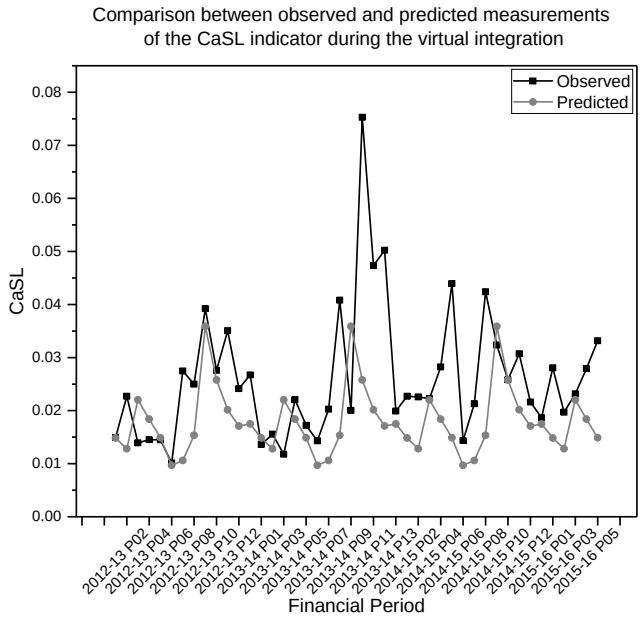


Figure 4: The observed and predicted measurements of the CaSL indicator during the virtual integration for SWT.

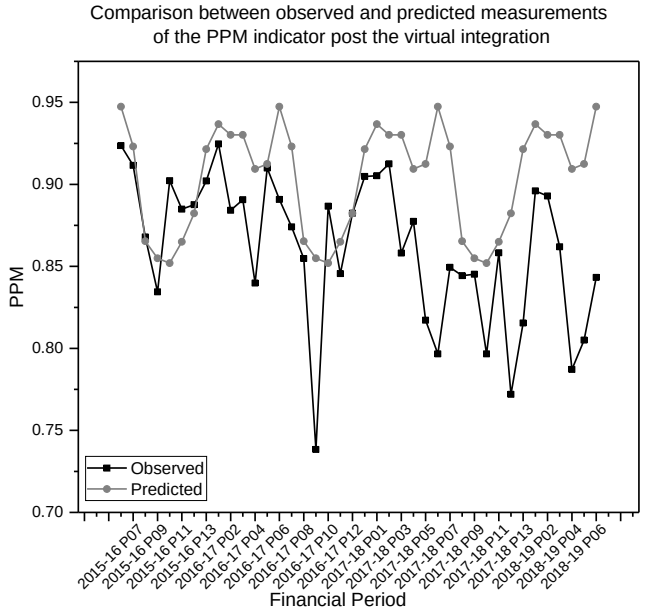


Figure 5: The observed and predicted measurements of the PPM indicator post the virtual integration for SWT.

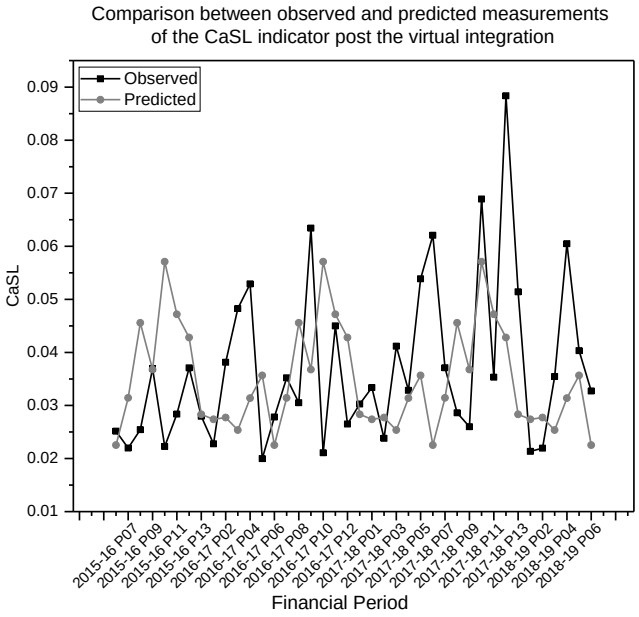


Figure 6: The observed and predicted measurements of the CaSL indicator post the virtual integration for SWT.

5 Conclusion

The assessment of the virtual integration started by evaluating the change in the performance quality of SWT. The results of this analysis pointed out the reduction in performance quality, as measured by the PPM and CaSL indicators for SWT. However, the change assessment was inconclusive, with SWT broadly performing better than GTR during the alliance period but performing worse than SE. The effect analysis was implemented in order to assess the effect of the virtual integration on the performance quality. The results of this analysis indicate that there is no evidence to support the effect of vertical integration on performance quality. The final analysis was predicting the performance quality of the SWT. The indication of this analysis is that the actual measurements of the performance quality are almost similar compared with the predicted measurements except for some periods that could be affected by other factors such as adverse weather. Overall, virtual integration does not seem to have had a significant impact on the performance quality of SWT.

Bibliography

Amaral, M. and Thiebaud, J.-C. (2015), Vertical separation in rail transport: How prices influence coordination, in '3rd Florence Conference on the Regulation of Infrastructures: Taking stock of current challenges', p. 50.

Armstrong, J. and Preston, J., (2017). 'Capacity utilisation and performance at railway stations'. *Journal of rail transport planning & management*, 7(3), pp.187-205.

Barron, A., Melo, P., Cohen, J. and Anderson, R., (2013). "Passenger-focused

management approach to measurement of train delay impacts", *Transportation Research Record: Journal of the Transportation Research Board*, 2351, 46–53.

Carey, M. (1998). 'Optimizing scheduled times, allowing for behavioural response', *Transportation Research Part B: Methodological*, 32(5), 329–342.

Carey, M. 1999. 'Ex ante heuristic measures of schedule reliability', *Transportation Research Part B: Methodological*, 33(7), 473–494.

Carey, M. and Carville, S. 2000) 'Testing schedule performance and reliability for train stations', *Journal of the Operational Research Society* 51(6), 666–682.

Cryer, J. D. and Chan, K.-S. (2008), *Time Series Analysis with Applications in R*, 2nd ed., Springer-Verlag. New York.

Drew, J. and Nash, C. (2011), 'Vertical separation of railway infrastructure: Does it always make sense?, *Institute of Transport Studies, University of Leeds* Working Paper 594.

Ferreira, L. and Higgins, A. (1996), 'Modeling reliability of train arrival times', *Journal of transportation engineering*, 122(6), 414–420.

Fox, J. (2015), *Applied regression analysis and generalized linear models*, 3rd ed., Sage Publications, London.

Gibb, C. (2016), 'Changes to improve the performance of the Southern network and train services, and restore passenger confidence [pdf]'. URL: https://goo.gl/Bv2gpx [Accessed date: 15 September 2017]

Goverde, R. M. (2005), 'Punctuality of railway operations and timetable stability analysis', PhD thesis, Technical University Delft.

Goverde, R. M. and Meng, L. (2011), 'Advanced monitoring and management information of railway operations', *Journal of Rail Transport Planning & Management* 1(2), 69–79.

Goverde, R. M., & Hansen, I. A. (2013). Performance indicators for railway timetables. In *2013 IEEE International Conference on Intelligent Rail Transportation Proceedings* (pp. 301-306). IEEE.

Harris, N. G. (1992), *Planning Passenger Railways: A Handbook*. Transport Publishing Co., UK.

Knowles, R. D. (2013), 'Railway franchising in Great Britain and effects of the 2008/09 economic recession', *Environment and Planning A*, 45(1), 197–216.

McNulty, R. (2011), 'Realising the potential of GB rail, report of the rail value for money study, summary report', *Department for Transport and ORR*.

Merkert, R. (2005), The restructuring and future of the British Rail system, Technical Report Working Paper 586, Leeds, UK. Copyright of the Institute of Transport Studies, University Of Leeds. URL: http://eprints.whiterose.ac.uk/2288/

Merkert, R., Smith, A. S. and Nash, C. A. (2010), 'Benchmarking of train operating firms–a transaction cost efficiency analysis', *Transportation Planning and Technology*, 33(1), 35–53.

Mizutani, F., Smith, A., Nash, C. and Uranishi, S. (2015), 'Comparing the costs of vertical separation, integration, and intermediate organisational structures in European and East Asian railways', *Journal of Transport Economics and Policy (JTEP)*, 49(3), 496–515.

NR and ORR (2017), 'Mandate L3 AR 004: Review of New Performance Metrics. [pdf]' URL: https://orr.gov.uk/__data/assets/pdf_file/0005/25376/review-of-new-performance-metrics-2017-07-18.pdf [Accessed: 12 April 2019]

Olsson, N. O. and Haugland, H. (2004), 'Influencing factors on train punctuality results from some Norwegian studies', *Transport policy*, 11(4), 387–397.

ORR (2016), 'Passenger Freight Rail Performance: Quality and Methodology Report7. [pdf]'. URL: https://goo.gl/5oyhu1 [Accessed: 25 March 2017]

ORR (2017), The National Rail Trends (NRT) Data Portal. URL: https://dataportal.orr.gov.uk/ [Accessed: 25 July 2017]

Parbo, J., Nielsen, O. A. and Prato, C. G. (2016), 'Passenger perspectives in railway timetabling: a literature review', *Transport Reviews*, 36(4), 500–526.

Preston, J. (1996), 'The economics of British rail privatization: an assessment', *Transport Reviews*, 16(1), 1–21.

Preston, J. (2002) 'The Transaction Cost Economics of Railways'. *Trasporti Europei*, 20/21, 6 – 15.

Preston, J., Wall, G., Batley, R., Ib´a˜nez, J. and Shires, J. (2009), 'Impact of delays on passenger train services: evidence from Great Britain', *Transportation Research Record: Journal of the Transportation Research Board*, (2117), 14–23.

Thompson, L. S. (2013), 'Recent developments in rail transportation services'. Organisation for Economic Co-operation and Development (OECD). URL: http://www.oecd.org/daf/competition/Rail-transportation-Services-2013.pdf [Accessed: 20 July 2017]

Veiseth, M., Olsson, N. and Saetermo, I. (2007), 'Infrastructures influence on rail punctuality', *WIT Transactions on the Built Environment* 96.

Vromans, M. (2005), Reliability of Railway Systems, PhD thesis, Erasmus University Rotterdam, Rotterdam School of Management. URL: http://hdl.handle.net/1765/6773 [Accessed: 20 July 2017]

Washington, S. P., Karlaftis, M. G. and Mannering, F. (2003), *Statistical and econometric methods for transportation data analysis*, CRC press.

Wetzel, H. (2008), European Railway Deregulation: The Influence of Regulatory and Environmental Conditions on Efficiency, Working Paper Series in Economics 86, University of Lneburg, Institute of Economics. URL: https://ideas.repec.org/p/lue/wpaper/86.html

Xia, Y., Van Ommeren, J. N., Rietveld, P. and Verhagen, W. (2013), 'Railway infrastructure disturbances and train operator performance: The role of weather', *Transportation research part D: transport and environment*, 18, 97–102.

Yaghini, M., Khoshraftar, M. M. and Seyedabadi, M. (2013), 'Railway passenger train delay prediction via neural network model', *Journal of advanced transportation* 47(3), 355–368.

Yuan, J. (2006), Stochastic modelling of train delays and delay propagation in stations, PhD thesis, Technical University Delft.

Yuan, J. and Hansen, I. A. (2007), 'Optimizing capacity utilization of stations by estimating knock-on train delays', *Transportation Research Part B: Methodological*, 41(2), 202–217.

Yuan, J. (2008), Statistical analysis of train delays, in I. A. Hansen and J. Pachl, eds, 'Railway Timetable Traffic: Analysis - Modelling - Simulation', Hamburg: Eurailpress, chapter 10, pp. 170–181.

Improving the Trade-Offs Between Network Availability and Accessibility

John Armstrong [a,b,1], John Preston [a], Tolga Bektas [c]
[a] Transportation Research Group, University of Southampton
Burgess Road, Southampton SO16 7QF, UK
[1] E-mail: j.armstrong@soton.ac.uk, Phone: +44 (0) 23 8059 9575
[b] Jacobs, Elms House, 43 Brook Green, London W6 7EF, UK
[c] University of Liverpool Management School, University of Liverpool
Chatham Street, Liverpool L69 7ZH, UK

Abstract
Passenger and freight traffic growth on Britain's railways has led to increased needs for maintenance, renewal and enhancement of the national railway network, and reduced opportunities for access to the network to conduct these engineering activities without disrupting operations. As a result, the costs of compensation to operators for service disruption and revenue loss have been increasing in line with traffic levels. There tends to be a trade-off between the cost efficiency of engineering activities and the compensation costs for the operational disruption caused, since longer track possessions are typically more efficient, but also more disruptive, reducing network availability for operations. There is thus a need to reduce and, ideally, minimise the total costs of engineering activities and compensation for the disruption caused. The current possession planning process does not actively aim to minimise service disruption and compensation costs, much less the combined engineering and compensation costs. This paper describes the detailed review of the current possession planning process, including data availability and needs, that is being undertaken. It also outlines a methodology that will be applied in order to (i) amend the current possession planning process to reduce its disruptive impact and compensation costs, thus increasing network availability for operations, and (ii) identify data requirements to enable the assessment of duration, engineering costs and timetable impacts/compensation costs associated with alternative possession strategies, and apply these in combination with scheduling techniques to reduce and, ideally, minimise combined engineering and compensation costs, and the detrimental impacts on railway users and funders.

Keywords
Railways, Maintenance and Renewals, Engineering Access, Network Availability, Possession Planning, Costs

Paper Type
Type B: Professional paper (i.e. applied research)

1 Introduction

In Britain, as elsewhere, growth in railway passenger and freight traffic in recent decades, while welcome, has presented the railway industry with various operational, management and performance challenges. Among these is the increased need for network access for

infrastructure maintenance, renewals and other engineering activities as a result of greater traffic volumes and infrastructure wear and tear, combined with reduced opportunities to carry out these necessary works, as user expectations move towards 24/7 network availability for travel and transport, and the network is more intensively used. Further compromises are required between the efficiency with which engineering activities can be conducted (typically maximised by lengthy engineering 'possessions' of the track, or 'blockades'), and network availability to users (typically maximised by short, overnight possessions).

This paper reviews the current situation regarding engineering access planning in Britain and identifies needs, opportunities and means for improvement. Following this introduction, the problem statement and objectives of the work are set out, and relevant literature is briefly reviewed. Our intended methodology is then summarised, including data sources and needs. Finally, the practical relevance of the work is described, followed by a list of references.

2 Problem Statement and Objectives

In common with some other countries, railway traffic levels in Britain have increased dramatically over the past 25 or so years, following decades of decline. This otherwise welcome growth in traffic, as well as presenting capacity challenges, results in increased infrastructure wear and tear and associated maintenance and renewals (M&R) needs, while also reducing opportunities for access to the infrastructure for M&R and enhancement purposes. As summarised by Andrew McNaughton (2018), the strategic technical adviser to HS2 Ltd., the company responsible for building High Speed Two, the second phase of Britain's high-speed railway network,

> *the challenge now facing the UK is how to transform the capacity and efficiency of our network to support future growth within the available financial resources without creating wholesale disruption for millions of passengers. The UK will need a variety of solutions that provide greater capacity, improved reliability and better value for both passengers and taxpayers.*

This challenge statement mirrors the strategic goals for Britain's railways, sometimes summarised as the '4Cs', as explained by the Technical Strategy Advisory Group (TSAG, 2009):

1) Reduced Costs
2) Increased Capacity
3) Improved Customer satisfaction
4) Reduced Carbon emissions

As well as being essential for the maintenance, renewal and enhancement of the network, engineering access to the railway infrastructure affects at least three of the 4Cs: it increases costs (via compensation to train operators for loss of network availability for operations, as well as directly-incurred engineering costs); it temporarily reduces capacity; and it can seriously affect the customer experience, since users may be subjected to service cancellations or extended journey times via diversionary routes, including, in some cases, the use (and further inconvenience) of substitute road transport. While M&R and network

enhancements are necessary to maintain and increase network capacity, it is clearly in the interests of the railway industry and its users to reduce the costs and temporary capacity loss associated with these works, and to reduce their impact on users.

In Britain, train operators are compensated for the disruptive effects of engineering possessions of the infrastructure, and their potential long-term impact on user demand and revenue, by means of the Schedule 4 Compensation System (S4CS; Network Rail, 2018a), as set out in Schedule 4 of operators' Track Access Contracts (TACs) with Network Rail, the infrastructure manager (IM) of Britain's heavy rail network. There are three main components of the S4CS payments and calculations, determined by means of a comparison between the normal and possession-affected train timetables: cancellations of scheduled stops; extended journey times; and changes to operating costs. The first two directly affect and potentially deter users, and usually result in payments from the IM to operators; the third affects the operators only, and usually results in a 'negative payment' from the IM to the operators, set against the first two elements, since the total number of train km operated is typically reduced as a result of full or part cancellations of trains, reducing operating costs. Other costs, such as the running of replacement bus services, are also considered in the compensation process.

The effects of increasing traffic levels on S4CS costs can be seen from Figures 1 and 2, based respectively on data produced by the Office of Rail and Road (ORR, 2018, Table 12.13) and Network Rail (2018b): Figure 1 shows annual passenger train km (excluding Heathrow Express (HEx) airport train services) between 2011/12 and 2016/17 inclusive, while Figure 2 shows the annual Schedule 4 payments made by Network Rail during the corresponding time period. It can be seen that, despite declines in both from 2015-16 to 2016-17, (i) the annual S4CS payments are large, at approximately £300m per annum for the most recent data shown (although this constitutes only approximately 2.7% of total annual expenditure (Network Rail, 2018c)), and (ii) their pattern is similar to that of the annual passenger train km values.

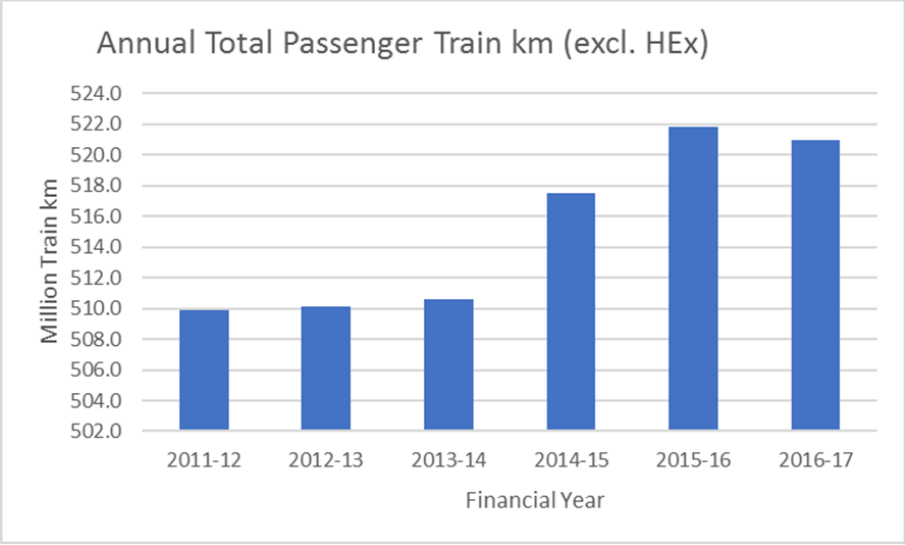


Figure 1: Annual Passenger Train km, 2011-2017
HEx = Heathrow Express

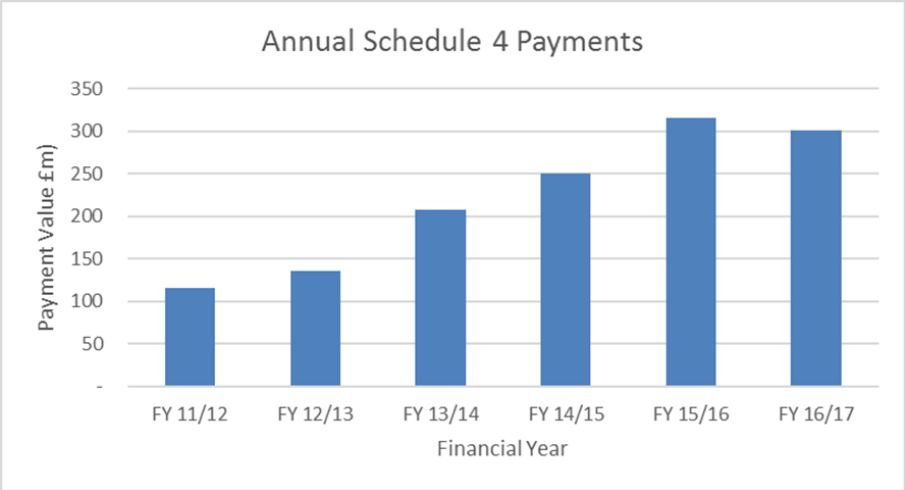


Figure 2: Nominal Annual Schedule 4 Payments, 2011-2017

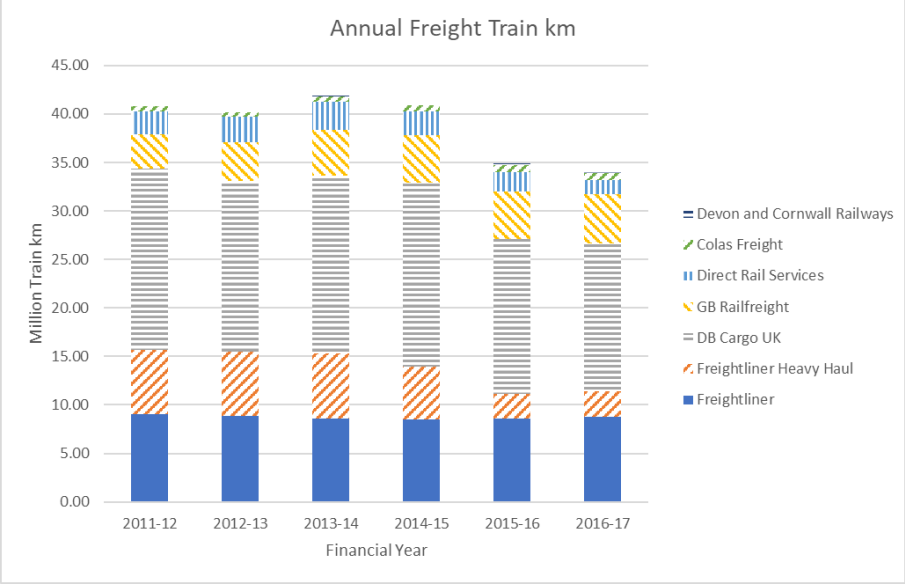


Figure 3: Annual Freight Train km, 2011-2017

Annual freight train km during the same period are shown in Figure 3. It can be seen that freight traffic, as well as being an order of magnitude smaller in volume than passenger train km, has declined in recent years, due mainly to a reduction in coal traffic as a result of the de-commissioning of coal-fired power stations, which has particularly affected DB Cargo UK and Freightliner Heavy Haul. Freightliner intermodal traffic has increased

slightly in recent years, and these services are relatively time-critical and tend to use busy passenger routes at night and weekends, and are thus vulnerable to engineering-related disruption.

If and when overall traffic growth is resumed (a desirable outcome for the railway industry, and for society, if modal shift from more polluting and less safe forms of transport is to be achieved), engineering-related compensation costs for both passenger and time-critical freight services are likely to increase further, in the absence of measures to prevent this. There is thus a need for improved planning and scheduling of M&R and other work requiring access to the infrastructure, to reduce disruption to users and the associated S4CS costs. However, reducing the duration of individual track possessions may also have an effect on the efficiency with which engineering activities can be undertaken, since a higher proportion of the time available will typically be required for the processes of taking possession of the infrastructure and subsequently restoring it to operational use, reducing the proportion of productive time on site. Consideration therefore also needs to be given to the trade-off between network availability for operations and the productivity with which engineering activities can be undertaken. This issue also presents challenges in terms of the availability of (i) cost and construction programme and duration data for alternative possession approaches, and (ii) the associated amended timetable data upon which the S4CS calculations are based.

The work described here thus has two main objectives:

1) Improve the planning and scheduling of engineering possessions to reduce (i) their impact on network availability for operations and (ii) the resulting S4CS payments, including the scheduling in parallel of activities affecting the same sections of the network, where possible
2) Develop means of including the timescales, costs and timetable impacts of alternative possession approaches, and include these in the planning and scheduling process, with a view to reducing, and ideally minimising, the combined engineering and compensation costs, and thus maximising the overall benefit:cost ratio of civil engineering activities and the necessary associated possessions and network availability restrictions

3 Review

The then-current approach to engineering access planning on Britain's railways was reviewed by Armstrong et al. (2015), who noted that the available measures of network availability for operations were being calculated retrospectively to reflect the effects of engineering possessions, rather than being used pro-actively, to assess, review and reduce the impact of planned possessions. However, they also observed that the Industry Access Programme (IAP) then being put in place had considerable potential to remedy this issue. A subsequent report by Europe Economics (2017) confirmed the 'lag variable' nature of the network availability calculations, and also their complexity and inflexibility (e.g. the calculations cannot be performed at a disaggregate level for individual network routes, despite the fact that possession planning takes place at this level, and responsibility for network operation, maintenance and performance is being devolved by the IM to individual routes). The report also observed that implementation of the IAP appeared to have stalled. Network Rail confirmed this, and indicated that their Transformation and Efficiency Team (TET) is continuing to work in this area, and is receptive to useful input and contributions.

The Europe Economics report acknowledges that possession planning is a complex

optimisation process, and it confirms that the current approach is unlikely to produce an optimal outcome, which is a cause of particular concern in the context of diminishing opportunities for engineering access and increasing concerns about M&R costs. The report notes that the possession planning system is based upon staff experience (and is thus potentially vulnerable to staff turnover) rather than possession planning tools, and that route-based possession planning tends to be undertaken in isolation, rather than considering potential synergies with work being undertaken elsewhere on the network. This increases the likelihood of sub-optimal outcomes, and (p9) "may lead to the overall volume of possessions being higher than it needs to be", whereas

> *reducing the number of possessions should be driven by the Schedule 4 [S4] incentive, whereby planners are incentivised to optimise the use of possessions (e.g. by using them for more than one type of work where this is efficient) in order to reduce the number of possessions and resulting S4 payments.*

The report considers alternative measures of network availability, including route-based metrics and comparisons of normal and possession-affected timetables (already the basis of the Schedule 4 calculations), and measures of possession efficiency, to ensure that possessions are used productively. However, as noted above, engineering efficiency tends to be maximised in longer possessions, and the effects on network availability also need to be considered. Ideally, and as also proposed by Li et al. (2013), such an improved metric should consider both factors by including both the engineering costs and the Schedule 4 costs (as a measure of the operational disruption caused) for individual pieces of engineering work and overall, for individual routes and, ultimately, for the network as a whole.

The chances of achieving optimal outcomes are not necessarily improved by the fact that the process is based upon negotiation and compromise between Network Rail, as IM, and (sometimes multiple) train operators, as well as being generally undertaken on a route-by-route basis, as noted above, without usually considering wider network effects. Considerable work in this area has been done elsewhere though, dating back at least to the 1960s (e.g. Wagner et al., 1964), and including various approaches to the solution of the Preventive Maintenance Schedule Problem (PMSP), and the combination or clustering of maintenance tasks, as described by Peng and Ouyang (2014).

Li (2017) presents a broad overview of railway maintenance scheduling, and proposes two decision support systems (DSSs). The first DSS includes five phases: data collection; technical optimisation to identify minimum maintenance requirements; economic optimisation to minimise the cost of the identified minimal maintenance requirements; constrained optimisation to include the effects of operational conditions and enable input parameter adjustment; and, finally, evaluation. The second DSS takes account of life-cycle costs in planning and evaluating possession strategies. Both were found to have considerable potential for reducing total infrastructure-related costs, while maintaining infrastructure quality, and these approaches appear to have considerable potential for application in Britain, adapted as required to local conditions, and subject to the availability of the necessary data.

4 Methodology

The planned methodology builds and improves upon the current approach to possession

planning in Britain, drawing upon international research and practice, while taking account of and complementing the work done for IAP and subsequently by TET. It uses S4CS/network availability measures to plan possessions pro-actively for reduced impact, rather than using them solely as retrospective measures and means of compensation for their disruptive effects. It includes four main elements and stages of work:

1) A review of the existing processes, planned improvements (as applicable, including outputs from the IAP and work being undertaken by the TET) and available data. This includes the potential for and possible means of extending datasets to include alternative possession and timetable options, and/or opportunities to relax these requirements and adopt a simplified approach (avoiding, for example, the need for the production of detailed timetable data to assess the S4CS costs associated with alternative engineering and possession strategies)
2) The development of a simplified network model for possession planning purposes and use in stages 3 and 4, identifying the required extent of route closures corresponding to possession locations (and thus the potential for the scheduling of simultaneous possessions on those route sections), and available diversionary routes, taking account of constraints such as electrification, loading gauge and route availability for different axle load categories
3) The development of a method and tool for improving the scheduling of possessions based upon current engineering workbank data, with a view to reducing S4CS payments and the associated disruption as a first step in the improvement process – this will include consideration of the simultaneous scheduling of possessions on affected route sections where possible. The results obtained will be compared with those produced by the current possession planning system, to assess the scale of potential benefits and efficiency gains
4) The extension of the stage 3 methodology on the basis of alternative possession and timetable scenarios, employing extended/simplified programme, cost and timetable datasets, using these to reduce and, ideally, minimise the total engineering and compensation costs

This methodology will be developed, applied, reviewed and refined as necessary in cooperation and collaboration with Network Rail staff.

5 Data types and sources

Three main categories of data are required:

1) Historic and planned possessions data: dates (and constraints/interdependencies between different elements of work), durations, locations and costs, and the associated timetable impacts in terms of train diversions and full/part cancellations of services, and thus the effects on train km operated, and operating costs
2) Network data: information needed to generate a representation of the national network sufficient for possession planning purposes, including electrification status, loading gauge, route availability (by axle load) for freight, identification of ‘isolatable’ route sections within which multiple pieces of work can be undertaken within a single possession, and potential diversionary routes (most of this data is already in the public domain and thus readily available)

3) Estimates of the durations and costs associated with alternative construction approaches, and the associated variations in their impacts upon normal train timetables – this data is likely to be the most difficult to obtain

The source for most, if not all of the data is Network Rail, in its capacity as IM. Some of the data (e.g. network characteristics and constraints) is freely available online, but the remainder will be obtained by discussion with Network Rail staff. (Note: since the abstract for this paper was submitted, less progress has been made than was originally anticipated in obtaining data from and agreeing methods and objectives with Network Rail; the authors anticipate being able to provide further updates in these respects at RailNorrkoping2019.) It may also be useful to employ actual, historic cost and timetable data for comparison with calculated alternatives, to facilitate the development, testing and demonstration of the planned approach and tools. In some cases, cost (as indicated above and also noted by Li and Roberti, 2017), duration and timetable data for alternative construction approaches may not be readily available, and it may therefore be necessary to generate artificial, realistic datasets for the purposes of developing, testing and demonstrating initial models and tools. This would build upon work previously done by the authors to produce estimates of future S4CS costs (Armstrong et al., 2015), as shown in Figures 4 and 5. As can be seen in Figure 4, the S4CS calculation process entails the comparison of two timetables, the 'Corresponding', or normal, timetable (T1) and the 'Applicable', or possession-affected timetable (T2), and the calculation for each affected train service group (SG) of the changes in the number of stops at the SG's specified monitoring points (MPs) in each direction of travel (the MPs are weighted by their historic proportions of alighting passengers, which vary by direction). Changes in journey times and operating distances are also calculated.

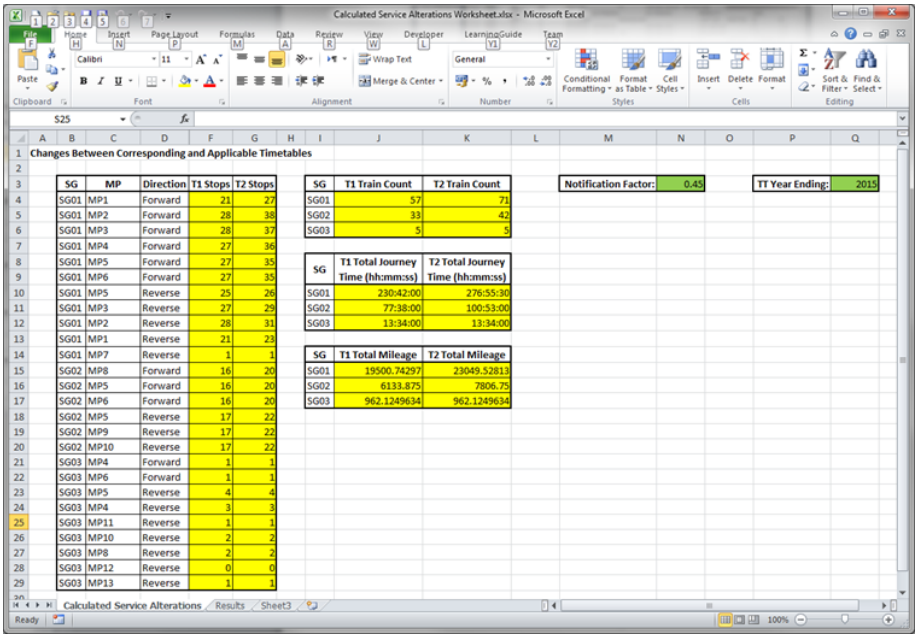


Figure 4: Train Service Alterations Worksheet

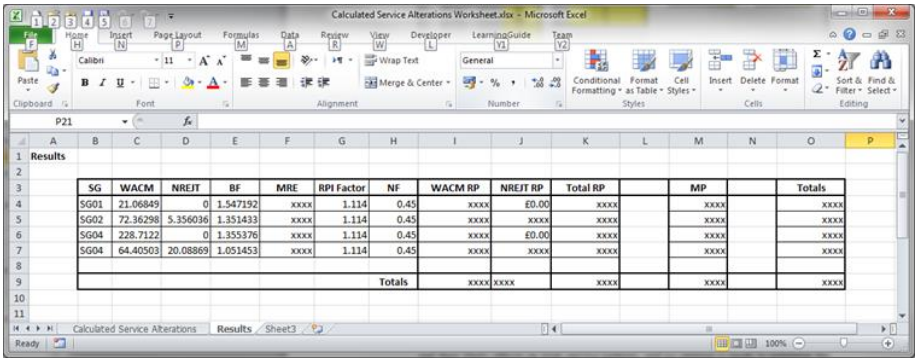


Figure 5: Summary Results Worksheet

In the results sheet shown in Figure 5, for each service group, the calculated weighted average cancellation minutes (WACM) and extended journey times due to Network Rail activity (NREJT) are shown. These are combined with a busyness factor (BF), marginal revenue effect (MRE) value, Retail Price Index (RPI) measure of inflation, and a notification factor (NF, reflecting the length of notice given by Network Rail to the operator of the planned disruption) to calculate the revenue payments (RPs) due to WACM and NREJT, and the total RP and the mileage payment (MP, usually negative, as noted above), and the resulting overall total payment (the calculation process is described in more detail in Armstrong et al., 2015).

6 Scientific and Practical Relevance of Planned Work

The focus of this work and professional paper is primarily on the practical application of existing knowledge in an industry context, but it does have some potential scientific relevance in terms of the extension and modification of techniques to meet the needs of the railway engineering and possession planning environment in Britain.

The work has considerable practical relevance in terms of its potential to enable and deliver improved planning of engineering activities and track possessions to reduce their impact on railway users and their overall costs to the industry. This is consistent with the objectives of the Rail Safety and Standards Board (RSSB, 2014) Operational Philosophy for Britain's railways, one of whose requirements is for the 24/7 operation of passenger and freight trains. Meeting this requirement will necessarily "significantly reduce access to the network for maintenance and renewal of assets", requiring improved operational flexibility, including bi-directional operation and the use of diversionary routes, and efficient access arrangements, and the work described in this paper should make a useful contribution to the achievement of that goal.

References

Armstrong, J., Preston, J and Hood, I., 2015. "Possession Compensation and Network Availability on Britain's Railways", In: *Proceedings of the 6th International Seminar on Railway Operations Modelling and Analysis (RailTokyo2015)*, Tokyo, Japan.

Europe Economics, 2017. “Availability Output Measure Review” [online]. Available from http://orr.gov.uk/__data/assets/pdf_file/0015/25305/availability-output-measure-review.pdf [Accessed 23 January 2019].

Li, R., Landex, A., Nielsen, O.A. and Madsen, S.N., 2013. “The potential cost from passengers and how it impacts railway maintenance and renewal decisions”, In: *Proceedings from the Annual Transport Conference (Trafikdage)*, Aalborg University, Denmark.

Li, R., 2017. PhD Thesis: “Phase-based Planning for Railway Infrastructure Projects” [online]. Available from http://orbit.dtu.dk/files/143903449/RuiLi_PhDThesis.pdf [Accessed 3 October 2018].

Li, R. and Roberti, R., 2017. “Optimal Scheduling of Railway Track Possessions in Large-Scale Projects with Multiple Construction Works”, *Journal of Construction Engineering and Management*, Vol. 143(6): 04017007.

McNaughton, A., 2018. “Future-proofing our railways: Smart thinking and innovative technology will deliver a service fit for the 21st century” [online]. Available from https://prospectmagazine.co.uk/sponsored/future-proofing-our-railways [Accessed 28 September 2018].

Network Rail, 2018a. “Payments for planned disruption on the railway” [online]. See https://www.networkrail.co.uk/industry-commercial-partners/information-operating-companies/payments-for-planned-disruption-on-the-railway/ [Accessed 1 October 2018].

Network Rail, 2018b, “Payments for planned disruption on the railway made under schedule 4” [online]. Available from https://cdn.networkrail.co.uk/wp-content/uploads/2017/12/Payments-for-planned-disruption-on-the-railway-made-under-schedule-4-and-the-corresponding-ACS.xlsx [Accessed 9 October 2018].

Network Rail, 2018c, “Network Rail expenditure in 2016/17” [online]. Available from https://cdn.networkrail.co.uk/wp-content/uploads/2017/12/Annual-expenditure-2016-17.pdf [Accessed 9 October 2018].

ORR, 2018, “Passenger Rail Usage 2017-18 Q4 Statistical Release” [online]. Available from http://orr.gov.uk/__data/assets/pdf_file/0014/28013/passenger-rail-usage-2017-18-q4.pdf [Accessed 9 October 2018].

Peng, F. and Ouyang, Y., 2014. “Optimal Clustering of Railroad Track Maintenance Jobs”, *Computer-Aided Civil and Infrastructure Engineering*, vol. 29, pp. 235-247.

RSSB, 2014. *Operational Philosophy for the GB Mainline Railway* [online]. Available from https://www.rssb.co.uk/research-development-and-innovation/research-reports-catalogue/pb022011 [Accessed 3 October 2018].

TSAG, 2009. “Technology route-mapping to support the planning for Rail’s 30 year vision” [online]. Available from https://www.sparkrail.org/_layouts/15/Rssb.Spark/Attachments.ashx?Id=75NEMTS3ZVHP-8-2954 [Accessed 15 January 2019].

Wagner, H., Giglio, R. and Glaser, R., 1964, “Preventive Maintenance Scheduling by Mathematical Programming”, *Management Science*, vol. 10(2): pp316-334.

Timetable Rules and Strategies for Double Track Maintenance Work

Magnus Backman[a], Emma Solinen [ab]
[a] Trafikverket, SE-172 90 Sundbyberg, Sweden
[b] Department of Science and Technology, Linköping University, SE-601 74 Norrköping, Sweden
E-mail: magnus.backman@trafikverket.se, Phone: +46101232221

Abstract
When large maintenance work is done at a double track line, it is often possible to have one of the two tracks open for traffic. The traffic then run with single track operation which heavily affects the capacity and need to be planned in an early stage, before the yearly timetable is finalized. Today, in Sweden, there are some difficulties when planning for maintenance works and how to adapt the reduced capacity in the timetable. Due to an increased demand for capacity and for better punctuality from train operators, there is a need for more well thought-out strategies for how to handle the capacity restriction and for how much robustness is needed in the timetable to preserve a certain quality.

In this paper, we present a study which assess strategies for double track maintenance work leading to single track operations. A simulation study is performed in which three different timetable strategies are tested and evaluated. The aim is to find strategies and timetable rules to better handle capacity reductions at double track lines so that trains can run with high quality even though there are maintenance works at the same time. In the paper we discuss the advantages and disadvantages with the three strategies and how they affect train slots, runtimes and punctuality.

Keywords
Railway timetabling, Robustness, Capacity reduction, Simulation, Punctuality

1 Introduction

The railway infrastructure is from time to time in need of an upgrade. For example, the tracks or the contact line need to be replaced to prevent it from break down and cause larger disruptions. When large maintenance work is done at a single track line, no traffic can use the line during the work. For a double track line it might be possible to still have one of the two tracks open for traffic. The traffic is then run with single track operation which heavily affects the capacity and need to be planned in an early stage, before the yearly timetable is finalized. Lidén (2015) presents a survey of problems and conducted research in the area of railway maintenance planning in which several problem areas are discussed. In previous research, e.g. Vansteenwegen et al. (2016) and Van Aken et al. (2017) models and algorithms are proposed to re-schedule trains in case of planned maintenance works.

However, the models tend to include complicated calculations and do not always include the robustness aspect. Until there is a complex system support to guide the timetable planners, there is a need for more well thought-out timetable strategies and suggestions of suitable headways and time supplements to preserve a certain quality.

Today, in Sweden, there are some practical difficulties when planning for maintenance works and how to adapt the reduced capacity in the timetable. There are no general guidelines for single track operations due to maintenance works in the official timetable rules presented in Trafikverket (2016). The Swedish implementation of the timetable planning tool Trainplan (Trapeze, 2019) used by Trafikverket does not support different infrastructure variants and most of the capacity restrictions have to be adapted manually. How much time supplement that is needed to handle single track operations is estimated by timetable planners and effects on headway times due to the signal system layout is often ignored. Traditionally, temporary single track operations have been handled with reduction of train paths and one single time supplement for trains passing the work section. This approach has shown to result in poor punctuality and due to an increased demand for capacity and for better punctuality from train operators, the capacity restriction and robustness needs to be studied further.

In this paper we present a study which assess strategies for double track maintenance work leading to single track operations. The aim with the study is to give better knowledge to Trafikverket in deciding how to maintain punctuality during maintenance work. We want to find strategies and timetable rules to better handle capacity reductions at double track lines so that trains can run with high quality even though there are maintenance works at the same time.

The outline of the paper is that in Section 2 the three most used timetable strategies are described together with the observed punctuality effect of today's timetable construction. Section 3 contains general rules that originates from the observations of previous timetable constructions. These rules has been developed in this study since we discovered that today's construction often resulted in infeasible timetables in practice. Section 4 describes the simulations study in which the three timetable strategies are evaluated with disturbances. Also how much time supplement is needed for each strategy to maintain punctuality is evaluated. In Section 5 we have a concluding discussion on the result of the simulation study and also on the advantages and disadvantages of the different timetable strategies.

2 Timetable Strategies for Handling Temporary Single Track Operations

Different timetable strategies have been used in Sweden to handle single track operations. The strategies can basically be grouped in to three different approaches; full re-scheduling, trains scheduled in groups and only time supplements. At first we analyse some single track operations and how they have been handled in the timetable are studied. Then the three strategies are described more in detail.

2.1 Effects of Todays' Timetable Construction

The studied examples are real-world examples where timetable planners have modified the timetables fully with respect to the temporary single track operation. These timetables showed that the planners had squeezed in as much trains as possible over the single track sections without violating the theoretical feasibility. This lead to the timetables becoming very sensitive for delays and the trains easily disturbed each other. In some cases even smaller disturbances would affect the operations for hours after they had occurred due to lack of recovery time in the timetable. Another common method used by the timetable planners was to add a few minutes extra time supplement at the track sections just before

the station where the single track operation started. The reason for this was to get the train to arrive to the single track section in exactly the right time, when a train in opposite direction left the single track. This extra time supplement did not add any real value to the timetable, it was just a way of puzzle the trains together. In fact, the time supplement instead often led to trains arriving too early to the already occupied single track section, causing even more disorder.

The study of the real-world timetables indicates that there is a need for restrictive timetable rules to prevent timetable planners to construct too optimistic timetables. In purpose of finding good timetable rules and to analyse how much additional time supplement needed to maintain a high quality, a simulation study is performed where three different strategies are evaluated.

2.2 Three Timetable Strategies for Evaluation

The three chosen strategies are common strategies for handling single track operations. All strategies have to some extent been used previously for planning single track operations, but in general, the result shows poor punctuality regardless of strategy. The three strategies are:

1) *Full re-scheduling of all trains to make a feasible timetable including the single track operation*

   With full re-scheduling of all trains every single track section needs to be planned with a separate timetable. For larger maintenance works that moves along the line this can give up to 10 different timetables over a year. This strategy requires a lot of planning resources but will give the dispatchers a conflict free timetable for every stage of the maintenance work.

2) *Trains scheduled in groups running in the same direction over a longer part of the line including the temporary single track*

   To avoid many timetable variants trains can be arranged in groups before the single track operation begins. First, a group of trains run in one direction and doesn't meet other trains until the whole group has passed the chosen section. Then a group of trains running in the other direction can pass the section. This strategy can be used to maximize capacity utilization and also to construct a timetable that covers more than one track section without scheduled meetings between single trains. This reduces the number of timetable variants needed when the maintenance work moves along the line.

3) *No re-scheduling other than adding time supplements for the trains passing the single track section*

   With no re-scheduling the only modification done is adding time supplements in the timetable. The trains can be in conflict in the timetable and it is up to the train dispatcher to solve them as they happen. This strategy requires only one timetable variant but it increases the amount of work for the train dispatchers.

3 General Rules

Due to previous experiences with poor punctuality regardless of strategy used, a set of general rules was developed even before the evaluating simulations. Real-world examples of timetables with temporary single track operations was studied and it was clear that characteristics of the signal system and driver behaviour often were ignored which made most of the timetables only theoretically feasible. Each area that generates need for extra timetable rules is presented below.

3.1 Time Supplements for Reduced Speed

The timetable planning tool used by Trafikverket does not support different infrastructure variants. This means that the reduced speed for trains passing a maintenance work has to be added manually as time supplement. The time supplements need to be calculated separately for each train category and then added for each train.

3.2 Construction with Start/Stop Supplements

A common method used in the studied timetables was to add a few minutes extra time supplement in the track sections just before the station where the single track operation started. The reason for this is to get the train to arrive to the single track section in exactly the right time, when a train in opposite direction left the single track. However, train drivers are not always aware of this time supplement. To adapt their speed to the planned timetable the train drivers have to use their experience of how long time this track sections usually takes to drive in full speed. In the majority of the cases, train drivers do not notice the deviation, run with full speed towards the single track section and ends up stopping before the entry signal. Since it takes more time for a train to accelerate from zero, than if it was running with a reduced speed, the planned runtime on the single track section is not enough and the train gets delayed. See how the delays are spreading in Figure 1.

This combination of timetabling and driver behaviour create a chain reaction that will last until there is a gap in the timetable wide enough that trains leaving the single track section do not affect trains entering this section.

To avoid this situation, time supplement should not be added on the last section before the temporary single track. Trains should instead, if necessary, have a planned scheduled stop at the border station of the single track. With this planned stop the timetable planning tool will automatically calculate the run time needed, included the extra time for acceleration from zero.

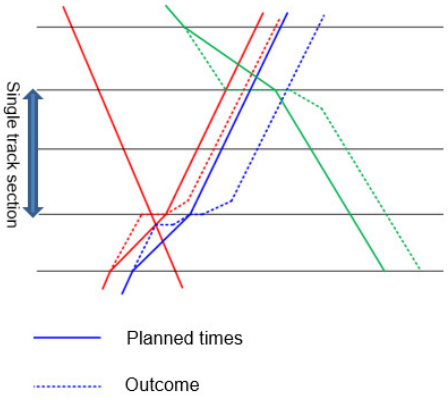


Figure 1: Illustration of how delays are spreading when trains have time supplement added before the single track section. Instead of slowing down they run with full speed and have to stop before the section.

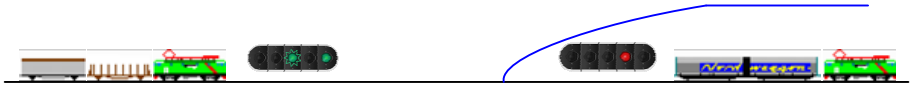


Figure 2: The train to the right has started to accelerate (blue curve) and the following train to the left get the signal aspect “expect stop” in the next signal.

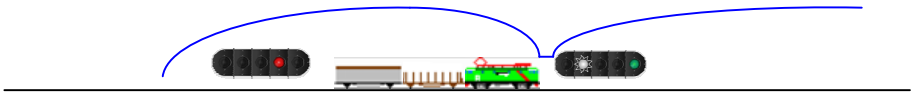


Figure 3: The first train has left the block section which makes the signal green, but the following train gets no update of the signal aspect and has to slow down to surveillance speed (blue curve).

3.3 Headway on the Single Track Section

All Swedish double track lines have full signalling in both directions, headways are not restricted by a reduced number of block signals even if trains run on the opposite track.
The automatic train control (ATC) system in Sweden gives the possible headway between trains (the distance between two trains following each other on the same infrastructure). The ATC system transfers signal information on certain points through balises. Balises are always present at the location of signals, but they can also be present before the signals to

transfer signal information from the next main signal. Contact with such repeating balise will then update the on-board computer with a new signal aspect. On smaller stations repeating balises are not common since these stations are primary used for maintenance works and operational train dispatching and not for everyday traffic. This will influence the headway if more than one train have to stop before entering the section with single track operation. When two or more trains have to queue up to wait for an oncoming train to leave the single track section the second train will start with signal aspect "expect stop" in the next signal, see Figure 2. Since there are few repeating balises, the change in signal aspect can´t be communicated to the ATC system between the signals. The train will therefore not get information on an update in the signal aspect, which will force the train to slow down to the surveillance speed of the block section before passing the next signal, see Figure 3.

This characteristics of ATC makes it impossible to maintain the standard headway used for trains on this line. The general rule is therefore that one minute must be added to the standard headway used for normal train operation.

3.4 Time Between Trains on Border Stations

In the studied timetables with single track operations many trains had the same timetable time for entering the single track section as oncoming trains had when leaving the section. This means that there is no margin time between the trains for the switch to change position and for the signal system to reverse block direction. To have enough time for such necessary actions at least one minute must be planned between trains in different direction leaving and entering the single track section.

4 Simulation Study

In purpose of finding good timetable rules and to analyse how much additional time supplement needed to maintain a high quality, a simulation study is performed. We use the microsimulation tool RailSys, (RMCon, 2019) commonly used in both industry and research to perform simulation studies. In the study a sequence of simulations was done to analyse the three different strategies to handle single track operation. Also a reference simulation with both tracks in operation is used to establish a comparable punctuality with normal everyday delays caused by other circumstances than the single track operation. The needed additional time supplement is calculated by comparing the results with the results from the reference simulation. When same punctuality is achieved at commercial stops surrounding the single track section, the amount of needed time supplement is defined.

The chosen line for the simulations is the Swedish Southern mainline, a double track line with dense traffic consisting of fast long-distance trains, regional trains, commuter trains as well as freight trains. The location for the single track section is between Tunneby and Osby (see Figure 4 for a cut-out from the actual timetable) but the whole simulation area is from Katrineholm, south of Stockholm, to Malmö.

To analyse how the length of the single track section affects the strategies, three different lengths are tested for each of the strategies. To make the studies more general, runtime is chosen to define the length instead of kilometres. Single track sections that takes 5, 10 and 15 minutes for fast trains to pass are used in the simulation. In the scenarios the maximum speed for the trains is restricted to achieve the right passing time. Also, the general rules stated in section 3 are used for the strategies they can be applied.

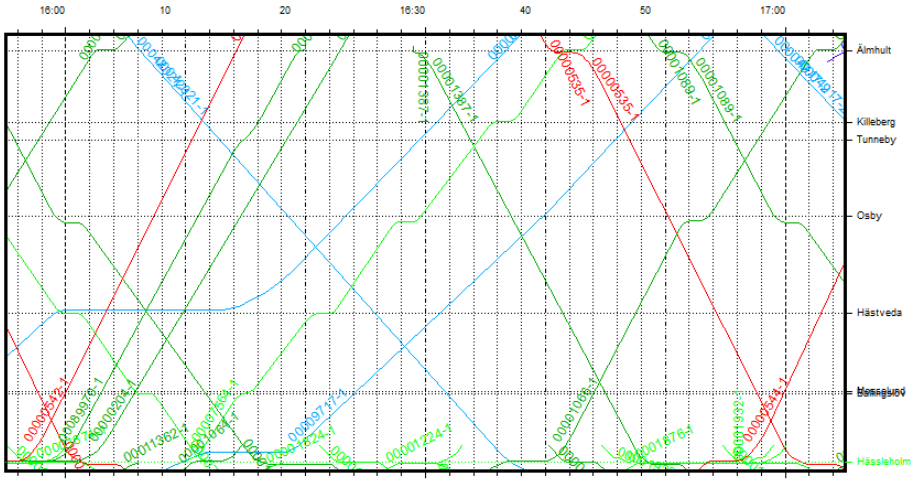


Figure 4: Peak hour traffic between Älmhult and Hässleholm, Red lines represent fast long-distance trains, dark green lines are intercity and fast regional trains, light green lines are commuter trains and blue lines are freight trains.

In the simulation, trains are disturbed with stochastic disturbances. These disturbances are inserted as entry delays, dwell time delays and line delays. Entry delays are taken from empirical delay data for the different train categories at the station where they enter the simulation area. Dwell time delays are based on empirical data from The Royal Institute of Technology (KTH) in Stockholm and inserted at each station where trains have a planned stop. The data is grouped in to different categories depending on number of passengers using the station. These delays represents the dwell time uncertainty, for example longer passenger exchange time due to train door failure. The third category of delays is line delays, i.e. delays that appear on the line between stations. These delays represent for example failure of traction units, transmission failure between signal equipment at the tracks and on-board computer that forces trains to run at reduced speed or temporary speed restrictions. These three types of delays are commonly used in Sweden in a way so that the simulations can represent realistic operational traffic in the Swedish network.

4.1 Results of the Three Strategies

After simulation of all strategies for all three scenarios with different single track length, it is possible to attain some results to evaluate. In all simulations, additional time supplement is necessary to maintain the punctuality. Depending on strategy and length of the single track section, the needed amount of time supplement differ. The supplement is needed to handle the fact that trains are not running on time. They do not always arrive to the single track section on time and need time supplement after the single track section to recover from this delay that might have increased due to the single track operation. If the time supplement is placed on the single track section instead of after, capacity would decrease.

In the following sections the summarized results of the three strategies are presented one by one.

1) Full re-scheduling

With this strategy all timetables have to be conflict free and fully adapted to the single track operation. Therefore, much effort needs to be put into creating the new timetables.

In the scenarios with 10 and 15 minutes of running time over the single track section, there is not room for all trains in the original timetable and traffic must therefore be reduced.

In all three scenarios an additional time supplement is needed to maintain punctuality for long distance and regional passenger trains. The amount of supplement is 50 % of the runtime over the single track section, e.g. if it takes 10 minutes for the trains to pass the single track section, the needed time supplement is 5 minutes. For local passenger trains and freight trains additional time supplements is not needed until the single track section takes more than 10 minute to pass. For local passenger trains the reason for this is that they have a much higher punctuality when arriving to the single track section than other trains and therefore often get prioritized. Freight trains have a lot of time supplements already in the original timetable for overtakings etc., and therefore they have less need for additional time supplements. For local passenger trains and freight trains an additional time supplement of 5 minutes is needed if the single track section takes longer than 10 minutes to pass.

2) Trains scheduled in groups

This strategy is only simulated for the scenarios with 10 and 15 minutes travel time over the single track section. The reason for this is that there is no need to group the trains together in the 5 minutes scenario. In the other two strategies all trains could fit in the timetable with acceptable amount of time supplements. If they were to be grouped we would deviate even more from the original timetable, give the trains longer runtimes and not gain anything with it.

In the simulated scenarios a large amount of additional time supplement, particularly for freight trains, is needed to gather the trains into groups. There is also a shortage of tracks near the single track section to store trains to form the groups. Passenger trains have to pass the freight trains before the single track section to be the first trains in each group. Else, they have to run after freight trains and much additional time will be necessary to lower the passenger trains' speeds to match the freight trains'.

The difference in runtimes between passenger and freight trains leads to large heterogeneousness within the train groups. This results in a timetable with a low ability to move the maintenance work to different sections along the line unless the trains are re-grouped once again. If the maintenance work consists of several moving stages a lot of timetable variants is needed in the end. However, the strategy will still reduce the number of needed timetable variants by half, compared to the strategy with full re-scheduling, because the single track section can often be moved at least one short section along the line without re-scheduling the timetable. If the traffic is more homogenous the need for several timetable variants with different train groupings decreases even more.

The simulation results show that the same amount of additional time supplement as for the strategy with full re-scheduling to handle late trains. Long distance and regional passenger trains need an additional time supplement which is 50 % of the runtime over the single track section and local passenger trains and freight trains an additional time supplement of 5 minutes if the single track section takes longer than 10 minutes to pass.

3) No re-scheduling

In this strategy, no re-scheduling of trains is done, only additional time supplements are used to give trains the possibility to recover from the delays caused by the single track

operation. In the simulation process it became clear that this strategy shows a much larger need for time supplement than other strategies.

One major issue for the punctuality is trains that systematically have meetings on the single track section every hour. Passenger trains have a more or less periodic timetable in Sweden and in some cases these trains meet at the same time every hour. These trains will frequently cause disturbances and delays for all other trains as well. In such case, all trains will need an additional time supplement which is 200 % of the runtime over the single track section to maintain punctuality. Therefore an additional rule for this strategy was developed; it is not allowed to have systematically meetings on the single track section. Trains that unfortunately have a meeting there in the original timetable have to be removed or moved to another time.

With the new rule forbidding systematically meeting the amount of additional time supplement needed to maintain punctuality is 100 % of the runtime over the single track section. This means that if it takes 10 minutes for a train to pass the single track section, the needed time supplement after the section is 10 minutes.

5 Concluding Discussion

The results from the case study show that it is important not to schedule trains as close as possible when they leave/enter the single track section. There is a need of headway margin time between the trains in opposite direction, not to spread delays too easily. Also there is a need for additional time supplements that can handle the everyday delays combined with the increased disturbance risk of the single track. In all three strategies additional time supplements are needed to handle trains that are arriving delayed to the single track section. However, depending on which strategy used, the amount of time supplement needed differs. In general, as less accurate and not conflict free the timetable is, the more supplements are needed.

Since the strategy with full re-scheduling results in the least time supplements, it would be easy to draw the conclusion that full re-scheduling always is preferred. However, due to the amount of work needed to produce a lot of timetable variants, it might not be a well-chosen use of recourses. Until an advanced timetabling tool with automatic re-scheduling of trains is implemented, the work has to be done manually and that is very time consuming. On lines or at times with homogenous traffic where the trains are running with similar speed, the strategy with grouping of trains would be a more effective strategy. Then the extra time needed to arrange the trains into suitable groups doesn't have to be that large and we can take advantage of the fact that we don't need that many timetable variants. However, in the case study presented in this paper the traffic was too heterogeneous and the strategy would not give any clear benefits compared to full re-scheduling.

If the strategy with no re-scheduling and only additional time supplements is used, we deliberately allow the trains to be delayed during the maintenance work. We therefore have to add large time supplements in the timetable so that the trains can recover from the delays, which leads to a large increase in the trains' travel times. The benefit with this strategy is that the need for several timetable variant is small, even though the single track section is shifting within the maintenance work. As long as the additional time supplement is placed after all possible single track sections, the same timetable can be used for all of them. This, of course, is only possible if we are allowed to let the trains be delayed for a long part of the line, during all stages of the maintenance work. If the time supplement is placed directly after a single track section it might be of no use when the maintenance work moves to the next section of the line.

The result from this study gives knowledge to railway timetable planners for how and when to use different timetable strategies. The amount of additional time supplements needed to preserve punctuality for each strategy is related to the specific case presented here and that case can be seen as a worst case scenario. The traffic demand is high, with a large heterogeneousness combined with a high level of disturbances on the line. If there are less trains or if the trains are running more punctual from the beginning, the need for large additional time supplements decreases. Exactly how much time supplement that is needed for different cases is for future work to analyse, but we can conclude that by using the time supplements suggested in Section 4, the punctuality would be preserved in most cases.

Also, regardless of strategy, it is important to take the characteristics of signal system and driver behaviour into account when adapting a timetable for a single track maintenance work. The general rules presented in Section 3 should always be applied since they concern the practical feasibility regardless of the amount of trains and disturbances. Without the use of the general rules the timetable will not be completely conflict free and the trains cannot run on time even though they are re-scheduled.

References

Lidén, T., 2015. "Railway infrastructure maintenance – a survey of planning problems and conducted research", *Transportation Research Procedia 10*, pp. 574-583.

RMCon (2019), Rail Management Consultants GmbH website, www.rmcon.de/railsys-en (2019-04-06)

Trafikverket, 2016. *Konstruktion av körplaner för tåg* (Railway timetable construction, in Swedish), TDOK 2016:0128, Guideline from the Swedish Transport Administration.

Van Aken, S., Besinovic, N., Goverde, R., 2017. *"Designing alternative railway timetables under infrastructure maintenance possessions*", Transportation Research Part B 98, pp. 224-238.

Vansteenwegen, P., Dewilde, T., Burggraeve, S., Cattrysse, D., 2016. "*An iterative approach for reducing the impact of infrastructure maintenance on the performance of railway systems*", European Journal of Operational Research 252, pp. 39-53.

Trapeze (2019), Trapeze Group website, www.trapezegroup.eu/solution_sheet/trapeze-rail-system-timetable-planning (2019-04-06)

Data Reconciliation of Freight Rail Dispatch Data

William Barbour [a,1], Shankara Kuppa [b], Daniel B. Work [a]
[a] Department of Civil and Environmental Engineering,
Institute for Software Integrated Systems
Vanderbilt University
[b] CSX Transportation
[1] E-mail: william.w.barbour@vanderbilt.edu

Abstract
In order to enable widespread use of data driven analysis and machine learning methods for rail operations problems, large volumes of operational data are needed. This data has the potential to contain erroneous or missing values, especially given its size and dimensionality. In this work a data reconciliation problem for rail dispatch data is proposed to identify and correct errors, as well as to impute missing data. The data reconciliation problem finds the least-perturbed modification of the historical data that satisfies operational constraints, such as feasibility of meet and overtake events, safety headway, siding allocation, and running time. It also imputes missing values with estimates that satisfy all operational constraints. The data reconciliation method is applied to a large historical dataset from freight rail territory in Tennessee, United States, containing over 3,000 train records over six months. The method identifies and corrects errors in the historical data, and is able to impute data on a synthetically decimated version of the historical data. The quality of the imputed data from data reconciliation is compared to imputed data using naive interpolation. The results show that data reconciliation reduces timing error of imputed points by up to 15% and increases the number of meet and overtake events estimated at the correct historical location from less than 40% to approximately 95%. These findings indicate that the data reconciliation method is a useful preprocessing step for analysis and modeling of railroad operations that are based on real-world physical dispatching data.

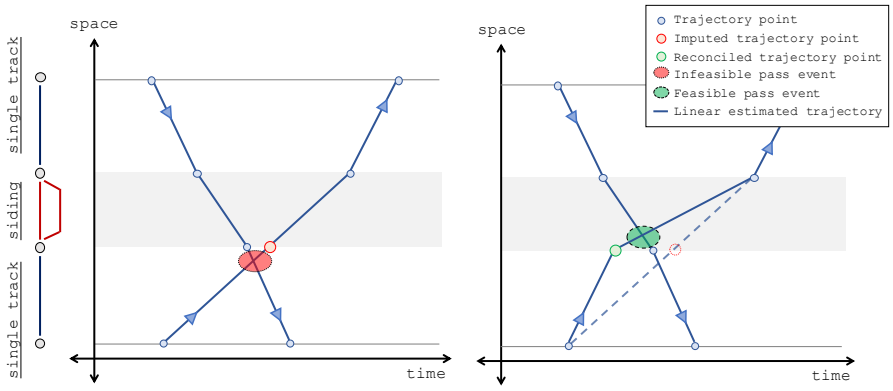
Keywords
data reconciliation, dispatching, modeling, optimization

1 Introduction

1.1 Motivation

Data-driven methods for railroad operations require abundant, high-quality sources of data for model building. Machine learning, and deep learning methods in particular, require large datasets for training. These methods will learn trends from input data, so if the data contains errors, then the errors may propagate into the trained model and the resulting analysis.

A common challenge in the emerging data science and data analytics fields is the amount of time spent on data cleaning and data preparation. Common tasks include standardizing and normalizing data, identifying faulty data and discarding or correcting it, and imputing values for missing entries. Especially when (reasonably) clean data from large systems



(a) Infeasible trajectories.

(b) Feasible trajectories.

Figure 1: A time-space plot example of two trains traveling in opposite directions. Train trajectories are denoted by the blue points and linearly estimated between points. It is obvious that these trains met at the siding track (shaded grey area). This example demonstrates how an erroneous trajectory point (red point) in (a) can result in an infeasible meet location (shaded red oval). This can be corrected by reconciliation of the timing to the green point in (b) to make the meet occur at a feasible location on the siding track (within the grey area).

that describe physical processes (e.g., freight rail flows) is needed, ad hoc and manual approaches to data preprocessing can easily be inefficient and can often be intractable.

To automate some aspects of data cleaning and data preparation, it is possible to use knowledge about the physical constraints of the system to identify and correct erroneous values. The process by which missing data is estimated and erroneous data is corrected using a model as a constraint is referred to as *data reconciliation* (Tjoa and Biegler, 1991). Given that railroad operations have obvious logical and physical constraints, useful data reconciliation problems can be posed and solved, which we demonstrate in this work.

1.2 Overview of Data Errors in Dispatch Data

Train trajectory data typically comes in the form of train arrival times at fixed locations on the rail network; in the United States these are often called *on-station points*, or *OS-points*. Most OS-points are located at the endpoints of passing sidings in single-track territory or at crossover points in multi-track territory. Track segments refer to the sections of track delineated by OS-points. The arrival times of a train at each OS-point between two rail yards or terminals on the network constitute a trajectory.

Data errors are identifiable as infeasible trajectories because they violate *meet* constraints (passing events between trains in opposing directions), *overtake* constraints (passing events between trains in the same direction), headway constraints (trains following, meeting, overtaking with insufficient time headway clearance), or other operational constraints. Data may also be missing, e.g., due to incomplete data fusion or sensor failures, which can further compound the difficulty of identifying and correcting errors.

Consider the data error in Figure 1a, a time-space plot, as an example. Three tracks

are shown on the spatial axis (y-axis): one siding track denoted by the grey shading and two single track segments. Trajectories of two trains traveling in opposing directions are denoted by the blue points, which represent known trajectory points, and the blue lines that are the linearly approximated trajectories between them. In Figure 1a, there is an imputed point shown in red along the trajectory. The timing value of this point results in a meet event (intersection between trajectories highlighted by the red shaded circle) between the two trains that occurs on a single track segment and is therefore infeasible. It is clear that the two trains must have passed each other on the siding and that the imputed trajectory point must be incorrect. Indeed, by relocating the erroneous imputed point to the location of the green point in Figure 1b, a feasible set of trajectories is found where the meet event occurs on the siding (highlighted by the green shaded circle). The amount by which the imputed point must be moved is dictated by the safety headway.

This simple example demonstrates one types of data error that can be encountered in rail operations data, noting that in real datasets the errors can be more complex, randomly distributed in the dataset, and can be compounded by missing values. As a consequence, ad hoc or manual approaches to diagnose and fix infeasible data are not viable on national-scale rail networks that move thousands of trains daily.

1.3 Problem Statement and Contribution

The main contribution of this work is the development of a method to perform optimization-based *data reconciliation* of railroad dispatching data. Given a set of train trajectories and as set of operational constraints, the data reconciliation problem simultaneously corrects any data that is infeasible, and also imputes any missing data.

A constraint set from a dispatch optimization problem that models single-track rail operations is used to perform data reconciliation and we note that the method may be generalized to other optimization-based dispatch tools and network topologies. To illustrate the performance of the method, the data reconciliation problem is implemented on a real freight rail dataset with synthetic omissions in the data. This is the first work to formalize data reconciliation for cleaning rail dispatch data, which can be a critical step for machine learning and data-driven rail operations and is a practical challenge in the transportation industry.

The remainder of the article is organized as follows. Section 2 reviews the related work on optimal dispatching, data driven rail operations, and data reconciliation problems posed in other domains. Section 3 provides a general forms of optimization-based dispatching and its relationship to the data reconciliation problem. Section 4 instantiates a specific data reconciliation problem used in the work on a real dataset from a US Class-1 railroad. In Section 5, we present and discuss results from applying the reconciliation to the historical dataset and to a synthetically incomplete dataset.

2 Background

First a selection of prior work pertaining to posing and solving the rail dispatch problem with optimization models is summarized. These models define the basis of logical and physical constraints for rail operations that are used by the data reconciliation problem. We then explain the need that data-driven rail analysis techniques have for large volumes of clean historical data and provide examples of such work. Finally, we discuss prior work on data reconciliation from other domains of study.

Optimization-based Rail Dispatch

Optimization-based rail dispatch is a common tool in passenger and freight railways around the world. Many rail dispatch and control schemes still require humans in the loop, but actions and plans are often suggested by *computer aided dispatching systems* (CAD) (Petersen et al., 1986). These systems are given the physical and logical constraints of the network, which include network topology, speed limits, signalling, train passing logic, and train physics, alongside railroad operating practices and preferences, which include train schedules, train priority, and delay recovery (Wang and Goverde, 2016; Khoshniyat and Peterson, 2015). The routing problem considers these many factors and constraints and, along with the size of the rail network, results in a large *mixed integer linear program* (MILP) (Bollapragada et al., 2018; Higgins et al., 1996). These problems are applicable at multiple levels of operations, including tactical planning, daily operations, and re-scheduling, as outlined by Törnquist (2006). Ultimately, we show how to extend these exact types of problem formulations to the data reconciliation problem that ensures feasibility of operational data that is collected.

One of the first formal definitions of the CAD problem was by Petersen et al. (1986). Higgins et al. (1996) focused on a similar CAD model as a decision support tool for single-line railways. Murali et al. (2016) used an *integer programming* (IP) model for tactical planning on the Los Angeles rail network. The intractably large MILP problems created by timetabling over a large geographical area and long time horizon are addressed with an incremental heuristic by Gestrelius et al. (2017). Wang and Goverde (2016) took a detailed approach to trajectory optimization from the energy conservation perspective with consideration of train performance calculations. Robust train timetabling was addressed with variable time headways by Khoshniyat and Peterson (2015). Törnquist and Persson (2007) studied the effects of different optimization objectives using a heuristic technique for disturbance re-scheduling on a mixed passenger/freight traffic corridor in Sweden. Bollapragada et al. (2018) describe a modern optimization-based system that handles train dispatching and other ancillary activities used at Norfolk Southern Railway in the United States. Much of the dispatching, scheduling, and disruption management literature is well-summarized by Fang et al. (2015) at the strategic, tactical, operational, and rescheduling levels.

Data-driven Rail Analysis

As previously discussed, train routing and control problems are difficult and nuanced, but are increasingly the focus of railroads seeking to further optimize and automate operations. Less work has been done on the post-hoc analysis of dispatching and dispatcher performance and this line of inquiry could have implications in safety, sustainability, and automation. See Ghofrani et al. (2018) for a review of many of the applications in this field. All of these data-driven methods require large volumes of reasonably clean historical dispatch data due to the unique topology of the rail network and the complexity of operations (Wang and Work, 2015; Barbour et al., 2018b; Oneto et al., 2019; Ghofrani et al., 2018).

Oneto et al. (2019) analyzed train behavior on a network scale using both prior network knowledge and historical data. Kecman and Goverde (2015) estimated passenger train running and dwell times in the Netherlands in real time using numerous machine learning models. Chapuis (2017) used artificial neural networks to produce arrival time estimates for French passenger trains. Wang and Work (2015) used historical data on Amtrak trains in the United States, which run with far higher variability than their international counterparts, along with vector regression to estimate arrival times. Support vector regression, ensemble

decision trees, and deep learning were all used to estimate arrival times of freight trains in the United States, which operate with a low degree of scheduling and high variability (Barbour et al., 2018a,b).

Data Reconciliation

The number works focused on data reconciliation is rather limited. An optimization-based data reconciliation problem was introduced by Tjoa and Biegler (1991), where chemical process measurement and control data was studied with respect to noise reduction and gross error correction. Leibman et al. (1992) develop a new method for data reconciliation using nonlinear programming targeted at dynamic and nonlinear environments. Tong and Crowe (1995) introduced the use of principle component analysis for gross error detection in data reconciliation, as an alternative to some statistical tests. Soderstrom et al. (2001) performed gross error detection and data reconciliation simultaneously by formulating and solving a mixed integer linear program for process flows.

In the transportation field, Zhao et al. (1998) used data reconciliation techniques for processing traffic count data under flow conservation constraints. Claudel and Bayen (2011) perform data reconciliation for highway traffic data, posed as a convex program based on constraints derived from a partial differential equation describing conservation constraints.

3 Optimal Dispatch and Data Reconciliation

In this section we first explain the generalized problem formulation of optimization-based dispatching and the corresponding data reconciliation problem formulation.

3.1 Optimal Dispatch Problem

The optimal dispatch problem takes a set of trains traveling on a section of the network (e.g., between major yards or terminals) and finds feasible trajectories that are optimal with respect to minimization of a function of weighted train runtime and satisfy physical and operational constraints. Broadly, many dispatch problems can be posed in the general form:

$$\begin{aligned} \underset{x,z}{\text{minimize:}} \quad & f(x,z) \\ \text{subject to:} \quad & A_1 x + A_2 z \leq b, \end{aligned} \tag{1}$$

where the decision variables are $x \in \mathbb{R}_+^p$ and $z \in \mathbb{Z}^q$. In a common formulation, the decision variables x encode times at which trains reach various points on the network, while the integer decision variables z encode dispatching logic that indicates if and where meets and overtakes occur on the network. The objective function $f(\cdot,\cdot)$ is a performance measure that quantifies the desirability of the dispatch solution, for instance with respect to delay or priority weighted delay of trains. Integer variables may factor into the objective function if, for example, one wishes to minimize the total number of meets and overtakes. The physical and operational constraints, such as the permissible locations of meet and overtake events, headway constraints, and train travel times, are encoded in the inequality constraints $A_1 x + A_2 z \leq b$. For simplicity the constraints are assumed to be mixed integer linear, although more general dispatch problems can also be considered.

3.2 Data Reconciliation Problem

With a generic form of the optimal dispatch problem defined, it is now possible to define the corresponding data reconciliation problem. The constraint set from the train dispatch problem plays a critical role in the data reconciliation problem. Accurate data reconciliation assumes that the constraint set correctly describes the operations of the rail network. Consider a historical trajectory dataset denoted by $\tilde{x}$ and $\tilde{z}$, possibly containing missing entries. Let $\tilde{x}_\Omega$ and $\tilde{z}_\Omega$ denote the subset of the historical dataset for which entries are present. The data reconciliation problem is written as:

$$\begin{aligned} \underset{x,z}{\text{minimize:}} \quad & g(x_\Omega - \tilde{x}_\Omega, z_\Omega - \tilde{z}_\Omega) + h(x_\Psi, z_\Psi) \\ \text{subject to:} \quad & A_1 x + A_2 z \leq b, \end{aligned} \tag{2}$$

where $x \in \mathbb{R}^p_+, z \in \mathbb{Z}^q$. The variables x_Ω and z_Ω are the subset of the decision variables that correspond to the historical dataset for which entries are present and x_Ψ and z_Ψ are the subset of the decision variables that correspond to missing historical entries.The reconciliation problem finds feasible trajectories, x, z, that are feasible and minimally-perturbed from the historical data according to the performance measure $g(\cdot, \cdot)$. An additional term $h(\cdot, \cdot)$ can be added to the reconciliation problem to further regularize the missing data that must be imputed by the data reconciliation problem. Importantly, while the historical data $\tilde{x}, \tilde{z}$ may or may not be feasible, and may or may not contain missing entries, the reconciled data indicated by the decision variables at optimality, x^*, z^*, are feasible and complete provided the constraint set is not empty.

A variety of possible performance measures can be designed for the data reconciliation problem. For example, a natural choice is an $\mathcal{L}_1$ penalty on the historical data:

$$g(x_\Omega - \tilde{x}_\Omega, z_\Omega - \tilde{z}_\Omega) = ||x_\Omega - \tilde{x}_\Omega||_1, \tag{3}$$

which promotes sparsity in the changes to the timing variables from the values in the historical data. Note an $\mathcal{L}_2$ penalty can also be considered, but it may return small changes to many of the entries rather than a few changes to a few entries. In (3), we do not consider a penalty on the integer variables z_Ω even though it is possible, because it requires more care to design and depends on the interpretation of the variables. For example, in the problems instantiated later in this work, the integer decision variables are uniquely determined once the continuous variables are fixed, and the primary objective is to match the timing data as much as possible.

In cases of missing historical data, the design of the regularization term influences the quality of the imputed values found when solving the data reconciliation problem. Supposing again that x denotes timing data, and x_Ψ denotes the vector of entries of x corresponding to the missing data in $\tilde{x}$, one can advance trains as quickly as possible with:

$$h(x_\Psi, z_\Psi) = ||x_\Psi||_1. \tag{4}$$

Letting w encode the priority of trains at the various timing points, one can advance the trains based on priority weights:

$$h(x_\Psi, z_\Psi) = w^T x_\Psi. \tag{5}$$

It is also possible to regularize based on desired timings x^{des} that allow for encoding desired segment speeds (e.g., average speeds) through the sections with missing data. This can be written as:

$$h(x_\Psi, z_\Psi) = ||x_\Psi - x^{\text{des}}||_1. \tag{6}$$

It is also possible to regularize based on the integer variables z_Ψ, to indicate a preference to avoid meets and overtakes, for example.

4 Instantiation of a Data Reconciliation Problem

This section provides an overview of the data reconciliation problem formulation including the parameters, decision variables, the objective function, and constraints.

We limit the discussion to terminology and constraints needed to understand the core functionality of the model. For clarity and brevity, in this abbreviated formulation we do not describe end of train clearance timing, trains entering and exiting in the middle of the network section, multi-track segments with crossing tracks, simultaneous meet and overtake events at sidings, and some features unique to this particular network section.

The dispatch optimization and data reconciliation problems share the same parameters, decision variables, and constraint set for a given network topology. Here a specific form of the MILP is used that is based primarily on the dispatch formulation of Petersen et al. (1986) and Higgins et al. (1996), but in principle the data reconciliation problem can be posed using constraints from other optimization-based dispatching problems.

4.1 Problem Setup

A track graph for the network section over which trains operate is delineated by OS-points that are located at the endpoints of multi-track segments or siding tracks (as discussed in Section 1.2). The set of all tracks segments is denoted M, with individual segments assigned integer labels beginning with track zero such that $M : 0, 1, 2, 3, ...$ Track segments containing a siding track or multiple tracks are included in the set S, where $S \subset M$. Each track segment $m \in M$ has length K_m. Trains travel in two directions on the network: direction 1 and direction 2. Define direction 1 to be the direction of increasing track integer labels and direction 2 to be the decreasing direction. Because track segments are denoted by integers, we can refer to the successor track in direction 1 relative to a segment $m \in M$ as segment $m + 1 \in M$. Likewise, the successor track in direction 2 relative to segment m is $m - 1$.

The set of trains traveling in direction 1 is denoted I and direction 2 trains are J. Individual trains are referred to as $i \in I$ or $j \in J$ and have unique identifiers such that $I \cap J = \emptyset$. Each train has a known length denoted L_i (or L_j).

An example network section is shown in Figure 2. This section has five track segments, $M : \{0, 1, 2, 3, 4\}$, two of which contain siding tracks, $S : \{1, 3\}$. The length of each track segment is labeled K_0, K_1, etc. Two trains are also shown: train $i \in I$ in direction 1 and train $j \in J$ in direction 2.

Additional parameters used in the objective function and in constraints must be provided. Historical data, as discussed in Section 3.2, is denoted $\tilde{x}$. Specifically, we define the historical completion time of each train i (and j) for each track segment m to be $\tilde{x}_{i,m}$ (and $\tilde{x}_{j,m}$). Note that completion time of a segment is relative to direction, so the values $\tilde{x}_{i,m}$ and $\tilde{x}_{j,m}$ for the same segment m refer to different endpoints of the track segment.

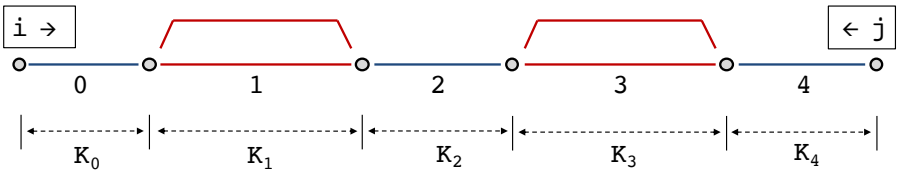


Figure 2: Depiction of notation used in data reconciliation problem for an example track graph with 5 segments. The set of all track segments in this example is $M : \{0, 1, 2, 3, 4\}$ and the set of siding segments is $S : \{1, 3\}$. The length of each track segment is denoted K_0, K_1, etc. The two trains in this example are labeled $i \in I$, which travels in direction 1, and $j \in J$, which travels in direction 2.

The free run (i.e., minimum) traversal time of each track segment is defined specific to each train. If train i takes the main line track on a segment m, its free run traversal time of the segment is $T_{i,m}$. If train i takes the siding track on a segment $s \in S$, its free run traversal time across the siding track is $U_{i,s}$. We assume that the siding free run time values are greater than the corresponding main line free run time (i.e., $U_{i,s} \geq T_{i,s}$). Trains $j \in J$ have corresponding parameters $T_{j,m}$ and $U_{j,s}$.

For meet or overtake events between pairs of trains, we define minimum clearance headways in terms of time (minutes). The minimum headway between trains traveling in the same direction is H_{i_1,i_2} (or H_{j_1,j_2}) for pairs of trains in $i_1, i_2 \in I$ (or $j_1, j_2 \in J$). For trains traveling in opposite directions, the headway time is $H_{i,j}$, where $i \in I$ and $j \in J$.

4.2 Decision Variables

The real-valued decision variables for the reconciliation problem are the reconciled trajectory timing values. The decision variables representing the reconciled data are denoted $x_{i,m}$ and $x_{j,m}$ for trains $i \in I$ and $j \in J$, respectively, corresponding to each track segment $m \in M$. These correspond to the historical data $\tilde{x}_{i,m}$ and $\tilde{x}_{j,m}$.

The integer-valued decision variables govern the interactions between trains. We use variables indicating train ordering (i.e., the order in which trains complete a track segment) to identify meet and overtake events. Let the set of track segments that are only single-track segments be denoted $M \setminus S$, which is the set M minus the set S. We define the ordering variables $\pi_{i,j,m}$ for all combinations of trains $i \in I$, trains $j \in J$, and track segments $m \in (M \setminus S)$, to be $\pi_{i,j,m} = 1$ if train i crosses segment m before train j, and $\pi_{i,j,m} = 0$ otherwise. For trains traveling in the same direction, we define $\phi_{i_1,i_2,m} = 1$ to indicate that train $i_1 \in I$ completed traversal of segment m before train $i_2 \in I$ (where $i_1 \neq i_2$), and $\phi_{i_1,i_2,m} = 0$ otherwise. Likewise, $\phi_{j_1,j_2,m} = 1$ if train $j_1 \in J$ completed traversal of m before train $j_2 \in J$, where $j_1 \neq j_2$.

The occurrences of meet events are indicated by binary values of $\mu_{i,j,s}$, which take the value $\mu_{i,j,s} = 1$ if a meet occurs between trains $i \in I$ and $j \in J$ along track segment $s \in S$, and $\mu_{i,j,s} = 0$ otherwise. The occurrence of overtake events for trains I in direction 1 are indicated by binary values of $\rho_{i_1,i_2,s}$, which takes the value $\rho_{i_1,i_2,s} = 1$ if a meet occurs between trains $i_1 \in I$ and $i_2 \in I$ (where $i_1 \neq i_2$) along track segment $s \in S$, and $\rho_{i_1,i_2,s} = 0$ otherwise. Values of $\rho_{j_1,j_2,s}$ encode overtakes for trains $j_1, j_2 \in J$ in direction

2.

When meet and overtake events occur, one of the trains in each event must take the siding track and one must take the main line track. Let $\sigma_{i,s} = 1$ if train $i \in I$ used a siding track at track segment $s \in S$, and $\sigma_{i,s} = 0$ if it did not. Likewise, let $\sigma_{j,s} = 1$ if train $j \in J$ used a siding track at $s \in S$, and $\sigma_{j,s} = 0$ if it did not.

4.3 Objective Function

The specific data reconciliation objective used in this work is as follows. We apply an $\mathcal{L}_1$ norm on the deviations from the historical data when the historical data is present, and regularize with an $\mathcal{L}_1$ penalty to a background term that encourages trains to travel at a constant speed in all sections for which data is missing. This is written as:

$$||x_\Omega - \tilde{x}_\Omega||_1 + ||x_\Psi - x^{\text{des}}||_1, \tag{7}$$

where x_Ω corresponds to a vector containing entries of $x_{i,m}$ and $x_{j,m}$ for all $i \in I$, $j \in J$, and $m \in M$ for which historical data is available. The vector x^{des} is arranged to have entries corresponding to the elements of x for which no historical data is available, and is set assuming trains in the historical dataset travel at constant speeds through sections with missing data, independent of other trains or physical constraints. Note that x^{des} may or may not be feasible, and is only used as a regularization term.

4.4 Constraints

Travel Time Constraints

Train timing at each OS-point is governed by prior OS-point timing data and minimum free run times. Precisely, the completion time $x_{i,m}$ for train i of segment m must be greater than or equal to the completion time $x_{i,m-1}$ of the preceding segment plus the minimum free run travel time $T_{i,m}$ specific to that train and segment. This is written as:

$$x_{i,m} \geq x_{i,m-1} + T_{i,m}, \tag{8}$$

where $i \in I$ and $m \in M$. For trains j traveling in direction 2, we have:

$$x_{j,m} \geq x_{j,m+1} + T_{j,m}, \tag{9}$$

where $j \in J$ and $m \in M$. Note that the segment preceding segment m is $m+1$ for direction 2, because the track segments labels are numbered in increasing order in direction 1.

When train i uses siding s (i.e., $\sigma_{i,s} = 1$), the completion time $x_{i,s}$ of the track segment s depends on the completion time of the previous segment $x_{i,s-1}$ and the minimum siding travel time $U_{i,s}$:

$$\texttt{IF } \sigma_{i,s} = 1, \texttt{ THEN } x_{i,s} \geq x_{i,s-1} + U_{i,s}, \tag{10}$$

where $i \in I$ and $s \in S \subset M$. Recall based on the numbering of the track segments, $s - 1 \in M$ refers to the track segment immediately before s and that the siding travel time $U_{i,s} \geq T_{i,s}$, indicating the minimum siding travel time is longer than the minimum main line travel time for each train at each segment.

A similar constraint on the completion time when trains $j \in J$ take the siding track handles trains in the opposite direction:

$$\texttt{IF } \sigma_{j,s} = 1, \texttt{ THEN } x_{j,s} \geq x_{j,s+1} + U_{j,s}. \tag{11}$$

Meet and Overtake Constraints

Meet and overtake events are constrained using logical properties of the binary ordering variables π and ϕ.

We constrain the arrival times of opposite direction trains at siding endpoints such that a train may not enter onto a single-track segment until the train in the opposite direction has cleared off the single-track segment, plus a safety headway. Recalling that $\pi_{i,j,m}$ indicates which train (i or j) first traverses a single-track segment $m \in (M \setminus S)$, and takes the value $\pi_{i,j,m} = 1$ if train i traverses first and 0 otherwise. Then the meet constraint is written as:

$$\texttt{IF } \pi_{i,j,m} = 1, \texttt{ THEN } x_{i,m} + H_{i,j} \leq x_{j,m+1}, \texttt{ ELSE } x_{j,m} + H_{i,j} \leq x_{i,m-1}, \tag{12}$$

where $m \in (M \setminus S)$, $i \in I$, and $j \in J$. Constraint (12) activates based on the value of $\pi_{i,j,m}$ and applies only to single track segments. If $\pi_{i,j,m} = 1$, then train i is arriving at the end of the single track segment before train j, and must have at least $H_{i,j}$ minutes of safety headway before train j proceeds onto the single-track segment. Note that because of directionality, the timing variable $x_{i,m}$ refers to the completion time of the single-track segment by train i and $x_{j,m+1}$ refers to the entry time of train j onto the same single-track segment. In the case that j traverses the single-track segment before train i ($\pi_{i,j,m} = 0$), then we require train j to finish the single-track segment, plus the safety headway, before train i may finish segment $m-1$ and enter onto the single-track segment m. Note that in the case that train j traverses m first, the constraint refers to the opposite end of the single-track segment where the completion time of train j is $x_{j,m}$ and the entry time of train i is $x_{i,m-1}$.

In the case of same-direction trains, we impose a following-headway to the completion times of each track segment depending on which train completes the segment first. Recall that $\phi_{i_1,i_2,m} = 1$ if train $i_1 \in I$ traverses segment $m \in M$ before train $i_2 \in I$, where $i_1 \neq i_2$ (i.e., train i_2 follows train i_1). In this case, the completion time $x_{i_2,m}$ of the segment for train i_2 must be at least H_{i_1,i_2} minutes (the safety headway) after the completion time $x_{i_1,m}$ of train i_1:

$$\texttt{IF } \phi_{i_1,i_2,m} = 1, \texttt{ THEN } x_{i_1,m} + H_{i_1,i_2} \leq x_{i_2,m} \tag{13}$$

A similar constraint handles the headway separation of trains traveling in direction 2:

$$\texttt{IF } \phi_{j_1,j_2,m} = 1, \texttt{ THEN } x_{j_1,m} + H_{j_1,j_2} \leq x_{j_2,m} \tag{14}$$

The next set of constraints allows overtakes only on siding segments, by forcing the order of same-direction trains to stay the same on single-track segments. For direction 1:

$$\phi_{i_1,i_2,m} = \phi_{i_1,i_2,m-1}, \tag{15}$$

where $m \in (M \setminus S)$, and direction 2:

$$\phi_{j_1,j_2,m} = \phi_{j_1,j_2,m+1}, \tag{16}$$

where $m \in (M \setminus S)$.

In a single-track network topology with a high volume of traffic, simultaneous meet and overtake events occurring at sidings with more than two parallel tracks do occur, albeit rarely. For example, if a train $i_1 \in I$ is overtaken by train $i_2 \in (I \setminus \{i_1\})$ and both i_1 and i_2 meet train $j \in J$, then three parallel tracks are required. To simplify the presentation, here we only describe the constraints that consider the case of two parallel tracks. Extensions to there or more parallel tracks result in additional meet and pass constraints that are tedious but also result in mixed integer constraints.

Recall that meet events are identified by $\mu_{i,j,s} = 1$ if a meet occurs between trains i and j at siding segment s, and $\mu_{i,j,s} = 0$ otherwise. Overtake events are identified by $\rho_{i_1,i_2,s} = 1$ if an overtake occurred between trains i_1 and i_2, and $\rho_{i_1,i_2,s} = 0$ otherwise. Consider train i_1 at track segment s. The total number of meet events train i_1 experiences with any opposite direction trains in J at s is $\sum_{j \in J} \mu_{i_1,j,s}$. Similarly, the total number of overtakes that train i_1 experiences with any same direction trains $i_2 \in (I \setminus \{i_1\})$ is $\sum_{i_2 \in (I \setminus \{i_1\})} \rho_{i_1,i_2,s}$. To avoid simultaneous meet and/or overtake events occurring on segment s, we would require:

$$\sum_{j \in J} \mu_{i_1,j,s} + \sum_{i_2 \in (I \setminus \{i_1\})} \rho_{i_1,i_2,s} \leq 1 \tag{17}$$

where $i_1 \in I$ and $s \in S$.

Likewise, for a train j_1 traveling in direction 2, we have an analogous constraint:

$$\sum_{i \in I} \mu_{i,j_1,s} + \sum_{j_2 \in (J \setminus \{j_1\})} \rho_{j_1,j_2,s} \leq 1 \tag{18}$$

with $j_1 \in J$ and $s \in S$.

Siding Assignment Constraints

For each meet event and overtake event that occurs, one of the trains must be assigned to take the siding track, which in turn imposes the minimum siding travel time constraint. These constraints are activated by the values of μ and ρ that indicate the occurrence of meets and overtakes, respectively.

Recall that the siding track indicator variable $\sigma_{i,s}$ takes the value $\sigma_{i,s} = 1$ if train i takes the siding track on segment s, and $\sigma_{i,s} = 0$ otherwise. The same is true for train j and the $\sigma_{j,s}$ variables. When $\mu_{i,j,s} = 1$, a meet occurs between trains i and j at siding segment s. As a result, one and only one of the siding indicator variables $\sigma_{i,s}, \sigma_{j,s}$ must be 1. This is written as:

$$\texttt{IF } \mu_{i,j,s} = 1, \texttt{ THEN } \sigma_{i,s} + \sigma_{j,s} = 1, \tag{19}$$

where $i \in I$, $j \in J$ and $s \in S$.

Likewise, for overtakes occurring in direction 1 between trains $i_1, i_2 \in I$ on siding $s \in S$ (indicated by the value $\rho_{i_1,i_2,s} = 1$), we have:

$$\texttt{IF } \rho_{i_1,i_2,s} = 1, \texttt{ THEN } \sigma_{i_1,s} + \sigma_{i_2,s} = 1. \tag{20}$$

A similar constraint holds for overtaking trains in direction 2:

$$\texttt{IF } \rho_{j_1,j_2,s} = 1, \texttt{ THEN } \sigma_{j_1,s} + \sigma_{j_2,s} = 1, \tag{21}$$

where $j_1, j_2 \in J$ and $s \in S$.

Finally, any trains using siding tracks must be short enough to physically fit on the available track length without interfering with switch points at the end of the siding track. If the length L_i of a train $i \in I$ is greater than the length K_s of a siding segment $s \in S$, then train i must not be assigned to take siding s (i.e., $\sigma_{i,s} = 0$). This is written as:

$$\texttt{IF } L_i > K_s, \texttt{ THEN } \sigma_{i,s} = 0, \tag{22}$$

with a similar constraint holding for trains $j \in J$:

$$\texttt{IF } L_j > K_s, \texttt{ THEN } \sigma_{j,s} = 0. \tag{23}$$

We note that variables identifying meets μ and overtakes ρ are set by additional constraints using logic derived from timing variables, which we do not enumerate here. Similar sets of constraints are also used to encode the `IF`/`THEN`/`ELSE` logic used to simplify the presentation of the constraints. The complete problem formulation results in a mixed integer optimization problem and does not require the use of a constraint programming solver.

5 Data Reconciliation Case Study on US Class-1 Freight Rail Data

In this section, the data reconciliation problem from Section 4 is run on data from a portion of a US class-1 freight railroad network. First, a description of the historical dataset and computational environment on which the data reconciliation problem is implemented are described. Two sets of experiments are run to assess the quality of the data reconciliation approach. In the first experiments, the data reconciliation problem is applied to a data which is complete but contains errors, for example due to upstream data cleaning steps to impute missing values. In the second set of experiments a synthetic dataset is created from the real original dataset by decimating entries of the complete dataset. Since the true entries are known, it allows assessment of the quality of the imputed solutions from the data reconciliation problem.

5.1 Description of Historical Dataset

The experiments described in this section use a real historical dispatch dataset from a section of the CSX Transportation rail network in the eastern United States between Nashville, TN, and Chattanooga, TN, described also in Barbour et al. (2018a,b). The time period used is six months between January 1, 2016, and June 30, 2016. The dataset contains 4368 hours of data and it includes more than 3,000 individual train trajectories. This section of the network is approximately 100 miles in length (160 km) and is highlighted by the yellow dashed box in the map in Figure 3. The test corridor is predominantly single track (blue sections on the map) with 11 passing sidings (red sections with dashed line delineations) of varying length. It is a highly congested area of the CSX network and trains must also contend with significant grade at multiple locations caused by mountains. The topology of the network combined with the high volume of traffic result in many meet and overtake events.

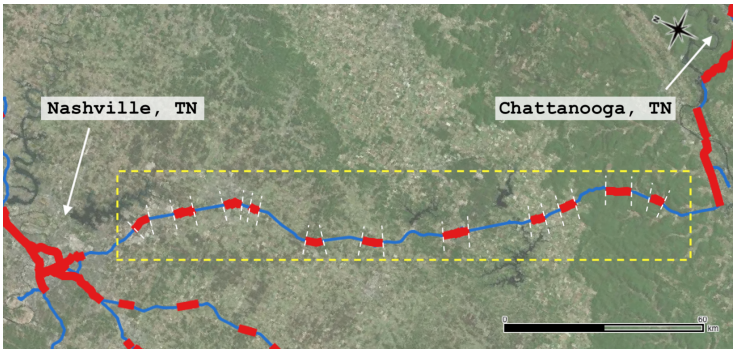


Figure 3: Map of the rail network territory is shown in the yellow dashed box between Nashville, TN, to Chattanooga, TN, United States. Multi-track sections are shown in red and single-track sections are shown in blue. The scale bar represents 60 kilometers.

5.2 Computational Environment

The data reconciliation problem is written in the AMPL mathematical programming language and solved using CPLEX 12, a commercial MILP solver. The model is connected to Python code that loads and transforms data, extracts results, and analyzes the output.

In order to maintain a reasonable size of MILP for the reconciliation problem, the data reconciliation problem is solved for datasets in a sliding window with a length between 8 and 24 hours (exact values explained in Section 5.4). A single 24 hour dataset containing approximately 20 trains yields a MILP of approximately 5,000 variables and 20,000 constraints, of which approximately 4,000 variables are binary and approximately 15,000 constraints encode logical constraints between the binary variables.

5.3 Experiment 1: Reconciliation of a Complete but Erroneous Historical Dataset

The first set of experiments are conducted on the six-month long historical dataset, which is complete, but contains errors. Any missing data points are imputed in upstream data cleaning steps which may or may not result in feasible trajectories. Using this dataset we apply the data reconciliation problem to identify and automatically correct erroneous data that do not satisfy operational constraints. The complete dataset is analyzed in a 12-hour shifting window until all data has been reconciled.

The results are as follows. On average, each 12-hour window of raw historical data contains approximately three errors that are corrected by the data reconciliation problem. Due to the proprietary and sensitive nature of the historical dataset, detailed descriptions and analysis of the errors (e.g., statistics on the types and the frequency at which they occur) specific to this dataset are not discussed in depth here. To qualitatively assess the quality of the reconciled data, after application of the data reconciliation problem, one week of the historical and reconciled data is manually inspected. The manual inspection verified that the reconciled data only deviates from the historical data in places where the historical data led to constraint violation. Common errors based on the manual inspection include infeasible meets and passes and headway constraint violations due to errors in the timing data.

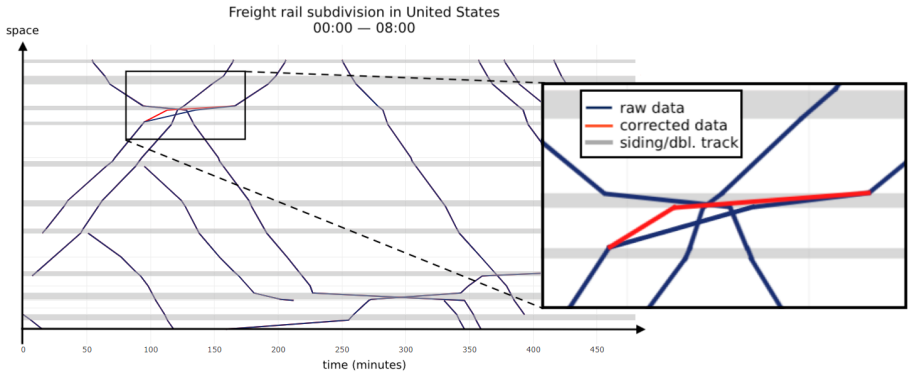


Figure 4: Stringline diagram of historical and reconciled data. Sidings and multi-track segments are shown as grey shaded areas. Raw train trajectory data is shown as blue lines. The raw data indicates that two meet and overtake events, magnified in the figure inset, occur on a single-track segment, which is infeasible. The red line is the reconciled data that results in feasible trajectories.

To give an insight into the type of errors that are automatically corrected, a representative example of an error in the historical data is shown in Figure 4 (The first eight hours of the 12 hour window are shown). Sidings and multi-track segments, where trains may pass each other, are denoted as grey shaded areas, with the white areas denoting single-track segments. The historical train trajectories (blue lines) have impermissible meet/overtake events that are magnified in the figure inset. The errors are evident in the stringline diagram because the expected meet and overtake events (i.e., the intersection point between trajectories) occur on a single track segment. In contrast, the data reconciliation problem produces the same trajectories as the historical dataset everywhere except in the neighborhood of the infeasible meet/overtake events. In that area, the reconciled data is indicated by the red line, and it results in a set feasible trajectories for all trains. There are three tracks at this passing siding, allowing both a meet and an overtake event to occur simultaneously. Note in Figure 4 that trajectories that do not cover the entire space correspond to local trains that complete routes between small intermediate destinations on this section of the network.

5.4 Experiment 2: Reconciliation of a Synthetically Decimated Historical Dataset

Next we quantitatively assess the performance of the data reconciliation problem when imputing missing data with feasible values. We begin with the historical data and create a dataset with missing entries by decimating (removing) a subset of the data entries. This is done to allow comparison between the imputed values produced by the data reconciliation problem with the true historical values that are known (but decimated in the data given to the data reconciliation problem).

To aid in interpretability of the results, the data is decimated only in areas far from any infeasible portions of the historical data, i.e., the historical data that is decimated is feasible. We clarify this is not a limitation of the method (i.e., it can be applied to a dataset containing both missing and erroneous data), but that it is not trivial to assess if differences between

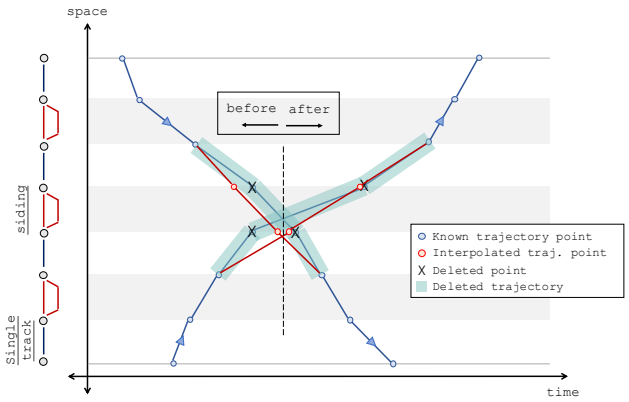


Figure 5: Four known trajectory points are selectively removed around meet/overtake events that are identified in historical data. One point immediately before the event and one point immediately after the event are removed for each train. These deleted points are shown as black 'X' markers and the missing trajectory segments are highlighted in light blue. A linear interpolation to impute the missing data (red points and lines) can result in infeasible meets.

the imputed and historical data are due to infeasibility of the historical data, or due to a poor imputed result from the data reconciliation problem. In the experiments conducted next, the synthetically decimated data is feasible so the ambiguity is avoided.

Generation of Synthetically Decimated Historical Datasets

The synthetically decimated historical dataset is created by removing known trajectory points around meet and overtake events in the reconciled historical data. At each of these events, a particular number of data points (per train) immediately before and immediately after the meet or overtake event are removed. This results in missing data centered around known meet and overtake events. Figure 5 shows an illustration of this removal process for a meet event between two trains. One point before the meet event in each trajectory and one point after the event in each trajectory are removed.

We assess and compare the quality of the imputed data from data reconciliation with imputed data from a naive linear interpolation approach. In Figure 5, the red lines and points represent the values imputed via linear interpolation. The interpolation uses the nearest known trajectory points to calculate the average speed across the missing trajectory section, from which the missing points are interpolated. There are many methods more complex than linear interpolation to which data reconciliation could be compared – speed-regularized interpolation, delay minimization, and energy conservation, to name a few – but linear interpolation is used here as a straightforward baseline method.

The quality of the imputed trajectory points is assessed by *i*) evaluating the location at which the recovered trajectories estimate meet and overtake events to occur and *ii*) calculating the time difference between each imputed value and the known trajectory value.

The location of each meet or overtake event found in the reconciled data is *feasible* if and only if it is on a siding or multi-track segment and does not violate other constraints. This location is *correct* if it matches the true location of the event indicated by the known data.

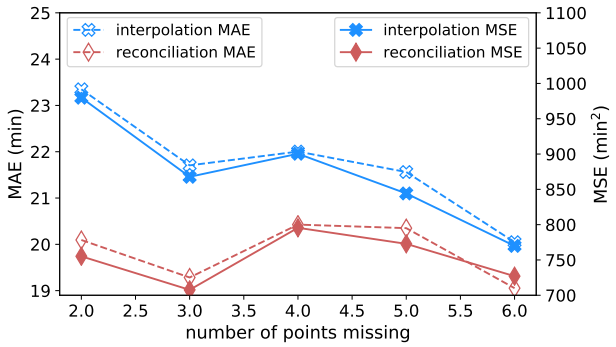


Figure 6: Mean absolute error and mean squared error of timing values for each missing point imputed by interpolated and reconciliation. MAE and MSE are averaged across trials, grouped by the total number of missing points around each meet or overtake event.

Note it is possible for linear interpolation to produce feasible or infeasible, and correct or incorrect imputed values. In contrast, data reconciliation always produces feasible imputed values which may or may not be at the correct location.

The quality of the timing data is assessed via the *mean absolute error* (MAE) and *mean squared error* (MSE) of the imputed values compared to the historical values that are decimated. Letting x^*_Ψ denote the vector of reconciled values, and with a slight abuse of notation, let $\tilde{x}_\Psi$ denote the historical data which is known but synthetically decimated in the experiment (i.e., the assumed ground truth). The quality of the imputed values are:

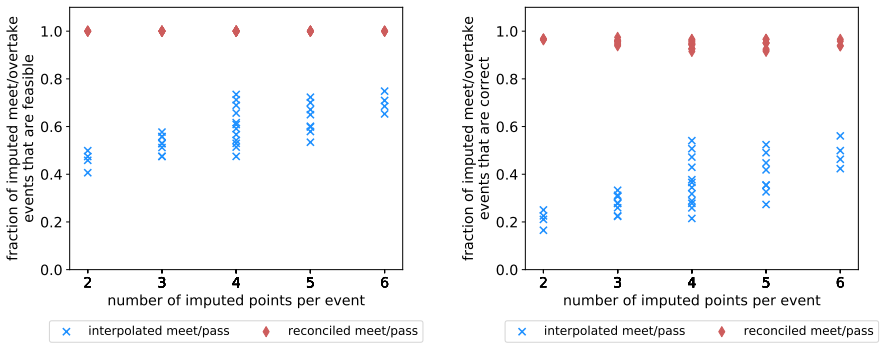
$$\text{MSE} = \frac{1}{|\tilde{x}_\Psi|}||\tilde{x}_\Psi - x^*_\Psi||_2^2, \qquad \text{MAE} = \frac{1}{|\tilde{x}_\Psi|}||\tilde{x}_\Psi - x^*_\Psi||_1, \tag{24}$$

where $|\tilde{x}_\Psi|$ denotes the number of imputed values.

Results on Synthetically Decimated Datasets

The results of the data reconciliation experiments on the synthetic, incomplete dataset are presented next. A total of 45 data reconciliation experiments are conducted on the six month dataset. Each experiment is defined by *i*) the number of points per train that are removed immediately before a meet or overtake event, *ii*) the number of points per train that are removed immediately after a meet or overtake event, and *iii*) the length of the sliding window. For example, the first experiment removes a single point per train before and a single point per train after each meet/overtake event, and the data reconciliation problem is solved on a sliding eight hour window through the six month dataset. The remaining experiments are defined by considering: *i*) the number of missing points per train immediately before a meet/overtake event (1, 2, or 3 points), *ii*) number of missing points per train after an event (1, 2, or 3 points), and *iii*) the sliding window length (8, 12, 16, 20, or 24 hours).

The MAE and MSE for trajectory points imputed by data reconciliation and linear interpolation are shown in Figure 6. The results are grouped by the number of total missing points around each meet or overtake event (i.e., the total points immediately before and after each event, resulting in between two to six missing points). Data reconciliation results in a 5-15% reduction in both MAE and MSE compared to linear interpolation.



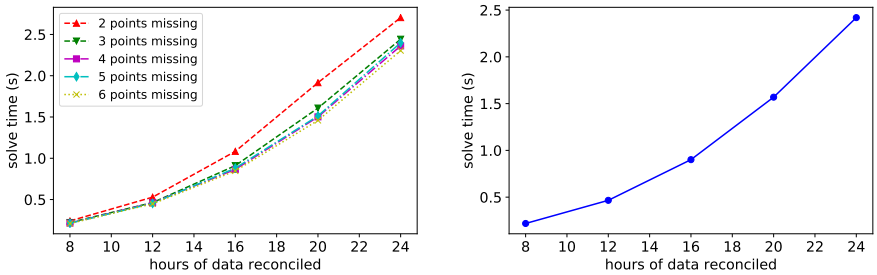
(a) (b)

Figure 7: Fraction of meet/overtake events found by data interpolation and reconciliation that are (a) found to occur at a feasible location, and (b) at the correct location.

The fraction of meet and overtake events that are found to occur at a feasible location when imputed by data reconciliation and by linear interpolation are shown in Figure 7a. The trial results are grouped by total number of points missing (i.e. points immediately before and after an event). The multiple values for a given number of points correspond to the various experiments run with differing points missing before or after each event but resulting in the same number of total missing points per event. Because the data reconciliation problem uses the physical constraints when interpolating the points, 100% of the imputed meet/overtake events are feasible. In contrast, linear interpolation results in feasible meet and overtake locations in only 40-70% of cases and exhibits variability across the different experiments for the same number of total missing points.

Figure 7b shows the fraction of meet and overtake events that are estimated to occur at the correct location as indicated by the known data, grouped by the number of missing points around each event. Reconciliation recovers the correct location for meet and overtake events in approximately 95% of cases, while linear interpolation recovers only 20-50%. Additionally, reconciliation preforms consistently across trials, with interpolation demonstrating higher variability in performance.

The reconciliation problem executes very quickly, even on large amounts of data. Solve time increases non-linearly as a function of the number of hours used for the shifting window, as seen in Figure 8b. The number of missing points per overtake event does not have a substantive effect on the solve time, as shown in Figure 8a. The solve times for two and three missing points per meet/overtake event are slightly longer than trials with larger numbers of missing points, but follow a similar trend to the larger numbers of missing points. Solve time of the reconciliation model is low due to the fact that the majority of constraints are already satisfied and the number of corrections required between historical and reconciled data is low. Based on the solve time for the reconciliation model, a year of data from a large rail network (e.g., track networks of freight railroad companies in the United States) could be reconciled in about 20 hours of total CPU time with a 24 hour sliding window.



(a) Average solve time by amount of missing data at each meet and pass event.

(b) Average solve time across all trials and amounts of missing data.

Figure 8: Solve time of data reconciliation model by length of shifting data window.

6 Conclusion

Given the growing emphasis on data driven analysis and algorithms to improve operational efficiency, tools are needed to automate the cumbersome data cleaning process. This work introduced the data reconciliation problem as a tool to correct errors and impute missing values in operational rail datasets. The data reconciliation problem leverages operational constraints that are commonly used in dispatch optimization in a new context that enables efficient reconciliation of infeasible historical data. To demonstrate the viability of the method, the data reconciliation problem is instantiated and applied to a real six-month dataset containing several thousand trains on a complex portion of a US Class-1 rail network. The data reconciliation problem is found to identify and correct erroneous data, as well as impute missing data in a way that is always feasible and often correct.

Numerous extensions to the data reconciliation problem are possible. For example, a detailed design and comparison of different performance measures in the data reconciliation problem objective function might lead to improved accuracy of the reconciled data. It will also be interesting to investigate the sensitivity of the data reconciliation problem to different constraint formulations. In addition to the optimization model discussed in this work, we also intend to test the data reconciliation model on an optimization-based dispatching formulation for multi-track network topologies. Finally, we note that the data reconciliation problem posed here does not identify inefficient but operationally feasible errors. Extensions to identify these errors would be a valuable addition to the rail data cleaning toolbox.

Acknowledgments

The authors acknowledge support by the Roadway Safety Institute, the University Transportation Center for USDOT Region 5, which Includes Minnesota, Illinois, Indiana, Michigan, Ohio, and Wisconsin. Financial support was provided by the United States Department of Transportation's Office of the Assistant Secretary for Research and Technology (OST-R) and by the Federal Highway Administration's Office of Innovative Program Delivery (OIPD), via the Dwight David Eisenhower Transportation Fellowship Program.

References

Barbour, W., Mori, J. C. M., Kuppa, S., and Work, D. B. (2018a). Prediction of arrival times of freight traffic on us railroads using support vector regression. *Transportation Research Part C: Emerging Technologies*, 93:211–227.

Barbour, W., Samal, C., Kuppa, S., Dubey, A., and Work, D. B. (2018b). On the data-driven prediction of arrival times for freight trains on us railroads. In *Proceedings of the IEEE 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 2289–2296.

Bollapragada, S., Markley, R., Morgan, H., Telatar, E., Wills, S., Samuels, M., Bieringer, J., Garbiras, M., Orrigo, G., Ehlers, F., Turnipseed, C., and Brantley, J. (2018). A novel movement planner system for dispatching trains. *Interfaces*, 48(1):57–69.

Chapuis, X. (2017). Arrival time prediction using neural networks. In *Proceedings of RailLille2017: 7th International Conference on Railway Operations Modelling and Analysis*, pages 1500–1510. International Association of Railway Operations Research (IAROR).

Claudel, C. G. and Bayen, A. M. (2011). Convex formulations of data assimilation problems for a class of Hamilton–Jacobi equations. *SIAM Journal on Control and Optimization*, 49(2):383–402.

Fang, W., Yang, S., and Yao, X. (2015). A survey on problem models and solution approaches to rescheduling in railway networks. *Transactions on Intelligent Transportation Systems*, 16(6):2997–3016.

Gestrelius, S., Aronsson, M., and Peterson, A. (2017). A MILP-based heuristic for a commercial train timetabling problem. *Transportation Research Procedia*, 27:569–576.

Ghofrani, F., He, Q., Goverde, R. M., and Liu, X. (2018). Recent applications of big data analytics in railway transportation systems: A survey. *Transportation Research Part C: Emerging Technologies*, 90:226–246.

Higgins, A., Kozan, E., and Ferreira, L. (1996). Optimal scheduling of trains on a single line track. *Transportation Research Part B: Methodological*, 30(2):147–161.

Kecman, P. and Goverde, R. M. (2015). Online data-driven adaptive prediction of train event times. *IEEE Transactions on Intelligent Transportation Systems*, 16(1):465–474.

Khoshniyat, F. and Peterson, A. (2015). Robustness improvements in a train timetable with travel time dependent minimum headways. In *Proceedings of the 6th International Conference on Railway Operations Modelling and Analysis (RailTokyo2015)*.

Leibman, M., Edgar, T., and Lasdon, L. (1992). Efficient data reconciliation and estimation for dynamic processes using nonlinear programming techniques. *Computers & Chemical Engineering*, 16(10-11):963–986.

Murali, P., Ordóñez, F., and Dessouky, M. M. (2016). Modeling strategies for effectively routing freight trains through complex networks. *Transportation Research Part C: Emerging Technologies*, 70:197–213.

Oneto, L., Buselli, I., Lulli, A., Canepa, R., Petralli, S., and Anguita, D. (2019). A dynamic, interpretable, and robust hybrid data analytics system for train movements in large-scale railway networks. *International Journal of Data Science and Analytics*, pages 1–17.

Petersen, E., Taylor, A., and Martland, C. (1986). An introduction to computer-assisted train dispatch. *Journal of Advanced Transportation*, 20(1):63–72.

Soderstrom, T. A., Himmelblau, D. M., and Edgar, T. F. (2001). A mixed integer optimization approach for simultaneous data reconciliation and identification of measurement bias. *Control Engineering Practice*, 9(8):869–876.

Tjoa, I.-B. and Biegler, L. (1991). Simultaneous strategies for data reconciliation and gross error detection of nonlinear systems. *Computers & Chemical Engineering*, 15(10):679–690.

Tong, H. and Crowe, C. M. (1995). Detection of gross erros in data reconciliation by principal component analysis. *AIChE Journal*, 41(7):1712–1722.

Törnquist, J. (2006). Computer-based decision support for railway traffic scheduling and dispatching: A review of models and algorithms. In *Proceedings of the 5th Workshop on Algorithmic Methods and Models for Optimization of Railways (ATMOS'05)*, volume 2.

Törnquist, J. and Persson, J. A. (2007). N-tracked railway traffic re-scheduling during disturbances. *Transportation Research Part B: Methodological*, 41(3):342–362.

Wang, P. and Goverde, R. M. (2016). Train trajectory optimization of opposite trains on single-track railway lines. In *Proceedings of the International Conference on Intelligent Rail Transportation (ICIRT)*, pages 23–31.

Wang, R. and Work, D. B. (2015). Data driven approaches for passenger train delay estimation. In *Proceedings of the IEEE 18th International Conference on Intelligent Transportation Systems (ITSC)*, pages 535–540.

Zhao, M., Garrick, N., and Achenie, L. (1998). Data reconciliation—based traffic count analysis system. *Transportation Research Record: Journal of the Transportation Research Board*, 1625:12–17.

Sound evaluation of simulation results

Matthias Becker [a 1], Thorsten Büker [a],
Eike Hennig [a], Felix Kogel [a]
[a] VIA Consulting & Development GmbH
Römerstr. 50, 52064 Aachen, Germany
[1] E-mail: m.becker@via-con.de, Phone: +49 (241) 463 662-26

Abstract
Simulation is one of the powerful means within the toolset of railway operations research. In contrast to timetabling and to queuing theory, it supports a precise representation of interdependencies and has thus a large field of application. Since in today's railway operation many timetable concepts and even big investment-decisions are based on studies conducted with simulation tools, a focus should be set to the sound evaluation of simulation results, too. Nevertheless, the aggregation, validation and interpretation of simulation (raw) data can barely be found in literature. This fundamental task is subject of this paper.

A simulation consists of the following steps: model design, parametrisation and calibration, simulation, processing of raw data, interpretation and visualisation of results. First, various input parameters are manipulated and simulation results are manually evaluated in a simple closed-loop principle. As each simulation is subject to outliers, runs affected by dubious conflict solutions have to be identified and excluded automatically. In most cases, a special focus is on the comparison of different scenarios and the necessity of establishing comparability by forming intersections between the simulation runs. The remaining subset of simulation runs per scenario can be considered (statistically) representative, as soon as the key figure of each scenario series converges. Finally, the raw data can be processed for the evaluation of simulation results. Results of simulations are mostly complex but by producing results for different target groups the complexity has to be reduced without losing important details or provoking misinterpretation. For this reason, it is necessary to choose key figures which comprehensively represent the simulation results.

Keywords
simulation, evaluation, calibration, intersectioning, interpretation

1 Introduction

There are many different procedures to analyse railway operations. All of these approaches have different objectives. By some of them, it is possible to analyse real-time operational data to evaluate the current performance of a railway system, while others focus the calculation of capacity and operational quality by means of queuing theory or simulation. Simulation is one of the powerful means within the toolset of railway operations research. In contrast to pure timetabling and to queuing theory, it supports a precise representation of interdependencies and has therefore a large field of operation. Some of the benefits of simulation are:

1. Illustration of complex systems (infrastructure, timetable and operational procedure)
2. Cost-effective and fast analysis of different crucial questions

3. No need for real time tests on existing infrastructures

With the use of railway simulation tools, it is possible to analyse various different scenarios and evaluate the resulting effects. As an example, the scenarios may differ by infrastructure design (microscopic track layout), by command and control system (CCS) or by timetable-concepts. The results of a simulation run is a huge amount of data. These data mainly consist of planned and actual arrival-, departure- and passage-times of all trains (in all stations) within the simulation model. Afterwards all representative information (punctuality/delay) has to be gathered out. Since in today's railway operation many timetable concepts and even big investment-decisions are based on studies conducted with simulation tools, a focus should be set to the sound evaluation of simulation results, too.

In R&D tradition, there has been substantial work in the development of simulation tools of different nature. While either the simulation algorithm or the simulation evaluation is addressed within a variety of publications, the execution of studies relies on an important interim step: aggregation, validation and interpretation of simulation (raw) data. Barely no literature can be found. This fundamental task is subject of this paper. It is structured as follows: Paragraph 2 describes the motivation for this paper. In paragraph 3 we focus on the requirements for conducting a simulation and describe the possible key figures one can get from railway simulations. Afterwards chapter 4 covers the aggregation and interpretation of simulation raw data. Finally yet importantly, we conclude this paper in paragraph 5.

2 Motivation

There is long-lasting series of research on simulation of railway operation and only some exemplary publication can be listed (Penglin (2000), Gröger (2002), Gray (2013), Jensen (2014), Ochiai (2014), Lindfeldt (2015)). Some microscopic simulation tools, such as RailSys and LUKS, provide an explicit conflict detection and solution in a synchronous and/or asynchronous manner (Weymann (2008)). Recently optimisation components are applied within conflict solution (Weymann (2015)), too.

All publication mentioned above have in common, that they describe either the simulation algorithm or the evaluation of results (with a clear focus on the first aspect). Nonetheless, to achieve reliable outcomes an important interim step may not be discarded: aggregation, validation and interpretation of simulation (raw) data. Standard literature like (Hansen (2008)) leaves out this aspect, too. To close the gap, we try to give some insights within this paper.

Subsequently, the wording is related to simulations following the Monte-Carlo principles: Per scenario, a series of simulation runs is carried out in a deterministic manner. The delays per train and location are returned per run. Results of all runs form a sample that is evaluated by stochastic means to aggregate key figures related to the scenario. For simulation approaches, which rely on direct manipulation of distribution functions instead of Monte-Carlo principles, such as (Büker (2012)), the majority of subsequent considerations is also valid.

3 Requirements

Subject of any simulation are the timetable plus the underlying infrastructure. To achieve a realistic representation of in-field operation, various input parameters serve the calibration

of the simulation model:

- Simulation results depend on the magnitude of primary delays being "spread" into the simulation model. Usually, cumulative distribution functions describe the random primary delays, which are sampled to lists of realisations. Per simulation run, a list of realisations is used. (If various scenarios are under investigation, a superset of random variables has to be used to guarantee comparability.)
- Those primary delays cause delays, which may result in secondary delays due to delay propagation.
- The conflict-solution component (e.g. two-train approach or linear optimisation) aims to reduce the magnitude of secondary delays under the regime of a target function. Any conflict-solution component has to be configured by train- and route-dependent priorities as well as malus coefficients to ensure a behaviour close to real world.

3.1 Simulation Model

In order to produce valid simulation results, it is necessary to rely on resilient input data. The allocated primary delays, the available stopping and running time supplements, the dimensioning of the investigation area as well as the settling time are highly relevant. The simulation model has to be fine-tuned to such an extent, that the key figures are sufficiently accurate to draw conclusions either by comparison or in an absolute manner.

Primary Delays
The compensation of delays is probably the greatest challenge of railway operation. In order to represent real world disturbances different disturbance variables are considered in a simulation model:

- Primary delays at entry into the investigation area,
- Primary delays at commercial or operational stops
- Continuous running time extension

Ideally, primary delays at entry and commercial stops can be derived from operational data. If this is not possible – and this is the common situation – one has to make use of standard delays which are ideally differentiated according to type of train and utilization rate.

Supplements
Supplements enable a train to recover from possible delays and to approach the reference trajectory again. The success of this intention significantly depends on the available stopping and running time supplements. As the stopping time supplement is the share of stopping time that is not used for door operation, passenger exchange and dispatching time, it is very important to define this minimum stopping time with caution so that the stopping time supplement is not overestimated.

The running time can be differentiated into a technical minimum running time and into additional running time supplements, which are allocated either for timetable robustness or during timetable construction as part of the conflict solution. A delayed train is able to make use of its supplements with regard to interdependencies with other trains.

Investigation Area
It is necessary to border the investigation area sufficiently large so that partly far-reaching interdependencies can be evaluated within the simulation. In a first approximation, it is useful to limit the investigation area at least at the next larger main railway station. Furthermore, it is recommended to extend the investigation area if there are turnaround station in the closer proximity. As disturbance lists are always inflexible after calculated once, the assumed delays at entry cannot be reduced in scenarios with a better overall operational quality even if trains might enter the investigation area more punctual as a reaction of the more punctual system itself. By the integration of turnaround stations this disadvantage can be reduced, as the arrival delay is propagated to the next train run (minus stopping time reserves) and the number of fix entries is minimized.

Stable State of the System
Besides the geographical definition of the investigation area, it is also mandatory to define a time window to be analysed and as well to determine the necessary lead time in order to guarantee a stable state of operation during the examination time window. A tight lead time provokes that the operating programme has not yet completely started so that the simulation results overestimate the operational quality.

Turnaround and Passenger Changing Connections
In turnaround stations it is a must that consecutive train runs are linked so that the operation with one single train is considered. The simulation tool ensures that the following train run can only start after arrival of the first train and the following time demand for the turnaround, which is usually configurable. If purposeful, dependencies due to staff and passenger transfer times can be defined, too.

3.2 Preparation of the Model for the Simulation

A major driver to perform simulation studies is the adaptation of infrastructure (e. g. reduction, extension, changes to command and control technology). In most cases, the infrastructure model is prepared in a semi-manual manner. Afterwards the timetable is compiled in conjunction with the infrastructure model. Scheduling as well as first series of simulation serve to validate the basic model and to detect and fix modelling errors. (Daily practice underpins that even simulation on infrastructure models for productive train-path allocation requires error elimination, as merely such data is maintained which is necessary to schedule regular train paths.) This validation requires spending a close look to the results of the early simulations instead of blind trust into simulation outcomes. Once all errors have been corrected, the model calibration may be launched.

3.3 Key Figures

The primary output of a simulation is a huge amount of raw data, which have to be analysed, aggregated and interpreted in order to draw meaningful conclusions. Standard key figures are described in various publications. For this reason, key figures are only that shortly defined as it is necessary for the later paragraphs. Standard key figures are:

- Average lateness per train
- Average lateness per delayed train
- Average additional lateness per train

- Number of late trains
- Percentage of late trains
- Punctuality of trains
- Propagation of delays between selected stations

Most key figures cannot only be calculated for simulation results but also be derived from measurements in real world operation. The absolute delays as well as the punctuality, the development of delay over a train run and the travel-time quotients for operation can be calculated based on operational data. Simulation tools offer an additional key figure called infrastructure-related hindrances that usually cannot be derived from measured data as the dependencies cannot be reconstructed. The evaluation of operational raw data is already analysed (Graffagnino (2012)).

Table 1 gives an overview on the consecutively described key figures.

Table 1 Key figures in reality and simulation

Key figure	**Evaluable in reality**	**Evaluable in simulation**
Delay (+ related key figures)	x	x
Punctuality	x	x
Travel-time Quotient Operation	(x)	x
Infrastructure-related hindrances	-	x

Absolute Delay and Development of Delay

The base key figure of any simulation is the absolute delay of a train at each occupation element. Nearly all further key figures are based on the absolute delay. The difference of at least two absolute delays describes the development of delay over a section of a train run. The development of delays is a relative consideration, which helps identifying bottlenecks within the infrastructure that are places of a high delay propagation.

Punctuality

When an absolute delay is compared with a quality target that limits the acceptable delay by aid of a threshold, it is possible to attribute each train run to be punctual or not. The relative share of punctual trains results in the key figure punctuality. This threshold can be defined for each country, infrastructure manager or even system.

Vice versa it is possible to determine a delay that is not exceeded by a defined amount. Typical thresholds are the quantiles as well as the 95 percent probability that excludes five percent of the worst trains.

Travel-time Quotient

The travel-time quotient is represented by two running times. It is possible to differentiate between the travel-time quotient for timetables and for operation. In this paper only the travel-time quotient for operation (TTQ Operation) is relevant and further described. The TTQ Operation describes the quotient of simulated running time of a train and its scheduled running time. Consequently, a TTQ Operation smaller than one expresses a situation where a train realises a shorter running time than planned on the one hand. Usually this phenomenon can be observed, when a train is initially delayed, but has running and/or stopping time supplements that can be used for delay reduction. A TTQ Operation larger than one describes on the other hand that the planned running time is not sufficient so that

the train gains delays. Whereas a quotient smaller than one is unambiguously interpretable a quotient slightly larger or equal one may result of insufficient supplements or even a punctual train that has no need to run faster than planned.

Infrastructure-related Hindrances
Infrastructure-related hindrances reveal infrastructure elements that produce or propagate delays. These hindrances are accumulated over all events and their duration. Hindrances usually occur at signals, turnouts and stopping positions as result of a parallel demand of more than one train. The hindrances can be visualized within the track diagram and indicate the bottlenecks within a network.

4 Methodology

This paragraph describes the process from the calibration of a model and preparation of raw data to the interpretation and preparation of the results. A special focus is on the comparison of different scenarios and the necessity of establishing comparability by the identification of outliers and forming intersections between the simulation runs.

4.1 Calibration of the Simulation

In a simple closed-loop principle, various input parameters as well as settings are manipulated and simulation results are manually evaluated. The major mean of validation are time-distance graphs after simulation as they provide the best visualisation of a simulated operation with focus on the simulation specific conflict solution. In this step, calibration happens to the expectations of the user, who needs to have specific knowledge on railway operation in general and to the specific situation. Real-world key figures may serve as secondary reference, only, if available for the specific situation at all. To ensure an overall comparability of outcomes, the calibration principles should thus be as standardised as possible – as well throughout setting up various models as throughout working by different users.

It is in the responsibility of the user to adapt the settings of the simulation tool or the whole model in order to define a proper solution space for the conflict solution. Concerning the infrastructure model it can be useful for instance to remove opposite track movements from the solution space or reduce the costs for alternative track occupations at the same platforms where it is practicable and useful in reality. Furthermore, it can be recommendable to adapt the priorities of a train family, if a train is much discriminated by conflict solution otherwise. This may happen for instance, if a freight train shares its infrastructure with highly prioritized long-distance trains and is unrealistically long directed into sidings due to the target function of conflict solution. From the perspective of operation, it has to be clarified if the simulation may use additional operational stops ahead of junctions in order to reduce the length overtaking sections and enable an earlier departure of the trains from the previous station. This calibration reduces the number of unrealistic conflict solutions and prioritizes real-world conflict solutions even if they might be worse than computed conflict solutions.

4.2 Individual Evaluation of each Simulation Run

After calibrating the model, simulations are ideally performed in a mostly automatic setup. In most tasks, studies do not cover just one simulation series for one (calibrated) model but are of comparative nature. For instance, the optimum combination of infrastructure and timetable shall be found and proven. Subsequently we name a combination of input parameters a scenario. If either timetable or infrastructure vary between the scenarios, disturbances and configuration should be as constant as possible between the scenarios. For each scenario, a series of runs is simulated. To assure comparability of figures between the scenarios, any evaluation has to follow certain principles. Figure 1 provides a first insight into the process of the preparation of raw data.

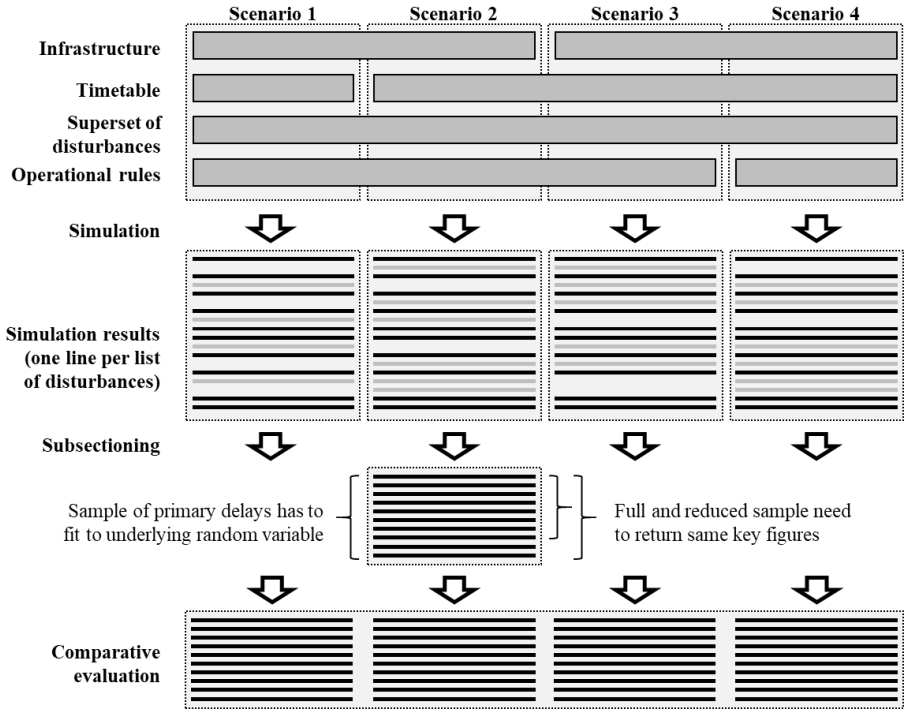


Figure 1 Simplified flow-chart of comparative analysis

Even in case of diligent calibration of input parameters, there are simulation runs whose results differ from actual behaviour severely. This happens as the conflict-solution component, as any human dispatcher, either does not find the optimum solution or even misbehaves. This mostly results from the reduced set of options for conflict solution compared to reality (e. g. partly cancellation of service, discordance of stops). The dubiety of such simulation runs in general correlates to a mix of:

- High primary delays
- Ambitious timetable concepts (only few supplements and low buffer times)
- Complicated/limited infrastructure (e. g. many crossings, single track)

An example of a set of simulation runs including those dubious ones is visualized in Figure 2.

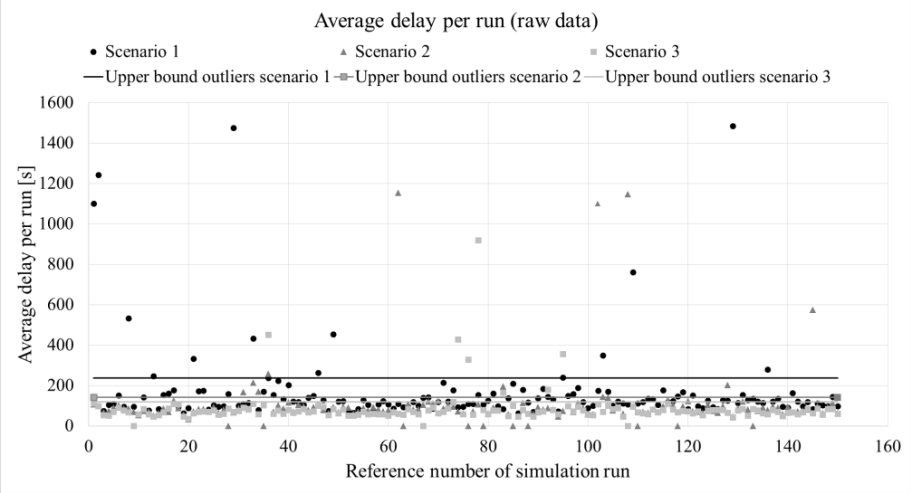


Figure 2 Exemplary set of raw data

As simulations are executed mostly automatic, those runs out of the series being affected by dubious conflict solutions have to be identified. This identification has to happen either on the level of the whole run or on the level of the train. Again, identification has to follow standardised principles to guarantee comparability. A useful key figure for identifying outliers is the average delay in all stations. This key figure is exemplary visualized in Figure 2 for each run for each scenario. For instance, it is obvious that simulation run number 129 of scenario 1 with an average delay of 1485 seconds is an outlier (upper right corner). This statement is supported by the fact that the average delay of scenario 2 and 3 is only 46 and 42. Statistically these outliers can be excluded for each scenario by an upper bound, which can be defined as the 1.5-fold of the range between the 25 and 75 percent quantile on top of the 75 percent quantile. This upper bound is also visualized in Figure 2 by a solid line.

If results of various scenarios shall be compared, a scenario-comprehensive intersection of runs with similar disturbance sets has to be created, firstly. The elimination of outliers and intersecting the remaining simulation runs of each scenario afterwards sometimes reduces the number of remaining evaluable simulation runs considerably. Figure 3 shows the remaining simulation runs after the previously described steps. In this case, the number of evaluable simulation runs is reduced from 150 to 102 simulation runs.

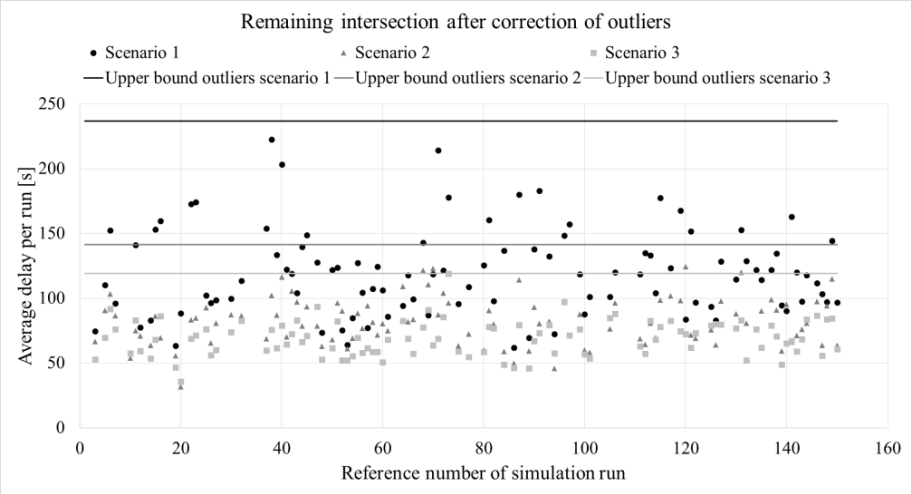


Figure 3 Remaining simulation runs after correction of outliers and intersecting

The related remaining subset of simulation runs per scenario can be considered (statistically) representative, as soon as the key figure of each scenario series converges. This subset is analysed per scenario:

- Firstly, the resulting disturbances have to fit to the distribution function of primary delays
- Secondly, key figures have to converge. For that purpose, the key figure average delay per run of each scenario is averaged over an increasing amount of simulation runs. In case they differ by less than an acceptable epsilon, results are considered to be statistically sound.

Figure 4 shows the effect of converging simulation results over the quantity of simulation runs. It is neccessay that this key figure is statistically distributed as simulated and not sorted according to size. The crosshatched lines represent the upper and lower boundary in relation to the average of all simulation runs plus/minus an epsilon. The epsilon represents the accepted dispersion of the results related to the expected value of the entire sample and is defined as 1.5 %. In this example scenario 1 converges when evaluating at least 60 simulation runs, sencario 2 after 74 and scenario 3 after 76 simulation runs. As all three scenarios shall be compared at least 76 (identical) runs have to be compared. Of course it is necessary to have a sufficient quantity of simulation runs in order to reliably determine the convergence.

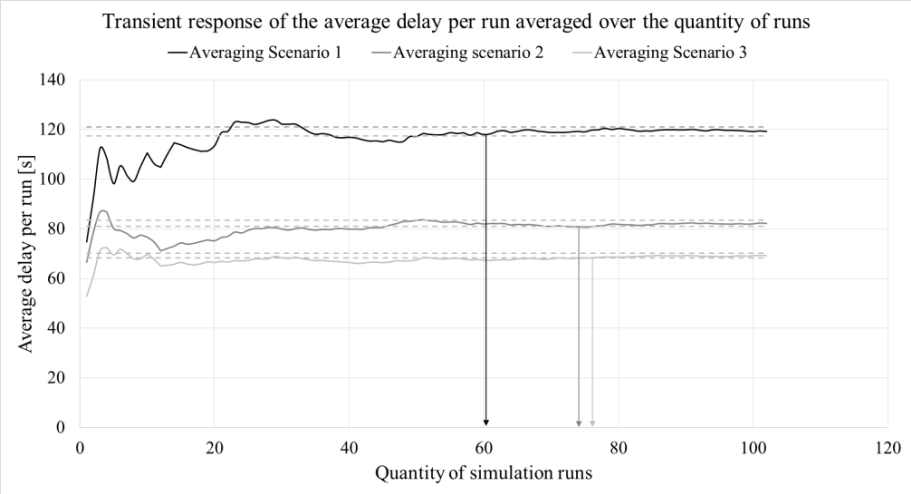


Figure 4 Exemplary convergence of the results of a simulation over the quantity of runs

After the subset of runs per series for evaluation has been identified, further checks can be performed on the layer of trains. For instance, such trains can be excluded from evaluation, which would be subject to cancellation in in-field operation. To ensure a standardised application, real-world delay threshold can be taken into account in this step.

4.3 Evaluation of Simulation Raw Data

After preparation of raw data by eliminating either whole runs or specific trains, the actual evaluation starts and key figures are aggregated. Within the presentation of results it is necessary to produce meaningful key figures for the single scenarios, which represent a general statement. For this, the previously described key figures have to be aggregated in a sound manner. It has proven reliable to interpret any key figure within its context and the return to the roots (infrastructure and timetable) to validate its statement. Results of simulations are mostly complex but by producing results for management level the complexity has to be reduced without losing important details or provoking misinterpretation. For this reason, some key figures are more able to express and to simplify the results than the others. The central issue of the evaluation of simulation results is whether scenario A or scenario B has a better operational quality. This question can be broken down by the evaluation of a model of two trains. Results of a detailed consideration of two identical trains in two different scenarios can be aggregated and for instance be averaged again for a holistic evaluation of all train runs. There are more or less four cases that describe the relation of two trains from the perspective of operational quality. The following figures underline that the expression of one isolated key figure is not necessarily in line with the overall evaluation of a scenario. Furthermore, the suitability of a key figure also depends on the target criterion to be optimized. In the standard case the operational quality shall be improved which means that the sum of delays shall be minimized. Figure 5 illustrates the delay of an identical train in two different scenarios.

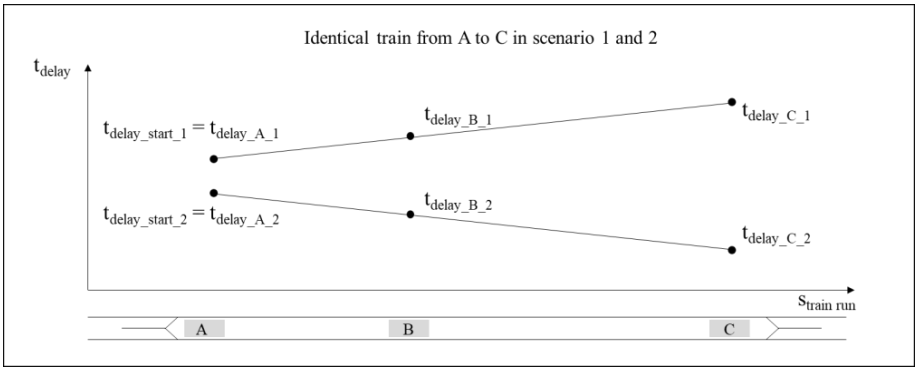


Figure 5 Case 1: Evaluation of the operational quality for a comparison of two scenarios

In the second scenario the train has continuously less delay than in the first scenario. It is obvious that scenario 2 has a better operational quality. This simple analysis is supported by the key figures introduced in paragraph 3.3 and qualitatively summarized in Table 2.

Table 2 Key figures and their statement in case 1

Key figure	**Evaluation**	**Advantage**
Absolute delay start	$t_{delay\_start\_1} > t_{delay\_start\_2}$	Scenario 2
Development of delay	$t_{delay\_development\_1} > t_{delay\_development\_2}$	Scenario 2
Absolute delay end	$t_{delay\_start\_1} + t_{delay\_development\_1} > t_{delay\_start\_2} + t_{delay\_development\_2}$	Scenario 2
TTQ Operation	$t_{simulated\_running\_time\_1} > t_{simulated\_running\_time\_2}$	Scenario 2
Average delay arrival	$t_{delay\_A\_1} + t_{delay\_B\_1} + t_{delay\_C\_1} > t_{delay\_A\_2} + t_{delay\_B\_2} + t_{delay\_C\_2}$	Scenario 2

Case 2 is an example for a situation where the reduction of the delay of a starting train may result from a longer turnaround time or a more punctual arrival from the previous ride whereas the development of delays remains identical as there were no measures met on the line. A related representation of this constellation is visualized in Figure 6.

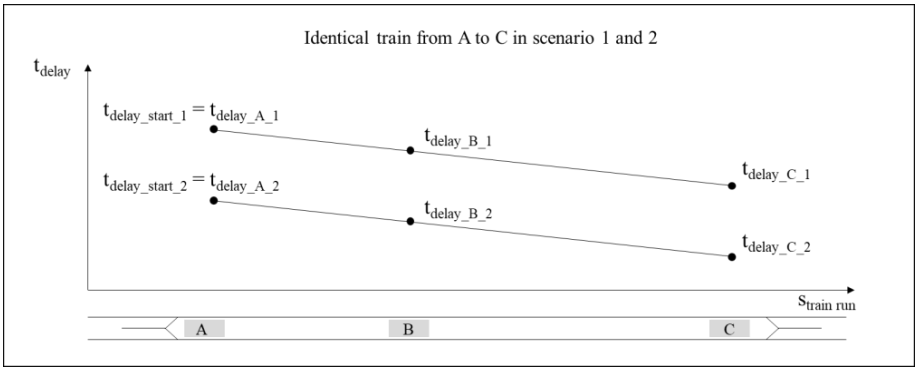


Figure 6 Case 2: Evaluation of the operational quality for a comparison of two scenarios

As the train has continuously less delay in scenario 2 it is undeniable that this scenario is preferable to scenario 1. Table 3 summarizes that the absolute delay in the first and in the last stop as well as the average arrival delay at each stop attest scenario 2 a better operational quality. At the same time the key figure "development of delay" and "TTQ Operation" are not able to detect the better scenario or even lead to wrong conclusions because these key figures only consider the development and not the absolute delay. In this case, a parallel consideration of absolute and relative delay as a combination of absolute delay at start and the development of delay could help to correctly interpret the situation.

Table 3 Key figures and their statement in case 2

Key figure	**Evaluation**	**Advantage**
Absolute delay start	$t_{delay\_start\_1} > t_{delay\_start\_2}$	Scenario 2
Development of delay	$t_{delay\_development\_1} = t_{delay\_development\_2}$	none
Absolute delay end	$t_{delay\_start\_1} + t_{delay\_development\_1} > t_{delay\_start\_2} + t_{delay\_development\_2}$	Scenario 2
TTQ Operation	$t_{simulated\_running\_time\_1} = t_{simulated\_running\_time\_2}$	none
Average delay arrival	$t_{delay\_A\_1} + t_{delay\_B\_1} + t_{delay\_C\_1} > t_{delay\_A\_2} + t_{delay\_B\_2} + t_{delay\_C\_2}$	Scenario 2

The third possible case is still evaluable by visual checking. In this case the delay of the starting train in scenario 2 is significantly reduced but therefore the delay increases over the course of the train. In reality, this may happen if the turnaround is improved like in case 2 but there is more traffic on the line so that more delays are propagated. The according developments of delay are illustrated in Figure 7.

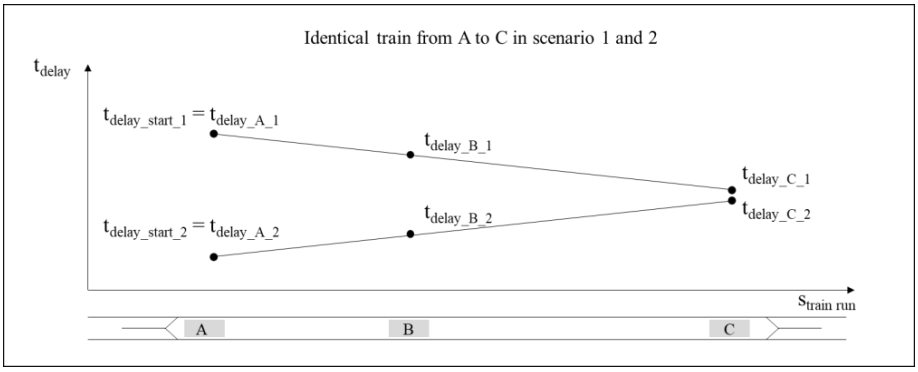


Figure 7 Case 3: Evaluation of the operational quality for a comparison of two scenarios

It is still easily evaluable that the train has a better operational quality in scenario 2 as the amount of delays is continuously smaller than in scenario 1. In this case, the standard key figures have greater difficulty in determining the better scenario than in case 1. The development of delay and the "TTQ Operation" are misleading as they only consider the relative change of delay but do not have any reference to the absolute delay. Given this it seems helpful to combine the development of delay with the absolute delay at start in order to reference to an absolute value. This value corresponds with the absolute delay at the end (compare Table 4).

Table 4 Key figures and their statement in case 3

Key figure	**Evaluation**	**Advantage**
Absolute delay start	$t_{delay\_start\_1} > t_{delay\_start\_2}$	Scenario 2
Development of delay	$t_{delay\_development\_1} < t_{delay\_development\_2}$	Scenario 1
Absolute delay end	$t_{delay\_start\_1} + t_{delay\_development\_1} > t_{delay\_start\_2} + t_{delay\_development\_2}$	Scenario 2
TTQ Operation	$t_{simulated\_running\_time\_1} < t_{simulated\_running\_time\_2}$	Scenario 1
Average delay arrival	$t_{delay\_A\_1} + t_{delay\_B\_1} + t_{delay\_C\_1} > t_{delay\_A\_2} + t_{delay\_B\_2} + t_{delay\_C\_2}$	Scenario 2

Case 4 demonstrates that even the combination of absolute delay and the development of delays reaches its limits as soon as the functions of delay intersect. This situation is a blend of cases 1 and 2 and visualized in Figure 8. This case may appear "constructed" but reality shows that a punctual departure and a punctual operation over the train run are completely decoupled and may appear in all possible combinations. From the perspective of the editor of the simulation it is not directly visible in which of the four cases the trains behave between different scenarios. Additionally the cases are mixed within a comparison of scenarios so that it is not easily identifiable which case dominates in which comparison of scenarios.

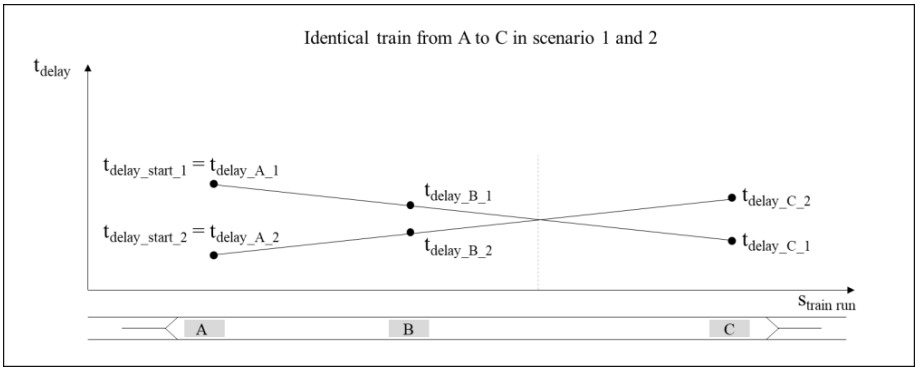


Figure 8 Case 4: Evaluation of the operational quality for a comparison of two scenarios

This setup leads to even more different statements of the key figures whereas it is not possible to evaluate this case only by taking a closer look. Three of five key figures in Table 5 indicate that scenario 1 has a better quality than scenario 2 because the train reduces its delay by making use of its supplements and arrives with less delay in its terminus. Nevertheless, the key figure average delay arrival indicates that scenario 2 has a better operational quality.

Table 5 Key figures and their statement in case 4

Key figure	**Evaluation**	**Advantage**
Absolute delay start	$t_{delay\_start\_1} > t_{delay\_start\_2}$	Scenario 2
Development of delay	$t_{delay\_development\_1} < t_{delay\_development\_2}$	Scenario 1
Absolute delay end	$t_{delay\_start\_1} + t_{delay\_development\_1} < t_{delay\_start\_2} + t_{delay\_development\_2}$	Scenario 1
TTQ Operation	$t_{simulated\_running\_time\_1} < t_{simulated\_running\_time\_2}$	Scenario 1
Average delay arrival	$t_{delay\_A\_1} + t_{delay\_B\_1} + t_{delay\_C\_1} > t_{delay\_A\_2} + t_{delay\_B\_2} + t_{delay\_C\_2}$	Scenario 2

The reason for this statement is reasonable when comparing the most important aim for each passenger namely the delay at his last stop. In Figure 9 the arrival delays of each station are accumulated. Indeed the sum of delays is smaller in scenario 2 than in scenario 1. As the number of stations is equal in both exemplary scenarios, the average behaves identical. As the number of stops may differ between different scenarios, it is recommended to evaluate the average delay at arrival instead of the sum of delay at arrival. In the comparison of all four cases the average delay arrival is the only key figures which continuously represents the results of the simulation correctly. That does not necessarily mean that the other key figures are unsuitable for representing the results of a simulation but they have to be stated carefully along with explanatory remarks so that a misinterpretation can be avoided.

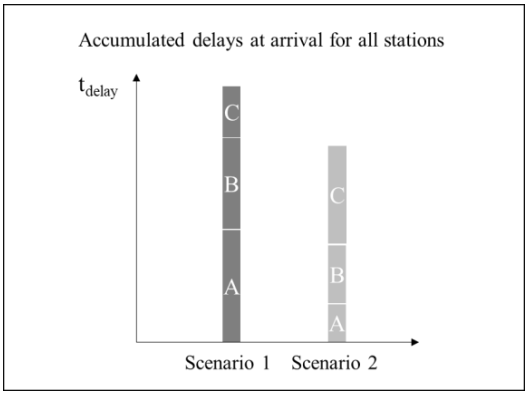


Figure 9 Comparison of accumulated delays at arrival for all stations

Furthermore, we strongly recommend analysing the deviation of the results so that the impact of outliers can be estimated. In case of a high number of outliers it can be helpful to make use of the median instead of the average. Additionally it is recommended to name quantile values in order to give a reference on the distribution of the single events. It is possible to visualize the results of the simulation as the development of delays including quantiles for each train over its train run. The visualisation of quantiles is very descriptive in diagrams as the effect of outliers can be eye-catchingly identified. In an interval timetable, the trains of a train family can be averaged as a compromise of number of diagrams and loss of information. It is also attractive to aggregate the development of delays for corridors. This aggregation has to be chosen very carefully as many effects that affect only one train family are merged into one diagram. By this, it can happen for instance that the punctuality unexpectedly rises at a junction within the corridor. In this case the delays are reduced within a corridor because the merging trains are very punctual. For this reason simulation results should not be aggregated for corridors when the share of trains changes over the considered section.

A further key figure, which is not yet discussed, is punctuality. The punctuality is probably the most famous key figure in relation to railway but it can be less meaningful than the previously discussed key figures. In most cases, the punctuality is measured in the terminus station of a train run and therefore derived from the absolute delay at the end of a train run. The same ad- and disadvantages of this previously discussed key figure remain valid for the key figure punctuality. Additionally it can easily happen that the distribution of delays changes dramatically but the punctuality remains constant. For instance, it is not unlikely that a scenario significantly reduces the number of high delays but as long as the delays are not reduced below the punctuality's threshold, they are not visible for the key figure punctuality. The same effect can happen for small delays below the threshold. For this reason, the key figure punctuality can be used for further argumentation but never as a standalone key figure to describe the results of a simulation or its operational quality.

Even more caution is required when the travel-time quotient for operation (TTQ Operation) is used for summarizing results of a simulation. In first line this key figure describes whether running time supplements can be used for the reduction of delays or not. On the first glance, a TTQ Operation smaller one suggests a better operational quality

compared to a scenario with a TTQ Operation larger or equal one because delays can be reduced. On the other hand, there is no need to reduce delays in scenario where the operational quality is comparatively high. Hence, the TTQ Operation has to be interpreted in context with a more meaningful key figure. As well, the target groups of the key figure TTQ Operation rather consists of timetable schedulers than of managers. A scheduler may adapt the timetable if the simulation reveals that running time supplements cannot be used for the reduction of delays. In a management level this key figure leads to wrong conclusions as the result of the supplements namely the variation of delay and punctuality are sufficient. Table 6 summarizes the advantages and disadvantages of the previously discussed key figures. For the target groups, "S" denotes scheduler, "E" denotes editor and "M" denotes management.

Table 6 Advantages and disadvantages of key figures

Key figure	**Strengths**	**Addressee**			**Signifi-cance**	**Interpreta-bility**	**Aggrega-teability**
		S	**E**	**M**			
Absolute delay	general optimization		X	X	++	++	+++
Delay development	disclosure of bottlenecks		X	X	++	++	+++
Graphical delay	detailed optimization		X	X	+++	+++	++
Punctuality	optimization for threshold			X	+++	++	+++
Average delay at arrival	holistic interpretation		X	X	+++	+++	+++
TTQ Operation	evaluation of reserves	X	X		++	+	+++
Infrastructure-related hindrances	disclosure of bottlenecks		X	X	+++	+++	+++

5 Conclusion

Conflict solutions of today's simulation tools are continuously approaching real world's operation. In order to draw the right conclusions of simulations it is necessary to evaluate and compare them correctly. Remaining weaknesses of simulation tools have to be detected and eliminated in the preparation of simulation raw data so that they do not affect the overall statement. Filtering out outlier simulation runs supports the evaluation of reliable and durable simulation runs. Secondly, it is important not to compare apples and oranges, which means that excluded simulation runs of one scenario have to be excluded in all other scenarios, too. The previous steps may lead to a massive reduction of evaluable simulation runs. For that reason, it is mandatory to proof the convergence of the results of each scenario for excluding the danger that the quantity of considered simulation runs affects the results. In the end it is important to interpret results of the simulation in sound manner and derive key figures which represent the results of the simulation. What might sound trivial is a rather

complex problem as most key figures can be misinterpreted when considered solitary. Graphical courses of delay support pale key figures and concentrate facts and circumstances.

6 References

Büker, Th., Seybold, B., 2012. *Stochastic modelling of delay propagation in large networks.* In: Journal of Rail Transport Planning & Management 2 (2012), pp. 34-50.

Gray, J., 2013. *Rail simulation and the analysis of capacity metrics*. In: Australasian Transport Research Forum 2013 Proceedings 2 - 4 October 2013, Brisbane.

Gröger, Th., 2002. *Simulation der Fahrplanerstellung auf der Basis eines hierarchischen Trassenmanagements und Nachweis der Stabilität der Betriebsabwicklung.* Dissertation at RWTH Aachen.

Graffagnino, T., 2012. *Ensuring timetable stability with train traffic data (Comprail),* SBB AG, Switzerland

Hansen, I.; Pachl, J., 2008. *Railway Timetable & Traffic: Analysis - Modelling – Simulation*, Eurailpress (2008).

Jensen, L.; Landex, A.; Nielsen, O. *Evaluation of robustness indicators using railway operation simulation.* In: Computer in Railways XIV (2014), pp. 329-339.

Lindfeldt, A., 2015. *Validation of a simulation model for capacity evaluation of double-track railway lines*. In: Proceedings of the 6th International Seminar on Railway Operations Modelling and Analysis (RailTokyo2015), Tokyo, Japan.

Ochiai, Y., Nishimura1, J., Tomii, N., 2014. *Punctuality analysis by the microscopic simulation and visualization of web-based train information system data*. In: Comprail (2014).

Penglin, Z., Schnieder, E., 2000. *Modelling and Performance evaluation of railway traffic under stochastic disturbances.* In: IFAC Control in Transportation Systems, Brunswick, 2000.

Weymann, F., Kuckelberg, A., Wendler, E., 2008. *Coupling of synchronous and asynchronous methods in the simulation railway operation.* In: Proceedings of the 10th International Conference on Application of Advanced Technologies in Transportation (AATT 2008).

Weymann, F., Nießen, N., 2015. *Optimisation processes to assist with fine compilation of timetables.* In: ETR International Edition 1 (2015) 1, pp. 24-27.

Optimization Model for Multi-Stage Train Classification Problem at Tactical Planning Level

Ivan Belošević [a1], Yun Jing [b], Miloš Ivić [a], Predrag Jovanović [a]
[a] Faculty of Transport and Traffic Engineering, University of Belgrade
Vojvode Stepe 305, 11000 Belgrade, Serbia
[1] E-mail: i.belosevic@sf.bg.ac.rs, Phone: +381 (0) 11 3091 201
[b] School of Traffic and Transportation, Beijing Jiaotong University
100044 Beijing, P.R. China

Abstract

Multi-stage train classification is a complex marshalling procedure that could be applied for simultaneous multi-group train formation. Simultaneous train formation is capable of processing a large volume classification insensitive on the number of outbound trains. Through multi-stage classification, wagons are moved several times to achieve desired outbound train sequences. The main optimization issue refers to finding a balance between the number of sorting steps and the total number of wagon movements. The optimization of the classification schedule could be addressed at different levels of the yard planning hierarchy. In this paper we develop mathematical formulation and two different heuristic algorithms to support tactical decisions for the multi-stage train classification problem. The main optimization issue refers to the allocation of tracks for performing multi-stage train classification minimizing annual operating costs. In order to validate the mathematical formulation and evaluate the efficiency of the proposed optimization model, we conduct computational tests and case study experimentations based on infrastructural and operational conditions applied in Belgrade marshalling yard in Serbia.

Keywords

Marshalling yards, Multi-stage train classification, Optimization model, Heuristic algorithms

1 Introduction

Marshalling yards, as consolidation nodes for rail wagonload transportation, play important role in railway freight networks (see e.g. Boysen et al (2012), Belošević et. al (2013) or Gestrelius et al. (2013)). Wagonload transportation, also called Single Wagon Load Service (SWL), consolidate loads composed of single wagons and wagon groups. These wagon loads are collected at different customer sidings and assembled in marshalling yards to full trains on the same routes. Multi-group trains have potential to take a substantial segment in wagonload service. Multi-group trains gather wagons into groups of wagons and serve two or more destinations. The order of the groups in such trains corresponds to the geographical disposition of destinations. The application of multi-group train service leads to the reduction of total layover time of wagons and to the concentration of shunting work on smaller number of main marshalling yards.

Multi-group trains are formed either applying single-stage or multi-stage classification procedures. Multi-stage train classification is a complex marshalling procedure and in

some recent research (Dahlhaus et al. (2010), Jocob et al. (2011), Bohlin et al. (2015) or Bohlin et al. (2018)) it is proven that belongs to the class of NP hard problems (nondeterministic in polynomial time). In the paper Jocob et al. (2011), a concise encoding of classification schedules is suggested for multi-stage classification procedure. This way of encoding is used in papers Marton et al. (2009) and Maue et al. (2009) for formulation of linear programming models applicable at the operational planning level. Generally, the models find out an optimal classification schedule with regard to the number of sorting steps as a primary objective and the number of movements as a secondary objective. The same encoding is also applied by Belošević and Ivić (2018). In contrast to above mentioned papers, Belošević and Ivić (2018) provide overall optimization simultaneously minimizing the number of sorting steps and the total number of movements. The formulated model is applicable at strategic planning level. The model integrates the creation of the classification schedule and design of sidings layout. As the multi-stage classification problem is computationally consuming problem, current literature also propose several usages of heuristics for efficient solving large scale instances (see Hauser et al. (2010), Belošević et al. (2013) or Belošević et al. (2018)).

Extending the existing research on this topic, this paper provides the optimization model applicable at tactical planning level. The main optimization issue refers to the allocation of classification tracks for performing multi-stage train classification over forthcoming period of the rail timetable validity. The model provides optimal track allocation minimizing annual operating costs. Based on the heuristic optimization approach presented in Belošević and Ivić (2018), this paper proposes two different Variable Neighborhood search (VNS) algorithms for solving large scale classification schedules. Developed algorithms perform a systematic search of variable neighborhoods either in deterministic or stochastic form. As a part of experiment evaluations, we randomly generate a set of various sized instances and analyze the performances of developed deterministic and stochastic VNS algorithms. The performance comparison of VNS algorithms is based on the objective value and running time of obtained solutions. Finally, case study experimentations are conducted to examine the efficiency of developed heuristic algorithms. The case study depicts infrastructural and operational conditions applied in Belgrade marshalling yard in Serbia.

2 Problem Statement

Multi-group trains are formed using sorting by train or simultaneous strategies. Sorting by train strategies result in a successive train formation procedure. Using a sorting by train strategy, wagons are initially sorted according to their outbound trains. After accumulating all wagons of a common outbound train, the wagons are resorted according to destinations. The duration of sorting by train formation procedures directly depend on the number of outbound trains (see e.g. Ivić et al. (2013)). On other hand, simultaneous strategies can greatly improve classification process, as they enable in parallel formation of several trains. Using a simultaneous train formation strategy, the sorting procedure is altered. Initially, wagons are sorted according to sorting blocks that encompass all groups with the same disposition in all outbound trains. Afterward, the wagons are resorted according to their target trains. This alternation makes the sorting procedure insensitive to the number of outbound trains and therefore capable of processing a large volume classification (see e.g. Belošević et al. (2012)).

Simultaneous formation strategies are based on the multi-stage classification procedure. The multi-stage classification includes iterative repetition of operations:

wagons pulling out from the track (pulling out operation) and wagons rolling in other tracks (disassembling operation). The sorting process starts with wagons pulling out from the first track in sidings for wagons accumulation and follows with their disassembling to other tracks. This process is repeated at all other tracks. At the sidings for wagons accumulation, sorting is performed according to the sorting blocks, while sorting according to outgoing trains is performed at the sidings for final train formation (Figure 1).

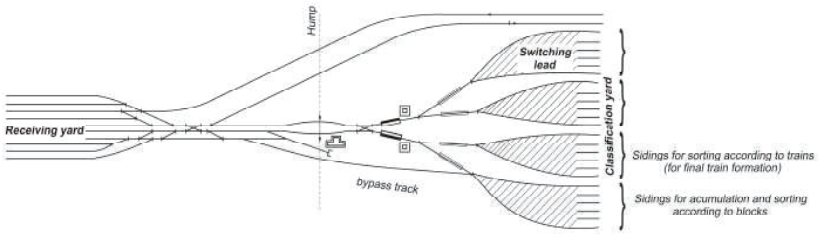


Figure 1: Sorting sidings layout

Through multi-stage classification, wagons are moved several times to achieve a desired order of wagons in an outbound train. The main optimization issue refers to finding a balance between the number of sorting steps and the total number of wagon movements indicating the length and complexity of the classification schedule. The optimization of the classification schedule should respect operational constraints and practical restrictions on the number and capacity of tracks. This optimization could be implemented at different levels of planning hierarchy (operational, tactical or strategic).

In this paper, we consider tactical planning level with the main optimization issue to allocate tracks that will be used for daily multi-stage train classification performed over forthcoming period of the rail timetable validity. Commonly once a year, marshalling yard managers make a decision referring to the distribution of classification tracks in the yard among different marshalling procedures (e.g. single-stage classification, multi-stage classification, empty wagons accumulation etc). In that sense, classification tracks have to be allocated in the way to minimize annual operating costs. As tactical decisions affect organizational processes, they have to be in accordance with the projected volume of daily work.

3 Optimization model

In this section we propose a mathematical programming formulation and a heuristic approach for solving the problem based on Variable Neiboruhood Search (VNS) strategy.

3.1 Mathematical formulation

The formulated mathematical model is based on the binary interpretation of multi-stage classification schedules proposed in Jacob at al. (2011). Let us assume that an arbitrary arranged ingoing sequence of N wagons have to be sorted into a properly sorted sequence $W = \{w_1, \ldots, w_N\}$ according to sorting blocks gs ($s = 1, \ldots, g_{max}$) where g_{max} denotes the

maximum number of groups among all outbound trains. Sorting blocks gather all groups with the same disposition in all outbound trains.

The model minimizes annual operational costs TC_{oper} and could be formulated as follows:

$$\min \ TC_{oper} \tag{1}$$

subject to:

$$\sum_{j=1}^{K} 2^{j-1} x_i^j - \sum_{j=1}^{K} 2^{j-1} x_{i-1}^j \geq 1, \qquad \forall\, i \in F \tag{2}$$

$$\sum_{j=1}^{K} 2^{j-1} x_i^j - \sum_{j=1}^{K} 2^{j-1} x_{i-1}^j \geq 0, \qquad \forall\, i \in W \backslash F \tag{3}$$

$$h_j - x_i^j \geq 0, \quad \forall\, j \in \{1, \dots, K\}, i \in \{1, \dots, N\} \tag{4}$$

$$h_j - h_{j+1} \geq 0, \qquad \forall\, j \in \{1, \dots, K-1\} \tag{5}$$

$$\sum_{i=1}^{N} x_i^j \leq C, \qquad \forall\, j \in \{1, \dots, K\} \tag{6}$$

where i and j are indices of wagons and sorting steps; F is a subset of the set W whose elements are starting a new group of wagons with the same sorting block index; K is upper bound on the number of tracks that technically could be allocated for multi-stage classification; and C indicates the track capacity limit expressed in the number of wagons.

Decision variables of the model are binary variables x_{ji} and h_j such that: $x_{ij} = 1$ if the wagon w_i participates in the sorting step h_j, or 0 otherwise; $h_j = 1$ if the sorting step h_j is realized, or 0 otherwise.

The objective function in the proposed model minimizes annual operating costs for performing multi-stage classification. The annual operating costs TC_{oper} refer to the costs of wagons' layover and fuel consumption. These costs are calculated on the daily base as a function of classification processing time. Specifically, operating costs could be estimated as a function of the number of sorting steps and total number of movements based on the classification processing time parameters t_h and t_x. The wagon layover cost is weighted with a freight rate e_w, while the fuel consumption is weighted with a fuel price e_c, fuel consumption rate c and the power of the engaged locomotive E_l. Finally, annual operating costs TC_{oper} could be expressed as:

$$TC_{\text{oper}} = 365 \left(\frac{e_w}{60} N + \frac{e_c\, c}{60} E_l\right) t_h \sum_{j=1}^{K} h_j \ + 365 \left(\frac{e_w}{60} N + \frac{e_c\, c}{60} E_l\right) t_x \sum_{j=1}^{K} \sum_{i=1}^{N} x_i^j \tag{7}$$

Constraints (2) and (3) define the schedule of wagon sorting in compliance with the basic principal of binary encoding. Specifically, wagons with a higher block index have higher code value (2), otherwise wagons with the same index may be assigned with the same code value (3). Constraints (4) and (5) define the relations between decision variables. Constraint (6) specifies the track capacity restriction.

3.2 Heuristic approach

VNS meta-heuristic is developed by Mladenović and Hansen (1997) and is based on the idea of systematic changes of neighborhood structures during a local search. The VNS

strategy starts with an initial solution and performs local search procedure within $N_k (1 \leq k \leq k_{max})$ sequential neighbourhoods. If local search results with an improvement, the solution is updated and the procedure is repeated to the new incumbent solution. The final solution presents a local minimum with respect to all k_{max} neighborhoods.

The initial classification schedule X is constructed based on triangular sorting (see Daganzo et al. (1983)). Transformations used to generate candidate schedules within the exploration of neighborhood are presented and explained in details in Belošević and Ivić (2018). The local search procedure performs an iterative evaluation of neighbors and strictly a better solution is returned as the incumbent for the succeeding search. The best improvement is used as a selection mechanism that returns into an incumbent a solution which results in the maximal improvement among all neighbors. We implement Variable Neighborhood Descent (VND) and Reduced Variable Neighborhood Search (RVNS) as two heuristics that differ in the applied strategy of exploring neighborhood structures.

VND is a deterministic heuristic that intensifies a search aiming to improve the solution greedily. Unfortunately, a persistent search of large neighborhoods is mostly time consuming. An attractive approach for improving the performance of VND is to keep the use of large neighborhoods but to reduce the exploration. RVNS is a stochastic heuristic which randomly selects points in a neighborhood and then updates an incumbent solution in the case of an improvement. RVNS diversifies a search aiming to disable improvement stagnation in broad neighborhoods. Steps of the VNS meta-heuristic strategy applied for multi-stage classification problem are presented in Figure 2.

Initialization
Select the set of neighborhood structures that will be used in local search
Find an initial classification schedule X
Repeat the following sequence until no improvement is obtained:
(1) Set $k \leftarrow 1$;
(2) *Repeat* the following steps until $k = k_{max}$:
(a) *Exploration of neighborhood*
Find the best neighbor X' of X $(X' \in N_k)$
(b) *Move or not*
If the obtained solution X' is better than X: set $X \leftarrow X'$ and $k \leftarrow 1$;
Otherwise: set $k \leftarrow k + 1$;

Figure 2: Steps of the VNS meta-heuristic

4 Numerical experiments

In order to evaluate the efficiency of developed VNS algorithms, we conduct numerical experiments. All numerical experiments are done on a working station featuring *Intel Core 2.30 GHz* Dual Core processor with 8GB of RAM memory. The proposed algorithms are coded in *Python 2.7.11*. The exact solving is performed using *CPLEX 12.6* with the running time restriction and memory restriction for the tree structure sets to 3,600 seconds and 8 GB, respectively.

The input data used in experiments are summarized in Table 1 citing costs and other relevant parameters reported in Belošević and Ivić (2018).

Table 1: Input data

Parameters	Value
Freight rate per wagon	1.3 [$/h]
Fuel price	0.3 [$/kg]
Specific fuel consumption	50 [g/kWh]
Locomotive power	650 [kW]
Classification processing time	$19.1h + 0.7x$ [min]
Track capacity	40 [wagons]
Upper bound on the number of tracks in sorting sidings	20

4.1 Computational tests

At first, we test the performance of developed VNS algorithms and show how they behave for a varied set of classification task examples. We vary the total number of wagons (100, 150 and 250 wagons) and the number of sorting blocks (6, 8 and 10 groups in outgoing trains). Combining different number of wagons and number of blocks, we obtain examples with the wide range of complexity. Three random instances are created per each example generating 27 instances in total.

All instances are first computed using the CPLEX solver and then evaluated by VND and RVNS heuristics. Due to the stochastic nature of RVNS, the evaluation procedure is repeated 50 times. The stopping condition for the evaluation is set as a function of N and g_{max}, specifically $t_{max} = Nlog_2 g_{max}$.

Results from computational tests are presented in Appendix. Appendix presents the objective value and running time of obtained solutions. Optimal solutions obtained by CPLEX are marked with bold objective function values. Non-bold values refer to the solutions returned in the moment when one of prescribed restrictions is reached, either regarding running time or memory tree. RVNS outputs are indicated by the best and average objective values, the standard deviation of objective values and the average running time. Also, we report a gap obtained by each algorithm on each instance with respect to the CPLEX value.

For all instances with 100 and 150 wagons in inbound flow, CPLEX proves optimality of returned solutions. For these sets of instances, CPLEX running times vary in range from 7 to 2200 seconds. For all instances with 250 wagons, CPLEX fails to prove the optimality of the returned solution. On other hand, VND and RVNS algorithms reach optimal or close optimal solution (with gap lower than 1%) for all instances with 100 and 150 wagons. In several instances, RVNS returns solutions with slight standard deviations due to the stochastic exploration of neighborhood structures. Due to the small neighborhood structures, running times of VND and RVNS algorithms are small and close to each other in most instances with 100 and 150 wagons. Computational tests demonstrate the quality of solutions returned by developed heuristic algorithms for the set of instances with 250 wagons, too. For the most of instances, developed heuristic algorithms return solutions that reach the best objective value. Analyzing the objective values, we can conclude that VND and RVNS algorithms perform almost similar. The minimal values for RVNS solutions are for the most instances equal to the objective values for VND solutions. Furthermore, RVNS solutions have a narrow spread around average values for all instances, so the highest gap for VND is 1.4%, while for RVNS is 1.8%. Finally, RVNS drastically outperforms VND considering running times. Running times for the VND algorithm vary in wide range reaching in some instances more than 30 minutes. On other hand, running times for the RVNS algorithm are mostly below one minute.

4.2 Case study experimentations

In order to construct meaningful experimentations we create a case study that depicts infrastructural and operational conditions applied in Belgrade marshalling yard. Belgrade marshalling yard is main marshalling yard on Serbian railway network. The layout of the yard is shown in the Figure 3. Belgrade marshalling yard features a hump with two parallel tracks. Receiving yard consists of 14 tracks with length in range from 680 to 841 meters. Classification yard consists of 48 tracks arranged in 6 groups with 8 tracks. The length of classification tracks chosen to conduct experimentations ranges from 850 to 1137 meters. Regarding the operation, Belgrade marshalling yard performs primarily single-stage classification. Multi-group trains are formed one by one within secondary sorting, grouping wagons for the same destination on a separate track. In this case study we analyze possibilities for applying the simultaneous train formation strategy. The projected volume of daily work in the forthcoming one year period of timetable validity is presented in Table 2 depicting the classification work on forming a set of outbound trains with maximum 10 groups per train. Each sorting block is assigned with the gamma distributed number of wagons with a rate set to 1.

We analyze the results obtained by developed VNS heuristic algorithms once again. Due to difficulties to use CPLEX for large scale examples, we compare obtained results with the results obtained by elementary simultaneous strategy, sorting by block. This strategy sorts wagons with same block number at a separate track and is close to the applied sorting strategy in the observed yard. The procedure is iterated 250 times. In addition to standard computational outputs regarding the objective function and running time, Table 3 shows key performance indicators in order to analyze the quality of returned classification schedules. Returned classification schedules are indicated with the number of sorting steps and total number of movements.

Table 2: The projected volume of daily work

	Sorting blocks									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Number of wagons	*35*	*34*	*32*	*29*	*28*	*25*	*21*	*17*	*15*	*12*

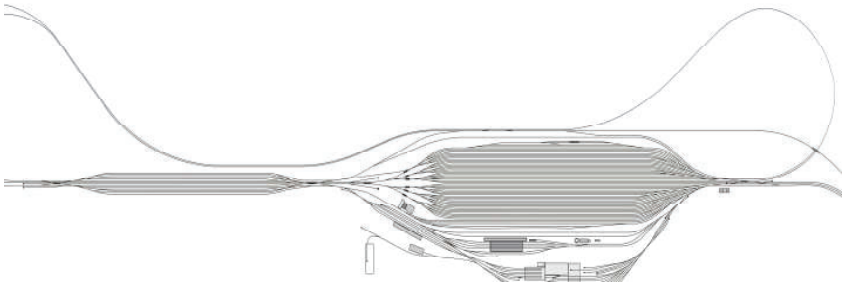


Figure 3: Layout of Belgrade marshalling yard

Table 3: Statistic summary of illustrative experimentations

	VND algorithm		**RVNS algorithm**		**Sorting by block**	
	Avg.	Range	Avg.	Range	Avg.	Range
Number of wagons	253	[208, 286]	*Same*		*Same*	
Obj. value [10^3 €]	744.8	[541.3, 974.1]	749.0	[542.7, 977.3]	776.4	[573.0, 995.6]
Gap [%]	0.0	[0.0, 0.0]	0.6	[0.0, 3.8]	4.4	[0.0, 12.8]
CPU [s]	551.7	[99.1, 2181.4]	40.9	[19.5, 216.7]	0.1	[0.0, 0.1]
Sorting steps	9	[8, 11]	9	[8, 11]	11	[10, 12]
Wagon movements	275	[234, 334]	276	[227, 337]	253	[208, 286]

Obtained results by VNS heuristics dominate over sorting by block strategy (see Figure 4). With respect to VND solutions, sorting by block strategy returns solutions with average gap 4.4 %, while the highest gap amounts up to 13 %. In contrast, developed VNS heuristic algorithms return close classification schedules with marginal differences in objective values. The average gap of RVNS solutions (with respect to VND solutions) amounts 0.6%, while the highest gap for RVNS is below 4%. Figure 5 shows the variations on average gap for solutions returned by RVNS and sorting by block. Experimentations confirm that performing complete local search of neighborhood structures could be high consuming. In that sense, VND algorithm requires averagely 551 seconds (in some instances more than 2000 seconds) for performing iterated descend. On other hand, RVNS does not have the problem with this computational issue and requires averagely about 41 seconds. Figure 6 shows these differences between VND and RVNS algorithms in terms of average running time.

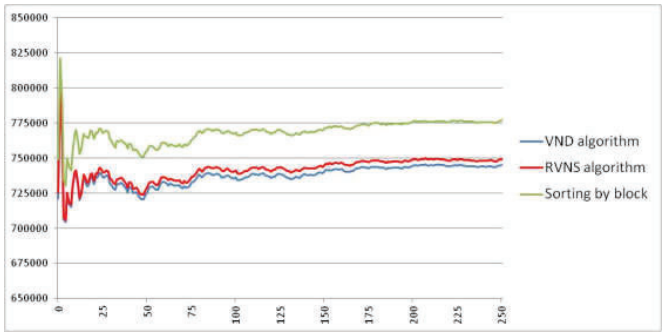


Figure 4: Average objective value

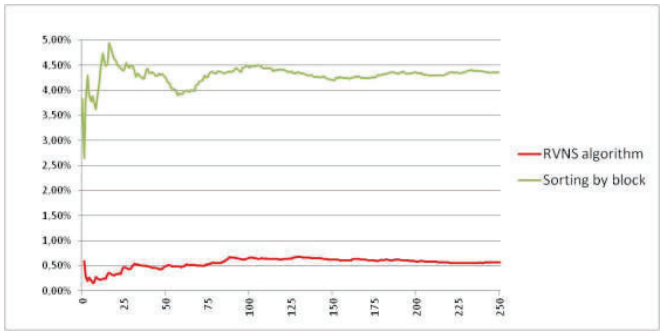


Figure 5: Average gap with respect to VND solutions

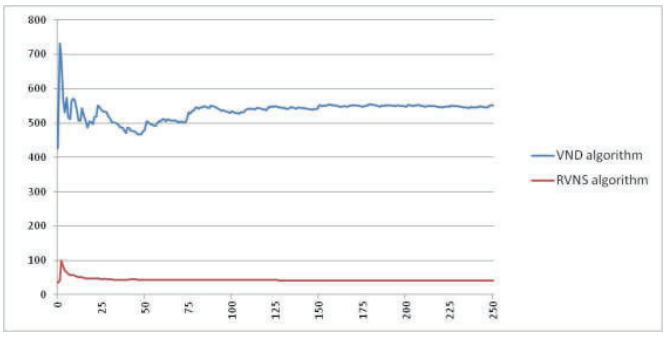


Figure 6: Average running time

Analyzing key performance indicator we can make similar conclusions regarding differences in quality of classification schedules returned by VND and RVNS algorithms. The results obtained by VND and RVNS are almost identical. The average number of sorting steps amounts 9 steps for both algorithms, while VND solutions averagely require 275 movements comparing to 276 movements in RVNS solutions. Comparing to the results from sorting by block, we can say that the heuristic optimization approach makes savings. Although the number of wagon movements is slightly increased, the average number of sorting steps is reduced for 2 steps. In order to define the required number of tracks that should be allocated for performing sorting procedure, we analyze cumulative distribution function (CDF) of the number of sorting steps. For developed heuristics the $0{,}90^{th}$ quantile amount 10 tracks, while for sorting by block amounts 11 tracks. The difference is higher considering $0{,}95^{th}$ quantile values. For developed heuristics the $0{,}95^{th}$ quantile remain 10 tracks, while for sorting by block it increases on 12 tracks.

5 Conclusions

The potential of wagonload transportation is undeniable and confirmed with the increasing interest among shippers in such services. Many environmental and economy of

scale benefits are advantages of wagonload transportation, but it needs to improve speed, reliability and cost competitiveness in comparison with other freight alternatives. In this paper, we consider multi-stage classification as one particular marshalling procedure that has the potential to improve the quality of wagonload transport allowing yards to simultaneously compound a set of multi-group trains. Multi-group trains are dominantly used for "last mile" service as local freight trains or industry trains.

In this paper we develop mathematical formulation and two different heuristic algorithms to support tactical decisions for the multi-stage train classification problem. The main optimization issue refers to the allocation of tracks for performing train classification in existing marshalling yards. Optimal track allocation is addressed by minimizing annual operating costs. Focusing on the computational complexity of the multi-stage train classification problem, in this paper we performed comparison of the efficiency of deterministic and stochastic heuristic approaches. The first one is a deterministic VND algorithm where the returned solution is a local optimum with respect to all predefined neighborhood structures. The second one is a stochastic RVNS algorithm which randomly selects points in a neighborhood and then updates an incumbent solution in the case of an improvement. RVNS diversifies a search aiming to disable improvement stagnation in large neighborhoods.

Conducted computational tests shown negligible differences between the solutions returned by developed heuristic algorithms and optimal solutions returned by CPLEX. The highest gap evaluated between the solutions returned by developed heuristics and optimal solutions amounts only 1.4%. Analyzing the objective values returned by developed heuristics, we can conclude that VND and RVNS algorithms perform almost similar. The computational tests show small dispersions of objective values evaluated by RVNS solutions. It is confirmed with low standard deviations obtained in all instances. On other hand, RVNS drastically outperforms VND considering running times. Running times for the VND algorithm reach up to 30 minutes in some instances, while running times for the RVNS algorithm are mostly below one minute. The obtained performances of developed heuristics are also confirmed in the conducted case study experiments. In the case study experiments, VNS heuristic results are compared with the results obtained by sorting by block strategy. Specifically, sorting by block strategy returns solutions with average gap 4.4 %, while the highest gap amounts up to 13 % with respect to VND solutions. These savings are confirmed analyzing key performance indicators, too.

Acknowledgements

This paper is realized and supported by The Ministry of Educations and Science of the Republic of Serbia, in the frame of the project: ''Research of technical-technological, staff and organizational capacity of Serbian Railways, from the viewpoint of current and future European Union requirements'', evidential number 36012.

References

Belošević, I., Ivić, M., Kosijer, M., Pavlović, N., 2013. "Planing Sorting Sidings using Binary Integer Programming Approach", In: *Proceedings of The 5th International Seminar on Railway Operations Modelling and Analysis (RailCopenhagen2013)*, Lyngby, Denmark.

Belošević, I., Ivić, M., 2018. "Variable Neighbourhood Search for Multi-Stage Train Classification at Strategic Planning Level", *Computer-Aided Civil and Infrastructure Engineering*, vol. 33, no.3, pp. 220-242.

Bohlin, M., Gestrelius, S., Dahms, F., Mihalák, M., Flier, H., 2015. "Optimization Methods for Multistage Freight Train Formation", *Transportation Science*, vol. 50, no. 3, pp. 523-540.

Bohlin M., Hansmann R., Zimmermann U.T., 2018. "Optimization of Railway Freight Shunting", In: Borndörfer R., Klug T., Lamorgese L., Mannino C., Reuther M., Schlechte T. (eds.), *Handbook of Optimization in the Railway Industry. International Series in Operations Research & Management Science*, vol. 268, pp.181-212, Springer, Cham.

Boysen, N., Fliedner, M., Jaehn, F., Pesch, E., 2012. "Shunting yard operations: theoretical aspects and applications", *European Journal of Operational Research*, vol. 220, no.1, pp. 1–14.

Dahlhaus, E., Horak, P., Miller, M. & Ryan, J. F., 2000. "The Train Marshalling Problem", *Discrete Applied Mathematics*, vol. 103, no. 1-3, pp. 41-54.

Gestrelius, S., Dahms, F., Bohlin, M., 2013. "Optimisation of Simulatneous Train Formation and Car Sorting at Marshalling Yards", In: *Proceedings of The 5th International Seminar on Railway Operations Modelling and Analysis (RailCopenhagen2013)*, Lyngby, Denmark.

Hauser, A., Maue, J., 2010. "Experimental evaluation of approximation and heuristic algorithms for sorting railway cars". In: Festa, P. (eds.), *Experimental Algorithms. SEA 2010. Lecture Notes in Computer Science*, vol. 6049, pp. 154-165, Springer, Heidelberg.

Jocob, R., Márton, P., Maue, J., Nunkesser, M., 2011. "Multistage methods for freight train classification", *Networks*, vol. 57, no. 1, pp. 87-105.

Maue, J., Nunkesser, M., 2009. Evaluation of computational methods for freight train classification schedules. Tech. Rep. TR-0184 ARRIVAL.

Marton, P., Maue, J. & Nunkesser, M. (2009), An improved train classification procedure for the hump yard Lausanne triage, In: *Proceedings of the 9th Workshop on Algorithmic Approaches for Transportation Modeling, Optimization, and Systems (ATMOS 09)*, Copenhagen, Denmark.

Mladenović, N., Hansen, P., 1997. "Variable Neighborhood Search", *Computers & Operations Research*, vol. 24, no. 11, pp. 1097-1100.

APPENDIX

Results from computational tests

No. wagons	Groups	Instance	CPLEX		VND			RVNS				
			Obj. value [$]	CPU [s]	Obj. value [$]	Gap [%]	CPU [s]	Min. obj. value $]	Aritm. obj. value [$]	Standard deviation	Gap [%]	CPU [s]
100	6	1	***142172,8***	*16,8*	*142172,8*	*0,0*	*11,0*	*142172,8*	*142232,3*	*181,4*	*0,0*	*8,0*
		2	***140983,8***	*13,7*	*140983,8*	*0,0*	*2,9*	*140983,8*	*140983,8*	*0,0*	*0,0*	*3,3*
		3	***143956,4***	*28,3*	*143956,4*	*0,0*	*10,7*	*143956,4*	*146150,4*	*2557,4*	*1,5*	*6,5*
	8	1	***148117,9***	*32,2*	*148117,9*	*0,0*	*11,1*	*148117,9*	*152958,9*	*4923,8*	*3,3*	*8,4*
		2	***152279,5***	*8,0*	*152279,5*	*0,0*	*4,5*	*152279,5*	*152279,5*	*0,0*	*0,0*	*5,0*
		3	***159583,5***	*21,5*	*159583,5*	*0,0*	*12,2*	*159583,5*	*160059,1*	*571,5*	*0,3*	*7,2*
	10	1	***165528,6***	*84,0*	*166123,1*	*0,4*	*8,3*	*166123,1*	*166202,3*	*258,1*	*0,4*	*7,4*
		2	***170284,7***	*22,7*	*170284,7*	*0,0*	*8,7*	*170284,7*	*170284,7*	*0,0*	*0,0*	*6,0*
		3	***165528,6***	*54,9*	*166123,1*	*0,4*	*8,7*	*166123,1*	*166281,6*	*346,8*	*0,5*	*7,3*
150	6	1	***273248,3***	*116,5*	*273248,3*	*0,0*	*49,6*	*273248,3*	*274293,5*	*772,3*	*0,4*	*13,2*
		2	***273248,3***	*98,9*	*273248,3*	*0,0*	*50,4*	*273248,3*	*274612,9*	*1706,9*	*0,5*	*13,0*
		3	***272999,4***	*17,6*	*273248,3*	*0,1*	*51,2*	*273248,3*	*275222,6*	*1885,3*	*0,8*	*13,3*
	8	1	***299378,6***	*1110,7*	*299378,6*	*0,0*	*51,2*	*299378,6*	*300046,4*	*747,6*	*0,2*	*20,4*
		2	***295894,5***	*130,7*	*295894,5*	*0,0*	*47,6*	*295894,5*	*296446,2*	*535,6*	*0,2*	*18,3*
		3	***297636,6***	*164,1*	*297636,6*	*0,0*	*42,6*	*297636,6*	*298043,0*	*676,0*	*0,1*	*19,5*
	10	1	***317669,8***	*836,3*	*317669,8*	*0,0*	*22,0*	*317669,8*	*317669,8*	*0,0*	*0,0*	*11,6*
		2	***308959,7***	*290,9*	*308959,7*	*0,0*	*52,4*	*308959,7*	*309743,6*	*979,8*	*0,3*	*21,6*
		3	***317669,8***	*2103,8*	*317669,8*	*0,0*	*107,3*	*317669,8*	*326151,8*	*3407,5*	*2,7*	*25,2*
250	6	1	*692475,7*	*3600,8*	*702443,8*	*1,4*	*401,0*	*702443,8*	*702491,3*	*260,0*	*1,4*	*28,7*
		2	*668267,6*	*3605,6*	*668267,6*	*0,0*	*426,3*	*668267,6*	*668789,7*	*698,0*	*0,1*	*29,5*
		3	*671115,6*	*3600,4*	*671115,6*	*0,0*	*375,2*	*671115,6*	*671447,8*	*612,6*	*0,0*	*26,0*
	8	1	*705698,7*	*3637,9*	*666843,5*	*-5,5*	*463,6*	*666843,5*	*668077,7*	*1661,2*	*-5,3*	*62,0*
		2	*666843,5*	*3600,7*	*666843,5*	*0,0*	*393,5*	*666843,5*	*666891,0*	*260,0*	*0,0*	*49,8*
		3	*669691,6*	*2240,2*	*669691,6*	*0,0*	*370,6*	*669691,6*	*671400,4*	*1467,5*	*0,3*	*37,9*
	10	1	*744553,8*	*3600,1*	*744553,8*	*0,0*	*1947,6*	*744553,8*	*750819,4*	*5541,7*	*0,8*	*71,7*
		2	*738450,9*	*3603,1*	*744553,8*	*0,8*	*1473,7*	*744553,8*	*746025,3*	*1650,7*	*1,0*	*65,6*
		3	*737026,9*	*3600,8*	*744553,8*	*1,0*	*1394,2*	*744553,8*	*750439,7*	*4340,8*	*1,8*	*53,0*

Strategic Planning of Rolling Stock Rotations for Public Tenders

Timo Berthold [a], Boris Grimm [b], Markus Reuther [c],
Stanley Schade [b,1], Thomas Schlechte [c]
[a] Fair Isaac Germany GmbH c/o Zuse Institute Berlin
Takustr. 7, 14195 Berlin, Germany
[b] Optimization, Zuse Institute Berlin
Takustr. 7, 14195 Berlin, Germany
[1] E-mail: schade@zib.de, Phone: +49 (0) 30 84185 336
[c] LBW Optimization GmbH
Obwaldener Zeile 19, 12205 Berlin, Germany

Abstract
Since railway companies have to apply for long-term public contracts to operate railway lines in public tenders, the question how they can estimate the operating cost for long-term periods adequately arises naturally. We consider a rolling stock rotation problem for a time period of ten years, which is based on a real world instance provided by an industry partner. We use a two stage approach for the cost estimation of the required rolling stock. In the first stage, we determine a weekly rotation plan. In the second stage, we roll out this weekly rotation plan for a longer time period and incorporate scheduled maintenance treatments. We present a heuristic approach and a mixed integer programming model to implement the process of the second stage. Finally, we discuss computational results for a real world tendering scenario.

Keywords
Rolling Stock Rotation Planning, Strategic Planning, Longterm, Maintenance

1 Introduction

Due to structural changes in the railway system all over Europe, the operation of trains does not lie solely in the hands of single state owned railway companies like it used to be. Instead railway companies compete in public tenders to receive public contracts to operate for example a certain railway line or the public railway network of a city, see Schlechte (2012) and Abbink et al. (2018). This market forces the companies to act as cost-efficient as possible to outperform others in the competition for public contracts. Since the planning of railway operations by a single company is already a complex problem, planning in a segregated market with several stakeholders becomes even more elaborate.

One of the arising problems is the estimation of costs to operate a railway enterprise for long-term periods. If a private company participates in a public tender for the operation of a railway line, it is critical to make a cost-effective offer, and, thus, to estimate the cost for all operational aspects. In this paper we analyze the costs for the rolling stock rotations including short-term, mid-term, and long-term maintenance schedules. A special feature of

this analysis is that we cannot solely rely on known optimization techniques or the planning of rolling stock rotations, because they are not suitable for periods of several years. To the knowledge of the authors the rolling stock rotation problem has not been addressed in the literature on such a long-term scale.

The objective of this paper is to obtain a vehicle and maintenance schedule for a given fleet on a given timetable that is to be operated and maintained over the course of ten years. The interesting aspect in this exposition is the presence of two scales. While the vehicles travel only several thousand kilometers per week, the maintenance procedures have to be performed after several ten thousand kilometers. Hence, they are not performed every week, and cannot be included into a cyclical weekly rotation plan. Over the course of several years it is still necessary to determine when and how often these maintenances have to be performed and it may also be of interest whether they can be distributed evenly in order to avoid peak workloads at the maintenance facilities.

In the following section we give a detailed description of the problem and our solution approach which is divided into two stages. We present the algorithmic details to solve the second stage separately in Section 3. Finally, we discuss the computational results that we obtained and summarise our findings.

2 Methodology

2.1 Problem and Data

As a problem, we use an anonymized real-world instance provided by the TransDev GmbH. The task is to plan the rolling stock rotations on a regional railway line between three cities GB, WB and HAM that are $32km$ and $42km$ apart from each other. The data specify a weekly timetable. There are 41 trips in each direction, which can be valid for different days of the week, with varying required passenger capacities. Furthermore, fuelling and maintenance intervals are provided and there are two kinds of vehicles available. One kind has 400 seats available, the other kind has 200 seats. All vehicles can be coupled to increase the capacity or to reduce the number of deadhead trips. Except for Saturday and Sunday there are usually no trips between 1am and 5am in the morning. Since only the timetable was provided, the cost for the vehicles, trips and so on had to be estimated. It was also given that refuelling has to occur every $1{,}000km$ and an IS maintenance has to occur every $40{,}000km$. The duration of the refuelling was assumed as $15min$. The duration of the IS maintenance depends on the level. It was estimated to take at least 10 hours for level 1. At multiples of $40{,}000km$ a higher level IS maintenance is required that may take more time. The highest level is 5, necessary after $640{,}000km$, which was estimated to take full 24 hours.

2.2 Approach

Due to the dual scale of the problem, we decompose it into two stages and use a sequential optimization approach. We will calculate a weekly rotation plan in the first stage and take care of the long-term maintenances in the second stage, where we track every single vehicle with the passage of time. If necessary or beneficial, we will break the rotation plan of the first stage apart and reassemble it to suit our needs.

First Stage
We use the rolling stock rotation optimizer ROTOR (Borndörfer et al. (2016); Reuther (2017)) to calculate a one week rotation plan for the first stage. For this part the long-term maintenance intervals are left out. ROTOR transforms the input timetable into a graph where trips are vertices that have to be connected by arcs that represent turns and empty trips, i.e. ROTOR models the problem using a mixed integer linear programming approach. To reduce the problem size the connecting arcs are created dynamically in the solution process. The pricing step that determines which arcs will be included into the problem is guided by a coarser version of the problem that simplifies the problem to an assignment problem. Furthermore, ROTOR uses heuristics like rapid branching to find primal solutions faster.

Table 1: Solution info

cycles	3
trips/turns	764
vehicles	4
total trip distance	27,974.31km
deadhead trip distance	242.83km
deadhead trips	6
solution time	47min
gap	0.00%

For the given timetable, ROTOR achieved an optimal solution using 4 of the vehicles with a capacity of 400 passengers each although vehicles with 200 passengers were available as well, see Table 1 for some statistics. The higher vehicle capacity was accounted for by a cost factor of 1.5. It was possible to include a 24 hour idle time for one vehicle on Sunday, since the trip density on that day was lower. This time is sufficient to include any of the required maintenances. The second longest idle time was less than 8 hours and, therefore, not even sufficient for an IS level 1 maintenance. ROTOR could be set up to include the refuelling in its calculations. However, it is easier to verify that there are sufficiently many time windows of $15min$ or longer in a postprocessing step, which we have done.

Since we want to roll out the rotation plan in the next phase, it is of importance that ROTOR provides a cyclical rotation plan. Apart from that it is conceivable to use other methods for the first stage, e.g. the method proposed by Frisch et al. (2018).

Second Stage
In the second stage, we intend to roll out the weekly rotation plan on a long-term planning interval of ten years and consider different strategies to incorporate the required maintenance treatments. A schematic drawing of the rotation plan is given in Figure 1 and the respective distances in Table 2. It can be seen that the cyclical rotation plan consists of 3 cycles. The opportunity for a maintenance lies at A_3 in the first cycle. Hence, only a single vehicle could undergo the required IS treatments. In order to enable the maintenance of the other vehicles, the rotation plan has to be modified. Our approach to connect every vehicle to a maintenance opportunity is to possibly swap vehicles if they are located at the same station overnight. It turns out that from Monday to Saturday either the vehicles on A and C or the vehicles on B and D are both located in GB. We ruled out Saturday morning, since

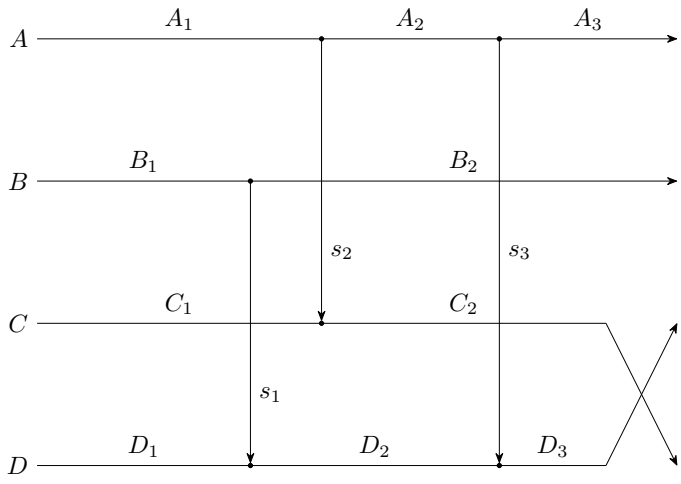


Figure 1: Schematic representation of the rotation plan computed in the first stage. The four rows labeled from A to D represent possible sets of schedules during a week for a vehicle. A vehicle that performs A or B will perform the same schedule in the next week. A vehicles that performs C will perform D in the following week and vice versa. The capital letters denote the various schedules/distances to be covered by the vehicles. The indexed capital letters denote subschedules and subdistances. Since C and D form a cycle together, the rotation plan consists of the 3 cycles in total. Since the vehicle on schedule B and D are both located in GB on Wednesday night, they can be exchanged at this point. The same applies to A and C on Thursday night and A and D on Sunday morning. These possibilities for switching the vehicles are denoted by s_1, s_2 and s_3.

the operations from Friday night to Saturday are more or less continuous. That would leave two options to swap between A and C or B and D. In order to keep the approach as simple as possible (also with regard to the practical implementation), we decided to take the latest possible options on Thursday and Friday only. These are denoted as s_1 and s_2 in Figure 1. Furthermore, at the beginning of the 24 hour idle time in GB, the vehicles on A and D are both located in GB. Therefore, either of them could idle or be maintained, while the other does the schedule D_3. This opportunity for a swap is denoted as s_3. Since railway planners are often interested in single cycles in one rotation, it is worth mentioning that by switching vehicles at s_1 and s_3 every time that these switches are possible, one can transform the three cycles into one single cycle.

What is now left is to determine an actual rotation plan that spans several weeks. This is the topic of the next section.

Table 2: Distances (in km) of vehicle schedule

A_1	3,830.400	B_1	3,222.030	C_1	4,296.040	D_1	2,947.070
A_2	2,316.580	B_2	4,261.810	C_2	3,304.210	D_2	3,147.760
A_3	0	B	7483.840	C	7,600.250	D_3	891.240
A	6,146.980					D	6,986.070

3 Solution Approaches

3.1 Backtracking Heuristic

Since the rotation plan from the first stage does not assure that every vehicle can be maintained, we have to modify it on a week by week basis. We have the option s_1 to intersect the paths B and D, s_2 to intersect A and C, and s_3 to intersect A and D. If we want to maintain the vehicle that starts the week on path B, we can apply the intersections s_1 and s_3 simultaneously. To determine which option should be applied in which week, such that every vehicle is maintained in time, we suggest a backtracking algorithm. For a given number of weeks, it performs a depth first search. The algorithm decides on one of the following options for a given week:

- run the rotation plan as determined by the first stage,
- apply the intersections s_1 and s_3,
- apply the intersection s_2,
- apply the intersection s_3 without s_1.

Furthermore, it decides if a maintenance shall be performed during said week. The general idea of backtracking is to have the algorithm select one option for a week and, then, to advance to the next week. This process is iterated until the desired number of weeks is reached or the maintenance interval of a vehicle is violated. In the first case, the algorithm has successfully determined a feasible solution and terminates. In the second case, it moves one week back and determines whether there is another choice out of the above options that was not already tested and test it, if so. If no options are left, it moves one more week back and so forth. If the options of the first week have run out, the algorithm has enumerated all possibilities and it will detect that there is no feasible solution. In the case of success, it will return the first solution that comes along.

We have two ways to control the heuristic. The first way is the range of decisions that we allow for every week. The options given above were selected, such that every vehicle could be maintained at the end of the week, regardless whether it started the week on A, B, C or D. The second way of control is the order that the algorithm uses to test the various options. We have used the order as stated above. Each of the options is tested without performing maintenance first. If that does not work, the algorithm tests the option with performing maintenance. If that does not work either, it moves on to the next option. If it was the other way, the algorithm would recommend to perform a maintenance every week, since it does not take the costs for a maintenance into account.

Table 3: Results of the backtracking heuristic. A periodic pattern that is repeated infinitely is given by weeks 5 to 9. k – number of week. v_i – position of vehicle i at the beginning of the week. block – km travelled since last maintenance. current – km travelled in the current week. total – total km travelled. m – perform maintenance at the end of the week?

k	v_1	block	current	total	v_2	block	current	total	swaps	m
0	A	0	0	0	B	0	0	0		n
1	A	6147	6147	6147	B	7484	7484	7484		y
2	A	0	6147	12294	B	14968	7484	14968	s_1, s_3	y
3	C	7038	7038	19332	A	0	6370	21337	s_1, s_3	y
4	D	14638	7600	26932	C	7038	7038	28376	s_1, s_3	y
5	B	21847	7209	34141	D	14638	7600	35976		n
6	B	29331	7483	41625	C	21625	6986	42962	s_1, s_3	y
7	A	0	6370	47995	D	29225	7600	50562	s_3	y
8	C	7038	7038	55033	A	0	6095	56657	s_1, s_3	y
9	D	14638	7600	62633	C	7038	7038	63695	s_1, s_3	y

k	v_3	block	current	total	v_4	block	current	total	swaps	m
0	C	0	0	0	D	0	0	0		n
1	D	7600	7600	7600	C	6986	6986	6986		y
2	C	14586	6986	14586	D	14586	7600	14586	s_1, s_3	y
3	D	22187	7600	22187	B	21795	7209	21795	s_1, s_3	y
4	B	29395	7209	29395	A	0	6370	28165	s_1, s_3	y
5	A	0	6370	35765	C	7038	7038	35203		n
6	A	6147	6147	41912	D	14638	7600	42803	s_1, s_3	y
7	C	13185	7038	48950	B	21847	7209	50012	s_3	y
8	D	20785	7600	56551	B	29331	7484	57496	s_1, s_3	y
9	B	27994	7209	63760	A	0	6370	63866	s_1, s_3	y

The heuristic was implemented in `Python` and it only takes a few seconds to run up to week 520. However, it is actually sufficient to run it for 14 weeks to see that the weeks 5 to 9 give a periodic pattern that is repeated subsequently. The results of the first 9 weeks are presented in Table 3. The only deviation from this pattern may happen in the last weeks where a maintenance may be left out. Since only one maintenance per week is possible, the algorithm has to plan ahead, so that only one vehicle reaches its maintenance interval and has to be maintained at a time. At the end of the time horizon, a maintenance may be left out, since the algorithm does not look ahead anymore. Thus, several vehicles can be close to their kilometer limit and require maintenance in the week after the end of the time horizon.

3.2 MILP Model

In addition to the backtracking heuristic, we have come up with a Mixed Integer Linear Programming (MILP) model. The model is basically a multi-commodity flow with one commodity per vehicle. The coupling of the commodities happens via the integral s vari-

ables that allow the transition of a flow between the four possible paths as shown in Figure 1. If an s is set to 1 for one commodity, it means that the flow of the commodity transitions downwards (A being on top and D being on the bottom). In that case, the corresponding s of another commodity has to be set to -1, which means that the flow of the corresponding commodity flows upwards.

Parameters

- N – number of weeks considered
- $A = A_1 + A_2 + A_3$, $B = B_1 + B_2$, $C = C_1 + C_2$, $D = D_1 + D_2 + D_3$ – distances of the different vehicle schedules, cp. Figure 1
- U – maximum maintenance interval in kilometers

Variables

We assume $i, j \in \{1, 2, 3, 4\}$ and $k \in \{1, \ldots, N\}$ if not specified otherwise. Also, if not specified otherwise, the variables are non-negative and continuous.

- $m_k \in \{0, 1\}$ – perform a maintenance in week k
- $s_{i,l,k} \in \{-1, 0, 1\}$ vehicle i transitions at s_l ($l \in \{1, 2, 3\}$) in week k, cp. Figure 1
- $y_{i,j,k} \in [0, 1]$ – vehicle flow of vehicle i on path j ($j = 1$ corresponds to A, $j = 2$ to B and so on) in week $k \in \{1, \ldots, N + 1\}$
- $x_{i,k} \in \mathbb{R}_+$ – kilometers travelled since last maintenance by vehicle i after week $k \in \{1, \ldots, N\}$ or before week 1 ($k = 0$).
- $t_{i,k} \in \mathbb{R}_+$ – transition variable to check that U is not exceeded
- $w_{i,k} \in \mathbb{R}_+$ – auxiliary variable to reset x if a maintenance is performed

Model

$$\min \sum_{k=1}^{n} m_k \tag{1}$$

s.t.

$$y_{i,j,1} = \delta_{i,j} \qquad \text{for } i,j \in \{1,2,3,4\} \tag{2}$$

$$y_{i,1,k+1} = y_{i,1,k} - s_{i,2,k} - s_{i,3,k} \qquad \text{for } i \in \{1,2,3,4\},\ k \in \{1,\dots,N\} \tag{3}$$

$$y_{i,2,k+1} = y_{i,2,k} - s_{i,1,k} \qquad \text{for } i \in \{1,2,3,4\},\ k \in \{1,\dots,N\} \tag{4}$$

$$y_{i,3,k+1} = y_{i,4,k} + s_{i,1,k} + s_{i,3,k} \qquad \text{for } i \in \{1,2,3,4\},\ k \in \{1,\dots,N\} \tag{5}$$

$$y_{i,4,k+1} = y_{i,3,k} + s_{i,2,k} \qquad \text{for } i \in \{1,2,3,4\},\ k \in \{1,\dots,N\} \tag{6}$$

$$\sum_{i=1}^{4} s_{i,l,k} = 0 \qquad \text{for } l \in \{1,2,3\},\ k \in \{1,\dots,N\} \tag{7}$$

$$y_{i,2,k} - 1 \leq \quad s_{i,1,k} \leq y_{i,2,k} \qquad \text{for } i \in \{1,2,3,4\},\ k \in \{1,\dots,N\} \tag{8}$$

$$y_{i,4,k} - 1 \leq -s_{i,1,k} \leq y_{i,4,k} \qquad \text{for } i \in \{1,2,3,4\},\ k \in \{1,\dots,N\} \tag{9}$$

$$y_{i,1,k} - 1 \leq \quad s_{i,2,k} \leq y_{i,1,k} \qquad \text{for } i \in \{1,2,3,4\},\ k \in \{1,\dots,N\} \tag{10}$$

$$y_{i,3,k} - 1 \leq -s_{i,2,k} \leq y_{i,3,k} \qquad \text{for } i \in \{1,2,3,4\},\ k \in \{1,\dots,N\} \tag{11}$$

$$y_{i,1,k} - s_{2,k} - 1 \leq \quad s_{i,3,k} \leq y_{i,1,k} - s_{2,k} \qquad \text{for } i \in \{1,2,3,4\},\ k \in \{1,\dots,N\} \tag{12}$$

$$y_{i,4,k} + s_{1,k} - 1 \leq -s_{i,3,k} \leq y_{i,4,k} + s_{1,k} \qquad \text{for } i \in \{1,2,3,4\},\ k \in \{1,\dots,N\} \tag{13}$$

$$\begin{aligned} t_{i,k} = {} & x_{i-1,k} + Ay_{i,1,k} + By_{i,2,k} \\ & + Cy_{i,3,k} + Dy_{i,4,k} \\ & + (D_2 + D3 - B2)s_{i,1,k} \\ & + (C_2 - A_2 - A_3)s_{i,2,k} \\ & + (D_3 - A_3)s_{i,3,k} \end{aligned} \qquad \text{for } i \in \{1,2,3,4\},\ k \in \{1,\dots,N\} \tag{14}$$

$$x_{i,k+1} = t_{i,k} - w_{i,k} \qquad \text{for } i \in \{1,2,3,4\},\ k \in \{1,\dots,N\} \tag{15}$$

$$w_{i,k} \leq Um_k \qquad \text{for } i \in \{1,2,3,4\},\ k \in \{1,\dots,N\} \tag{16}$$

$$w_{i,k} \leq Uy_{i,0,k+1} \qquad \text{for } i \in \{1,2,3,4\},\ k \in \{1,\dots,N\} \tag{17}$$

Description

The objective of the model is to perform as few maintenances as possible. The constraint (2) sets the starting positions of the vehicles. Here, $\delta_{i,j} = 1$ if and only if $i = j$, otherwise $\delta_{i,j} = 0$. The constraints (3-6) are flow conservation constraints and (7) couples the different commodities. With a pure flow model, it would be possible to have a flow back in time, e.g. a flow on D_1 and D_2 that reverses on s_3 and A_2 and then uses s_2. In order to forbid these kind of flows and keep causality in the model, we use the constraints (8-13). Moreover, (14) takes account of the travelled kilometers since the last maintenance. For example, if vehicle 2 is scheduled for A in week 3, then $y_{2,1,3}$ is 1. If the vehicle is supposed to change on schedule C during that week, then $s_{2,2,3} = 1$. Thus, the remainder of the first schedule $A_2 + A_3$ will be removed from the milage, while the second part C_2 of the third schedule will be added, cf. Figure 1. The constraints (15-17) ensure that the kilometers are reset if a maintenance is performed.

The model and the underlying problem could also be interpreted as a special case of the Train Dispatching Problem presented in Boccia et al. (2013) and Mannino (2011) or the Train Timetable Rescheduling Problem presented in Cacchiani et al. (2014). This leads to slightly different model formulations, which are comparable in the sense of tractable problem sizes.

The model was implemented and solved using the Python interface of the FICO Xpress solver version 8.5.7. The solutions of the backtracking heuristic were used as initial solutions. The tests were performed on a Dell Precision Tower 3620 with 30GB of main memory and 8 Intel® Xeon(R) CPU E3-1245 v5 @ 3.50GHz. For the instances with a time horizon of 15 or fewer weeks, it turned out that the solution provided by the backtracking heuristic was optimal, but it could take long to obtain the proof, see Table 4. For a 20 week instance, the memory did not suffice.

Table 4: Solution info

weeks	rows	cols	nodes	optimal value	time (s)
10	850	430	411,303	7	88
15	1265	511	47,045,565	11	17,867

3.3 Lower Bounds

Since proving the optimality of the heuristic solution using a general purpose MILP solver was not always possible or took long, one might try to find lower bounds oneself to directly prove the optimality of a backtracking solution or to provide these bounds to the solver. An easy bound is obtained as follows. The maintenance interval is $40{,}000km$ and the vehicles travel $27{,}974.31km$ per week (see Table 1). Hence, assuming a cyclical solution, the vehicles have to be maintained at least ceil$(N \cdot 27{,}974.31/40{,}000)$ times in an N week scenario, e.g. for a cyclical 5 week solution at least 4 maintenances are required. If the solution is not cyclic, we have to account for the fact that the vehicles could start with $0km$ travelled since the last maintenance but finish the scenario with almost $40{,}000km$ since the last maintenance. Hence, the lower bound is decreased by 4.

These bounds are not tight enough to prove the optimality of our solution for the 10 year scenario. One way to overcome this may be to use a cyclical solution. Borndörfer et al. (2008) and Borndörfer et al. (2018) argue why it is reasonable to do that for long periods. However, there are still ways to improve our bounds further. For example, we can determine the distance to the maintenance facility for every vehicle at the beginning of schedule A, B, C and D. Say a vehicle ends at position B. Then the vehicle needs to be able to travel $B_1 + D_2 = 6{,}369.79km$ before attaining its kilometer limit in order to reach the maintenance facility. (The vehicle will have to switch at s_1 and s_3 to reach the maintenance facility on Sunday.) We can determine analogous constraints for A, C and D. Using this adaption of our MILP model, we can show that if the vehicles are only maintained three times during five weeks, one vehicle will violate its maximum kilometer limit in the seventh week. To do so, we bound the maintenances in the first 5 weeks by three and solve a 6 week scenario, where we require all vehicles to be able to reach the maintenance facility in week

7. This MILP is infeasible. Hence, we know that at least four maintenances are required within every consecutive five weeks of the solution that do not include the last two weeks.

We now apply this knowledge to a solution of the backtracking heuristic for a long scenario, i.e. 517 weeks. Given any five consecutive weeks of the first 515 weeks, there have to be at least four maintenances within these five weeks. Hence, we can deduct that that in the weeks 1 to 515, $515/5 \cdot 4 = 412$ maintenances have to be performed. This way, there can be one week without maintenance in weeks 1 to 5, 6 to 10 and so on. If there are fewer maintenances, we would find an $\ell \in \{0, \ldots, 103\}$, such set there are two weeks within week $5\ell + 1$ to week $5\ell + 5$ where no maintenance is performed. Thus, one of the vehicles would violate its kilometer limit in week $5\ell + 7$ and the underlying solution would not be feasible. Hence, we obtain a lower bound of 412 maintenances for a 517 week scenario. Since the backtracking heuristic gives us a solution with 412 maintenances, we know that this solution is optimal. For the 520 week scenario, the backtracking heuristic provides a solution with 415 maintenances, but we do not have a strict mathematical proof of the optimality with this method. Obviously, there still have to be at least 412 maintenances. We can even improve this bound to 414 by arguing that there have to be at least 4 maintenances within the weeks 514 to 518. For the 522 week scenario, we know that 416 maintenances are optimal. Given the optimal solution values of 7 and 11 for $\ell = 10$ and $\ell = 15$, it is quite possible that 415 maintenances is the optimal value for the 520 week scenario, but we did not formally prove this claim.

4 Discussion

We have considered the problem of integrating a long-term maintenance schedule and a weekly rotation plan on a real-world scenario. The specific issues that had to be addressed were that

- a weekly rotation plan that includes a suitable idle time had to be found,
- the determined weekly rotation plan did only allow for the maintenance of one vehicle per week,
- the weekly rotation plan did consist of several cycles, of which only one contained the necessary idle time and
- that the time scale of the long-term maintenances made planning over several weeks necessary.

We recognized that the lower density of the timetable on Sundays makes it possible to include a 24 hour idle time for one vehicle and introduced transitions between the cycles of the rotation plan, so that every vehicle can undergo the long-term maintenance treatments. We provided a practical algorithm that quickly calculates a rotation plan on a scale of several weeks or even years, which distributes the required maintenances uniformly and even determines gaps in the maintenance calendar.

We could proof the optimality of the obtained solutions for a time horizon of up to 15 weeks. For longer time horizons, we obtained the optimal solutions for scenarios of length $k = 5\ell + 2$ for a positive integer ℓ. For k we found solutions that are at least close to the optimum, but we did not prove their optimality.

Altogether, we provide a starting point for an estimate of the cost for a 10 year railway enterprise.

Acknowledgement

This work has been supported by the Research Campus MODAL Mathematical Optimization and Data Analysis Laboratories funded by the Federal Ministry of Education and Research (BMBF Grant 05M14ZAM). The authors would like to thank Christof Schulz and Steffen Weider of LBW Optimization GmbH for providing the problem and dataset. Further thanks goes to Julian Bushe, as well as Alistair Benford, John Curtis Lynch, and Ariel Nikas for assisting in the preliminary work for this article during the G-RIPS program 2018. Also, thanks to one of the anonymous referees for their extensive comments.

References

Abbink, E., Bärmann, A., Besinovic, N., Bohlin, M., Cacchiani, V., Caimi, G.,Dollevot, T., de Fabris, S., Fischer, F., Fügenschuh, A., Galli, L., Goverde, R.M.P., Hansmann, R., Homfeld, H., Huisman, D., Johann, M., Klug, T., Kroon, L., Lamorgese, L., Liers, F., Mannino, C., Medeossi, G., Pacciarelli, D., Reuther, M., Schlechte, T., Schmidt, M., Schöbel, A., Schülldorf, H., Stieber, A., Stiller, S., Törnquist Krasemann, J., Toth, P. and Zimmermann, U.T. 2018. *Handbook of Optimization in the Railway Industry*. Vol. 268 1 ed.
URL: *http://www.springer.com/us/book/9783319721521*

Boccia, M. and Mannino, C. and Vasilyev, I. 2013. "The dispatching problem on multitrack territories: Heuristic approaches based on mixed integer linear programming." *Networks* 62(4):315–326.

Borndörfer, R., Karbstein, M., Liebchen, C. and Lindner, N. 2018. A Simple Way to Compute the Number of Vehicles That Are Required to Operate a Periodic Timetable. In *18th Workshop on Algorithmic Approaches for Transportation Modelling, Optimization, and Systems (ATMOS 2018)*, ed. Ralf Borndörfer and Sabine Storandt. Vol. 65 of *OpenAccess Series in Informatics (OASIcs)* Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik pp. 16:1–16:15.
URL: *http://drops.dagstuhl.de/opus/volltexte/2018/9721*

Borndörfer, R., Reuther, M., Schlechte, T., Waas, K. and Weider, S. 2016. "Integrated Optimization of Rolling Stock Rotations for Intercity Railways." *Transportation Science* 50(3):863 – 877.

Borndörfer, R. and Liebchen, C. 2008. When Periodic Timetables Are Suboptimal. In *Operations Research Proceedings 2007*, ed. Jörg Kalcsics and Stefan Nickel. Berlin, Heidelberg: Springer Berlin Heidelberg p. 449–454.

Cacchiani, V., Huisman, D., Kidd, M., Kroon, L., Toth, P., Veelenturf, L. and Wagenaar, J. 2014. "An overview of recovery models and algorithms for real-time railway rescheduling." *Transportation Research Part B: Methodological* 63:15–37.

Frisch, S., Hungerländer, P., Jellen, A. and Weinberger, D. 2018. "A Mixed Integer Linear Program for Optimizing the Utilization of Locomotives with Maintenance Constraints.".

Mannino, C. 2011. Real-time traffic control in railway systems. In *11th Workshop on Algorithmic Approaches for Transportation Modelling, Optimization, and Systems*, ed. Alberto Caprara and Spyros Kontogiannis. Vol. 20 of *OpenAccess Series in Informatics (OASIcs)* Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik pp. 1–14.
URL: *http://drops.dagstuhl.de/opus/volltexte/2011/3262*

Reuther, M. 2017. Mathematical Optimization of Rolling Stock Rotations PhD thesis.
URL: *https://depositonce.tu-berlin.de/handle/11303/6309*

Schlechte, T. 2012. Railway Track Allocation: Models and Algorithms PhD thesis.
URL: *http://opus.kobv.de/tuberlin/volltexte/2012/3427/pdf/schlechte_thomas.pdf*

A Concurrent Approach to the Periodic Event Scheduling Problem

Ralf Borndörfer [a], Niels Lindner [a, 1], Sarah Roth [a]
[a] Zuse Institute Berlin
Takustr. 7, 14195 Berlin, Germany
[1] E-mail: lindner@zib.de, Phone: +49 30 84185374

Abstract
We introduce a concurrent solver for the periodic event scheduling problem (PESP). It combines mixed integer programming techniques, the modulo network simplex method, satisfiability approaches, and a new heuristic based on maximum cuts. Running these components in parallel speeds up the overall solution process. This enables us to significantly improve the current upper and lower bounds for all benchmark instances of the library PESPlib.

Keywords
Periodic Event Scheduling Problem, Periodic Timetabling, Mixed Integer Programming

1 Introduction

The optimization of periodic timetables is a major planning task in public transit. The standard mathematical model is the formulation as a periodic event scheduling problem (PESP, Serafini and Ukovich (1989)). In 2005, the first mathematically optimized timetable has been put into operation (Liebchen (2008)) on the Berlin subway network. However, computing optimal timetables on country-sized railway networks is notoriously hard. Therefore, the focus lies usually on feasibility rather than on minimum passenger travel time or other optimization goals (Kümmling et. al. (2015)). In particular, solving the rather large benchmark instances of the library PESPlib [1] to optimality seems currently out of reach.

The state-of-the-art methods for solving periodic timetabling problems comprise satisfiability techniques (SAT, Großmann et. al. (2012)) for feasibility questions, branch-and-cut in the framework of mixed integer programming (MIP, Liebchen (2006)), and the modulo network simplex algorithm (MNS, Nachtigall (1998)) as a local improving heuristic. In fact, the current best solutions to the PESPlib instances have been found by running a MIP solver and an MNS implementation alternatingly for 8 hours in total (Goerigk and Liebchen (2017)).

We introduce a new PESP solver based on concurrency, and integrating all three approaches. The core idea is to run MIP, MNS and a new maximum cut based heuristic in parallel. This way, the global nature of the search procedure underlying the MIP solver enables the other algorithms to escape local optima. Moreover, our solver features additional ingredients, e.g., a cutting plane separator for cycle and change-cycle inequalities.

In Section 2, we introduce PESP and two mixed integer programming formulations. The architecture and the key ingredients of our PESP solver are illustrated in Section 3. The

[1] available at `http://num.math.uni-goettingen.de/~m.goerigk/pesplib`

features of the PESPlib set and the current solution status is described in Section 4. The subsequent Section 5 contains the results of applying our solver to the PESPlib instances. We conclude this paper with a short summary in Section 6.

2 The Periodic Event Scheduling Problem

The *Periodic Event Scheduling Problem* (PESP), going back to Serafini and Ukovich (1989), is the default approach to model periodic timetabling problems. It seeks for an optimum periodic slack respecting certain bounds in a given network. We refer to Liebchen (2006) for an exhausting overview. Formally, the input for PESP is the following:

- a directed graph G with vertex set V and arc set A,
- a period time $T \in \mathbb{N}$,
- lower bounds $\ell \in \mathbb{Z}^A_{\geq 0}$,
- upper bounds $u \in \mathbb{Z}^A_{\geq 0}$, where $\ell \leq u$,
- weights $w \in \mathbb{Z}^A_{\geq 0}$.

In this paper, we will consider only integer bounds and weights. The graph G is usually called an *event-activity network*, where vertices are considered as *events*, and arcs as *activities*. Given a PESP instance (G, T, ℓ, u, w), a *periodic timetable* is an assignment of values in $[0, T)$ to the events, i.e., a vector $\pi \in [0, T)^V$. A periodic timetable defines a *periodic slack* $y \in \mathbb{R}^A_{\geq 0}$ via

$$y_{ij} := [\pi_j - \pi_i - \ell_{ij}]_T, \quad ij \in A, \tag{1}$$

where $[\cdot]_T$ denotes the modulo T operator taking values in the interval $[0, T)$. Intuitively, a periodic timetable π fixes the duration of an activity $ij \in A$ modulo T to be $[\pi_j - \pi_i]_T$, and the actual duration of ij is computed as the smallest number $\ell_{ij} + y_{ij} \in [\ell_{ij}, u_{ij}]$ satisfying $[\ell_{ij} + y_{ij}]_T = [\pi_j - \pi_i]_T$.

The PESP is now formulated as the following mixed integer program:

$$\begin{array}{lll} \text{Minimize} & w^t y & \\ \text{subject to} & y_{ij} = \pi_j - \pi_i - \ell_{ij} + p_{ij} T, & ij \in A, \\ & 0 \leq y \leq u - \ell, & \\ & 0 \leq \pi < T, & \\ & p \in \mathbb{Z}^A_{\geq 0}. & \end{array} \tag{2}$$

The integer variables p_{ij} for each activity $ij \in A$ are called *periodic offsets*. Their purpose is to model the modulo T conditions (1). Using the incidence matrix $A \in \{-1, 0, 1\}^{V \times A}$ of the network G, these constraints may as well be written as

$$y = A^t \pi - \ell + pT.$$

This is why we will call (2) the *incidence matrix mixed integer programming formulation* of PESP in the sequel. Since the incidence matrix of a directed graph is totally unimodular,

so is A^t, which implies that there is always an optimal periodic timetable π taking values in $\{0, 1, \ldots, T-1\}$. We may therefore interpret the constraint $0 \leq \pi < T$ as $0 \leq \pi \leq T-1$. A fortiori, there is always an integral optimal periodic slack y.

Another formulation can be obtained as follows: Suppose that the network G has m activities. A *cycle matrix* of G is a full row rank matrix $\Gamma \in \{-1, 0, 1\}^{\mu \times m}$ whose rows form a maximal linearly independent set of incidence vectors of oriented cycles in G, i.e., a cycle basis. If Γ represents an integral cycle basis, i.e., any maximal minor is either 0 or ± 1, then the following mixed integer program is equivalent to (2):

$$\begin{array}{lll} \text{Minimize} & & w^t y \\ \text{subject to} & & \Gamma(y + \ell) = zT \\ & & 0 \leq y \leq u - \ell, \\ & & z \in \mathbb{Z}^\mu. \end{array} \tag{3}$$

This is the *cycle matrix mixed integer programming formulation*. The number $\mu = m - n + c$ of equality constraints resp. integer variables is also called the *cyclomatic number* of G and serves as one measure of difficulty for PESP instances.

Example 1. Consider the PESP instance $I = (G, T, \ell, u, w)$ depicted in Figure 1. The timetable π indicated by the vertex labels has weighted slack $1 \cdot 5 + 5 \cdot 10 + 3 \cdot 25 = 130$. We will see later in Examples 2 and 3 that π is indeed optimal.

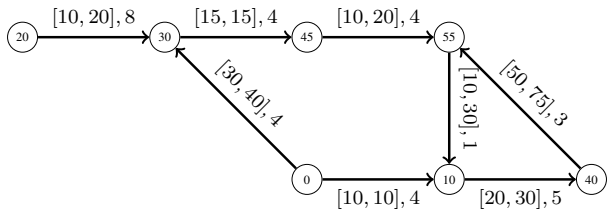


Figure 1: Example PESP instance with period time $T = 60$. The activities are labeled with $[l, u], w$. An optimal periodic timetable is given by the event labels.

3 Solver Architecture

3.1 Overview

The main idea of our PESP solver is to execute several well-performing algorithms in parallel. The solver operates in three phases, see also Fig. 2:

1. *Preprocessing phase*: Given a PESP instance, the network size is reduced by an exact preprocessing step and a subsequent heuristic preprocessing step. The exact method transforms the PESP instance into an equivalent instance – the *final problem* – with the same objective value. On the other hand, the heuristic preprocessing is allowed to slightly alter the instance and objective value, resulting in the *master problem*. The details are described in §3.2. In addition to creating the final and master problems,

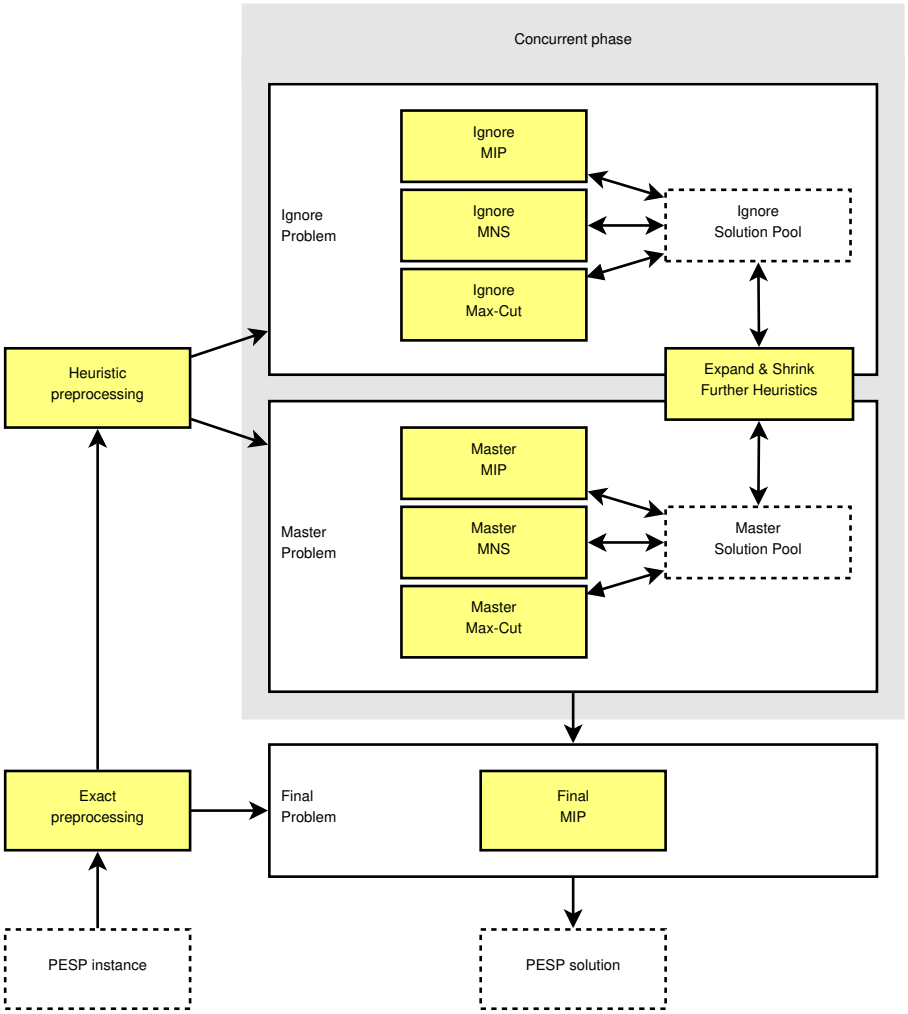


Figure 2: Architecture of our concurrent PESP solver

both preprocessing steps are also applied to a much smaller network, defining an *ignore problem*, see §3.3.

2. *Concurrent phase*: This is the main phase of the solver. Both the master and ignore problem are tackled using a MIP solver (§3.4), a modulo network simplex algorithm (§3.5), and a maximum cut heuristic (§3.6) each. These six threads run asynchronously in parallel. The incumbent solutions for each of the problems are shared among the threads by means of a common solution pool. A seventh thread transforms periodic timetables from the ignore problem's solution pool to the master problem's solution pool and vice versa, and additionally applies further heuristics (see §3.7) on both pools. While the master problem is kept for the whole concurrent phase, the ignore problem may change. The concurrent phase ends after a time limit or if the MIP solver terminates, either by detecting infeasibility or by proving optimality of the incumbent solution.

3. *Final phase*: The final problem is treated with a MIP solver, taking the best periodic timetable for the master problem as initial solution. This phase usually is aborted after a short time, as it typically does not make significant progress after the first few minutes. However, the internal heuristics of the MIP solver often detect better incumbents invisible to the master problem.

The solver is flexible in the sense that any subset of the methods in the concurrent phase can be switched off. On the other hand, the MIP solver may use several internal threads.

One advantage of concurrency is that the branch-and-cut process of the MIP does not need to wait for the other heuristics to finish and vice versa. Moreover, the communication of new solutions across all threads helps to overcome local optima.

3.2 Network Preprocessing

Our exact and heuristic preprocessing methods are based on the network size reduction strategies in Liebchen (2006) and Goerigk and Liebchen (2017). Let (G, T, ℓ, u, w) be a PESP instance. The preprocessing comprises the following steps, see also Figure 3:

1. Remove all bridges, i.e., all arcs that are not part of any oriented cycle.

2. Delete all isolated vertices.

3. Contract fixed arcs: If for an activity $a \in A$ holds $\ell_a = u_a$, then necessarily $y_a = 0$. We can hence delete a and its target j. All other arcs incident with j are replaced with arcs with the same weight from or to the source of a, adding or subtracting the lower bound ℓ_a, respectively.

4. Contract degree two vertices: If i is a vertex of degree two with an entering arc from j and a leaving arc to k, then delete i and its incident arcs, and add a new arc between j and k, adding up lower and upper bounds. The weight of the new arc becomes the minimum of the weights of the two old arcs. In exact preprocessing, this is only done if the two arcs incident to i share the same weight. However, in the case of heuristic preprocessing, also incident arcs with different weights are considered.

5. Finally, we normalize the lower and upper bounds such that both $\ell \in [0, T)$ and $u - \ell \in [0, T)$ by subtracting a suitable multiple of T. As a consequence, the periodic offsets p are then guaranteed to lie in $\{0, 1, 2\}$, reducing the size of the branch-and-bound tree the MIP solver has to search.

Both exact and heuristic preprocessing carry out all five steps. The only difference is that the heuristic preprocessing is allowed to contract more vertices of degree two. However, the preprocessing may introduce multiple arcs between two vertices.

Lemma 1. *Let $I = (G, T, \ell, u, w)$ be a PESP instance, and denote by I_{exact} and I_{heur} the PESP instances after exact and heuristic preprocessing, respectively. Then the optimal weighted slacks* OPT *satisfy*

$$\mathrm{OPT}(I) = \mathrm{OPT}(I_{exact}) \geq \mathrm{OPT}(I_{heur}).$$

Proof. As can be seen from the cycle matrix MIP formulation, there is no constraint on the periodic slack of a bridge. Hence in an optimal solution, any bridge $a \in A$ has periodic slack $y_a = 0$. Clearly, isolated vertices can be omitted. Since fixed arcs cannot have slack, they do not contribute to the objective value. Again by inspecting the cycle matrix MIP formulation, one checks that the normalization in step 5 does not affect the minimization.

It remains to check step 4. The contraction of degree two vertices does not affect the cycles of the graph, except that they contain less arcs. However, the weights may change: Let a_1, a_2 be arcs in I incident to a common degree two vertex with in-degree one and out-degree one. If their optimal periodic slacks are y_{a_1}, y_{a_2}, then the contribution to the objective value $\mathrm{OPT}(I)$ is given by $w_{a_1} y_{a_1} + w_{a_2} y_{a_2}$. The optimal solution to I can be transformed into a feasible solution to I_{heur} with the slack $y_{a_1} + y_{a_2}$ on the new arc a_{12} arising from contracting a_1 and a_2. Note that by optimality, we have $y_{a_1} + y_{a_2} < T$. However, a_{12} contributes $\min(w_{a_1}, w_{a_2})(y_{a_1} + y_{a_2})$ to $\mathrm{OPT}(I_{\text{heur}})$. This shows $\mathrm{OPT}(I) \geq \mathrm{OPT}(I_{\text{heur}})$. Observe that there is no change in objective value if $w_{a_1} = w_{a_2}$, thus $\mathrm{OPT}(I) = \mathrm{OPT}(I_{\text{exact}})$. □

Example 2. Figure 3 visualizes the preprocessing of the instance from Example 1.

In the result of the heuristic preprocessing, the bounds $[55, 75]$ and $[20, 55]$ immediately fix a duration of 55 for both arcs, and thus a duration of 5 for the third arc in reverse direction. In particular, there is only one feasible periodic slack, which is hence optimal with weighted slack 110.

Moving to the instance after exact preprocessing, there is still a single feasible periodic slack for the same reason, but now with an optimal value of 130. Tracing back the previous preprocessing steps, which only work on features not contributing to the objective value, we see that an optimal timetable for the original instance has indeed weighted slack 130.

The contraction process in step 4 can in principle be extended to all vertices of degree two. However, in this case, it is possible that $\mathrm{OPT}(I) \neq \mathrm{OPT}(I_{\text{exact}})$, see Figure 4.

As a final remark, note that the cyclomatic number of the network is preserved by all preprocessing steps. In particular, the number of equality constraints in the cycle matrix MIP formulation remains unchanged. Of course, the number of events and activities decreases and thus does the number of variables in both MIP formulations.

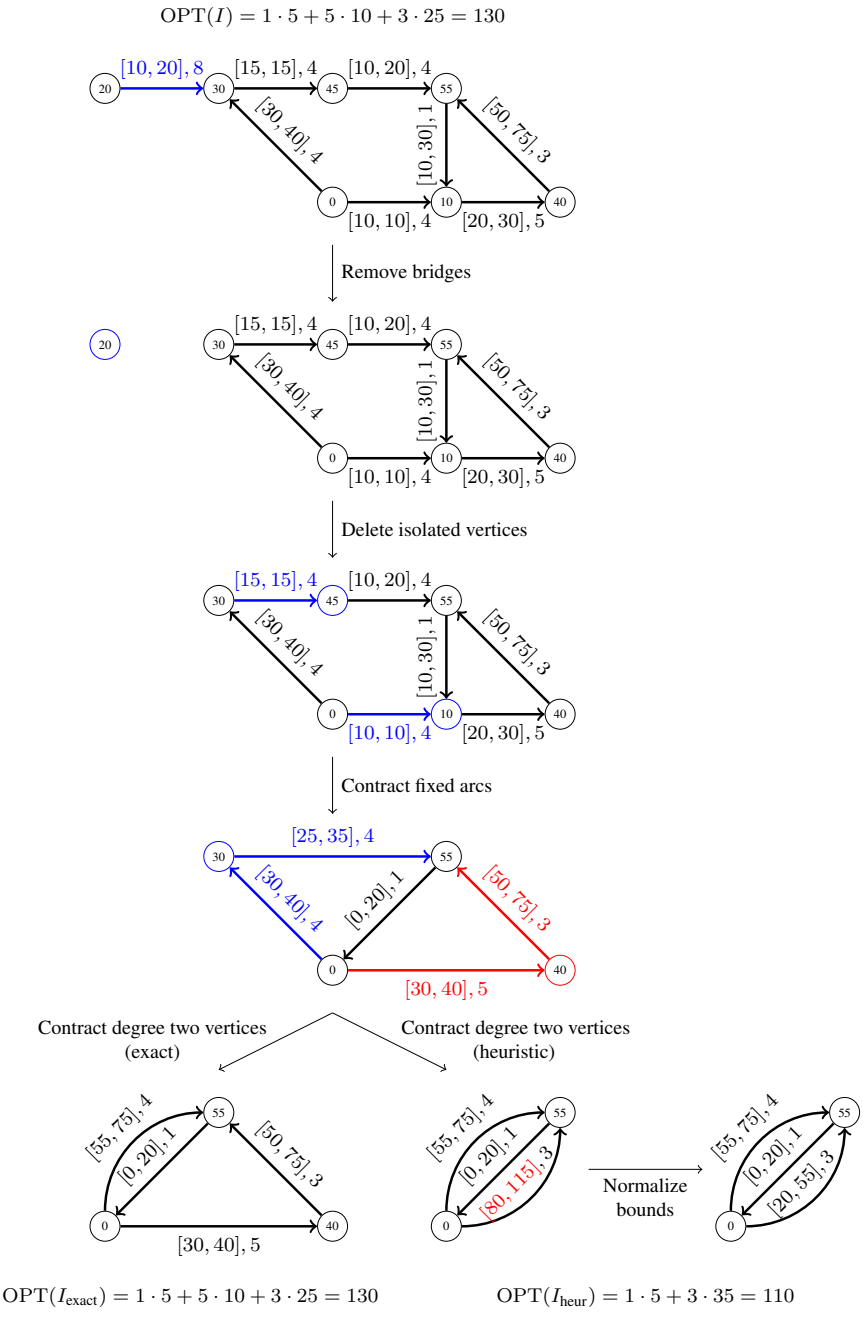
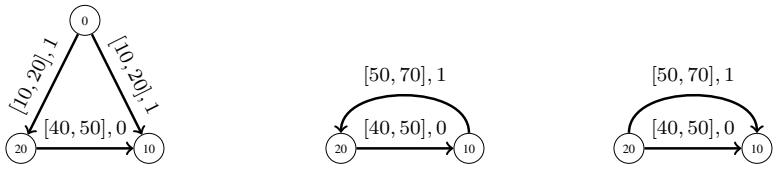


Figure 3: Exact and heuristic preprocessing



$\mathrm{OPT} = 1 \cdot 10 + 0 \cdot 10 + 1 \cdot 0 = 10$ $\mathrm{OPT} = 1 \cdot 20 + 0 \cdot 10 = 20$ $\mathrm{OPT} = 1 \cdot 0 + 0 \cdot 10 = 0$

Figure 4: Contracting arbitrary degree two vertices can lead to jumps in the optimal weighted slack.

3.3 Ignoring Free Arcs

Given a PESP instance $I = (G, T, \ell, u, w)$ and a real number $r \in [0, 1]$, we create the *ignore-r instance* I_r as follows (Goerigk and Liebchen, 2017, §3.1): Let A_{free} be the set of all *free* activities, i.e. the set of all arcs a with $u_a - \ell_a \geq T - 1$. Consider now the PESP instance arising from I by deleting the arcs from A_{free} in ascending order by weight w, until a total weight of $r \cdot \sum_{a \in A_{\text{free}}} w_a$ has been removed. Applying heuristic preprocessing to this instance defines I_r. In particular, $I_0 = I_{\text{heur}}$.

The networks I_r become smaller as r increases. Clearly, restricting a periodic timetable on I yields feasible timetable on I_r.

Lemma 2. *For any $r \in [0, 1]$ holds $\mathrm{OPT}(I_r) \leq \mathrm{OPT}(I)$.*

Proof. Any optimal solution to I is feasible for I_r, and removing arcs and heuristic preprocessing cannot increase the objective value by Lemma 1. □

Conversely, any timetable on I_r extends to a timetable on I, because the free activities are precisely the ones without any condition on their periodic slack.

Lemma 3. *Fix $r \in [0, 1]$ and let $A_{free,r}$ denote the free arcs of A removed when constructing I_r. Then*

$$\mathrm{OPT}(I) \leq \mathrm{OPT}(I_r) + \sum_{a \in A_{free,r}} w_a(T - 1) + E,$$

where E is an upper bound on the error in objective value that occurs during heuristic preprocessing of I_r.

Proof. Any optimal solution to I_r can be extended to the non-preprocessed network, causing an increase of at most E in objective value. Extending the timetable further to I adds at most $w_a(T - 1)$ for every activity a removed for constructing I_r. □

Since small terms $\sum_{A_{\text{free},r}} w_a(T - 1)$ can only be achieved with small values of r, good bounds are hard to obtain from the previous lemma. In practice, it often seems that $\mathrm{OPT}(I) \approx \mathrm{OPT}(I_r) + \sum_{A_{\text{free},r}} w_a(T - 1)/2$.

The *ignore problems* for our PESP solver stem from the ignore-r instances for different choices of $r \in [0, 1]$. The solver usually starts with a high r and decreases it after a certain amount of time. This way, the ignore problems become harder, but closer to the master problem.

3.4 Mixed Integer Programming Features

Let $I = (G, T, \ell, u, w)$ be a PESP instance. Our solver builds a mixed integer program using one of the following formulations:

- the incidence matrix formulation (§2 (2)),
- the cycle matrix formulation (§2 (3)) w.r.t. the cycle matrix of a strictly fundamental cycle basis arising from a minimum-weight spanning tree w.r.t. w,
- the cycle matrix formulation w.r.t. the cycle matrix of a minimum-weight undirected cycle basis w.r.t. $u - \ell$, if this basis is integral.

While computing a fundamental cycle basis is easily done using e.g. Kruskal's algorithm, our minimum-weight cycle basis algorithm currently relies on the rather time-consuming greedy algorithm by Horton (1987).

The MIP formulation can further be enhanced by

- *Cycle inequalities* (Odijk (1994)): Let $\gamma \in \{-1, 0, 1\}^A$ be the incidence vector of an oriented cycle. Then γ can be decomposed as $\gamma = \gamma_+ - \gamma_-$ into its non-negative and non-positive parts, i.e., $\gamma_+, \gamma_- \in \{0, 1\}^A$. The following inequality holds for any feasible periodic offset p and any feasible periodic slack y:

$$\left\lceil \frac{\gamma_+^t \ell - \gamma_-^t u}{T} \right\rceil \leq \gamma^t p = \frac{\gamma^t (y + \ell)}{T} \leq \left\lfloor \frac{\gamma_+^t u - \gamma_-^t \ell}{T} \right\rfloor$$

  If the cycle matrix formulation is used, and γ is a row of the cycle matrix Γ corresponding to an integer variable z_γ, then $\gamma^t(y + \ell)/T = z_\gamma$ and the cycle inequality yields bounds on the variable z_γ.

- *Change-cycle inequalities* (Nachtigall (1998)): Let $\gamma = \gamma_+ - \gamma_-$ be the incidence vector of an oriented cycle as above. Then the inequality

$$(T - \alpha)\, \gamma_+^t y + \alpha\, \gamma_-^t y \geq \alpha\, (T - \alpha), \qquad \text{where } \alpha = [-\gamma^t \ell]_T.$$

  holds for any feasible periodic slack y. Since the LP relaxations of both MIP formulations have 0 as their optimal values, the change-cycle inequalities are useful to provide a non-trivial lower bound for the slack variables y in the LP relaxation.

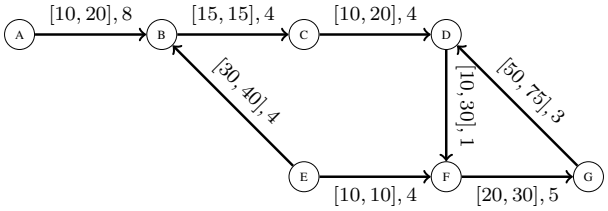


Figure 5: The PESP instance from Example 1 with named events, $T = 60$.

Example 3. We illustrate the helpfulness of the cycle inequalities on the instance of Example 1, see also Figure 5. An integral cycle basis consists of the two oriented cycles BCDFEB and DFGD, leading to the following cycle-based mixed integer program:

$$\begin{aligned}
\text{Minimize} \quad & 8y_{AB} + 4y_{BC} + 4y_{CD} + y_{DF} + 4y_{EB} + 4y_{EF} + 5y_{FG} + 3y_{GD} \\
\text{subject to} \quad & y_{BC} + y_{CD} + y_{DF} - y_{EF} + y_{EB} + 55 = 60z_1, \\
& y_{DF} + y_{FG} + y_{GD} + 80 = 60z_2, \\
& y_{BC} = y_{EF} = 0, \\
& y_{AB}, y_{CD}, y_{EB}, y_{FG} \in [0, 10], \quad y_{DF} \in [0, 20], \quad y_{GD} \in [0, 25], \\
& z_1, z_2 \in \mathbb{Z}.
\end{aligned}$$

The cycle inequalities for the two cycles read

$$\begin{aligned}
1 = \left\lceil \frac{15 + 10 + 10 - 10 + 30}{60} \right\rceil \quad &\leq z_1 \leq \left\lfloor \frac{15 + 20 + 30 - 10 + 40}{60} \right\rfloor \quad = 1 \\
2 = \left\lceil \frac{10 + 20 + 50}{60} \right\rceil \quad &\leq z_2 \leq \left\lfloor \frac{30 + 30 + 75}{60} \right\rfloor \quad = 2,
\end{aligned}$$

so that the above MIP simplifies to

$$\begin{aligned}
\text{Minimize} \quad & 8y_{AB} + 4y_{CD} + y_{DF} + 4y_{EB} + 5y_{FG} + 3y_{GD} \\
\text{subject to} \quad & y_{CD} + y_{DF} + y_{EB} = 5, \\
& y_{DF} + y_{FG} + y_{GD} = 40, \\
& y_{AB}, y_{CD}, y_{EB}, y_{FG} \in [0, 10], \quad y_{DF} \in [0, 20], \quad y_{GD} \in [0, 25].
\end{aligned}$$

The optimal solution is now to put as much slack as possible on the cheapest arc DF, i.e., to set $y_{DF} = 5$. Then $y_{FG} = 10$ and $y_{GD} = 25$, and the other slacks are 0. Consequently, the optimal solution has weighted slack $5 + 5 \cdot 10 + 3 \cdot 25 = 130$.

As a final ingredient to the MIP, we implemented a cutting plane separator. Since finding the maximally violating (change-)cycle cut is NP-hard and the best known algorithms are pseudo-polynomial dynamic programs (Borndörfer et. al. (2018)), we instead use a heuristic separator. Starting from a fractional solution to the LP relaxation, the separator adds violated (change-)cycle inequalities by inspecting the fundamental cycles of a minimum-slack spanning tree.

Currently, we interface the MIP solvers CPLEX[2] and SCIP (Gleixner et. al. (2018)).

3.5 Modulo Network Simplex

The modulo network simplex method (MNS, Nachtigall and Opitz (2008)) is an improving heuristic based on the idea that there is always an optimal PESP solution associated to a *spanning tree structure*. Formally, let $I = (G, T, \ell, u, w)$ be a PESP instance. Then there is an optimal periodic slack y^* and a spanning tree F of G such that for all arcs a in F holds either $y^*_a = 0$ or $y^*_a = u_a - \ell_a$. More precisely, the spanning tree structures correspond one-to-one to the vertices of the convex hull of

$$\{(y, z) \in \mathbb{R}^A_{\geq 0} \times \mathbb{Z}^\mu \mid \Gamma(y + \ell) = Tz, 0 \leq y \leq u - \ell\},$$

[2] https://www.ibm.com/analytics/cplex-optimizer

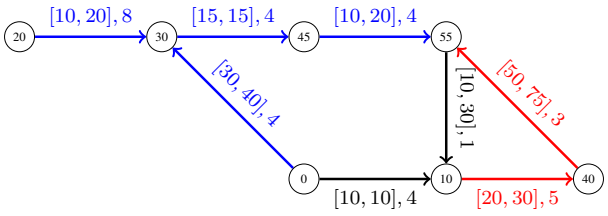


Figure 6: The PESP instance from Example 1 with an optimal spanning tree structure: Blue arcs have slack 0, red arcs have slack $u - \ell$.

i.e., the polytope associated to the cycle matrix MIP formulation (Nachtigall and Opitz, 2008, Theorem 2.1). Figure 6 illustrates an optimal spanning tree structure in our running example. Given a spanning tree structure, one can look for a better spanning tree structure by adding a co-tree arc a and deleting a tree arc along the fundamental cycle associated to a. This search can be performed with a simplex-style tableau, however, the change in objective value needs to be recomputed rather costly. After a few iterations, the MNS usually becomes stuck in a local optimum.

Our MNS implementation features the following:

1. *Initialization*: Given a feasible PESP solution, we fix the corresponding cycle offset variables z and solve the cycle matrix MIP formulation for the remaining – continuous – variables y. A strictly fundamental cycle basis computed from a minimum-weight spanning tree w.r.t. w produces the required cycle matrix. Since all integer variables are fixed, this is now a linear program, and any optimal vertex of the corresponding polytope yields a spanning tree structure.

2. *Inner loop*: Try to improve the current tree structure by exchanging tree arcs with co-tree arcs. The usual strategy is to apply steepest descent. However, the running time is drastically improved by a quality-first rule, i.e., the search for an improving move is already stopped as soon as a sufficient improvement has been achieved.

3. *Single-node cuts* (Nachtigall and Opitz (2008)): If the inner loop reaches a local optimum, then try to adjust the current timetable by modifying a single vertex.

4. *Multi-node cuts*: If inner loop and single-node cuts do not improve the timetable, then try to shift a bigger set of vertices in a random and greedy way, see Goerigk and Schöbel (2013).

5. *Restart*: If inner loop, single- and multi-node cuts, and rebuilding the spanning tree structure as in the initialization process do not lead to a better timetable, then we restart the MNS with a worse solution that was not computed by one of the MNS features. This is carried out in a tabu-search style.

The algorithm is regularly updated with the current incumbent solution, which might have been found by other algorithms, e.g., the MIP solver. The MNS turns out to be a powerful improving heuristic in the beginning of a solving process. In the later phase, improvements by the inner loop are rare, and mostly come from multi-node cuts and restarts.

3.6 Max-Cut Heuristic

Since the modulo network simplex method gets trapped in local optima too often, we developed a deeper improving heuristic dominating the MNS inner loop, single-node and multi-node cuts. Instead of searching for multi-node cuts in a heuristic way, we turn this into an optimization problem:

Problem 1 (Maximally improving delay cut)**.** Let $I = (G, T, \ell, u, w)$ be a PESP instance. We call any pair (S, d) with $S \subseteq V$ and $d \in \{1, \dots, T-1\}$ a *delay cut*. Given a feasible periodic timetable π for I, find a delay cut (S, d) such that the periodic timetable $\pi(S, d) \in [0, T)^V$ is feasible and has minimum weighted slack, where

$$\forall v \in V: \quad \pi(S,d)_v := \begin{cases} \pi_v + d & \text{if } v \in S, \\ \pi_v & \text{if } v \notin S. \end{cases}$$

Multi-node cuts are delay cuts, and clearly single-node cuts are delay cuts with a singleton subset S. Moreover, any move in the MNS inner loop can be seen as a delay cut, as removing a spanning tree arc induces a fundamental cut. In particular, if a periodic timetable cannot be improved by a delay cut, then the timetable is locally optimal for the MNS inner loop and single-/multi-node cuts.

Lemma 4. *For fixed d, the problem of finding a maximally improving delay cut is a maximum cut problem with possibly both positive and negative weights.*

Proof. We want to minimize the minimum weighted slack of $\pi(S, d)$, i.e.,

$$\sum_{ij \in A} w_{ij} [\pi(S,d)_j - \pi(S,d)_i - \ell_{ij}]_T.$$

Of course, since a fixed feasible timetable π is given, we can instead minimize

$$\sum_{ij \in A} w_{ij} \left([\pi(S,d)_j - \pi(S,d)_i - \ell_{ij}]_T - [\pi_j - \pi_i - \ell_{ij}]_T \right).$$

The summand vanishes for arcs where the endpoints are both in S or both in $V \setminus S$. Therefore, we can minimize

$$\begin{aligned} &\sum_{ij \in \delta^+(S)} w_{ij} \left([\pi_j - \pi_i - d - \ell_{ij}]_T - [\pi_j - \pi_i - \ell_{ij}]_T \right) \\ + &\sum_{ij \in \delta^-(S)} w_{ij} \left([\pi_j - \pi_i + d - \ell_{ij}]_T - [\pi_j - \pi_i - \ell_{ij}]_T \right) \end{aligned}$$

Here, $\delta^+(S)$ and $\delta^-(S)$ denote the sets of all arcs leaving S and entering S, respectively. For fixed d and π, we therefore have a minimization problem of the form

$$\sum_{ij \in \delta^+(S)} c^+_{ij} + \sum_{ij \in \delta^-(S)} c^-_{ij},$$

for fixed c^+, c^-, and we set c^+_{ij} (resp. c^-_{ij}) to ∞ if $[\pi_j - \pi_i - d - \ell_{ij}]_T$ (resp. $[\pi_j - \pi_i + d - \ell_{ij}]_T$) is not a feasible slack for the arc ij. Since in principle c^+, c^- can take any value in $(-T, T)$, we arrive at a minimum cut problem with positive and negative costs, which is at the same time a maximum cut problem by switching signs. □

The maximum cut problem as constructed in the proof is in general not solvable in polynomial time due to the presence of arcs with positive weight. To emphasize this difficulty, we prefer the term "maximum cut" over "minimum cut".

However, solving the maximally improving delay cut problem for a fixed d turns out to be well doable by a MIP solver in practice. Our PESP solver invokes SCIP to compute a maximally improving delay cut for $d = 1, 2, \ldots, \lceil (T-1)/2 \rceil$. Note that by symmetry, a delay cut (S, d) is as good as $(V \setminus S, T - d)$, and hence there is no need to check for all values of d up to $T-1$. If this max-cut heuristic (or another concurrently running algorithm) finds a better solution, it is restarted. Although this MIP-based approach is inferior to MNS in the early stage of solving, it provides better quality solutions in later phases. Using a MIP solver under the hood also enables to prove local optimality for the cut methods of MNS.

3.7 Further Ingredients

Furthermore, our solver is able to invoke the following strategies:

- *Reflow heuristic*: We apply the MNS initialization step to every new incumbent in a solution pool, not only the ones found by the modulo network simplex algorithm. Since this only involves solving a single linear program, dual to a minimum cost flow problem (Nachtigall and Opitz, 2008, §2), this is very fast. We currently use SCIP for this task.

- *SAT initial solution*: The problem of finding a feasible periodic timetable for a given PESP instance can be formulated as a boolean satisfiability (SAT) problem. This is done using the order encoding and rectangle covering strategy from Großmann et. al. (2012). Although this produces a pseudo-polynomial number of variables and clauses, a specialized SAT solver – we use *Glucose*[3] – is able to provide a feasible truth assignment typically within a second or less. The truth assignment is then transformed back to a feasible periodic timetable. We call this procedure before starting the master and the ignore problems to quickly obtain an initial solution. This is valuable since especially on larger instances, the MIP solver has difficulties to find a feasible solution in the beginning, and MNS and the max-cut heuristic require a feasible timetable as input.

- *SAT propagator*: If the MIP solver – like SCIP or CPLEX without dynamic search – applies a classical branch-and-cut algorithm to the incidence matrix MIP formulation, then the branching decisions are on the periodic offset variables p. We regularly query the current local bounds for p at the nodes of the branch-and-bound tree, and transform the PESP feasibility problem into a SAT problem as above. If the instance is infeasible – this is usually detected after a few milliseconds by glucose – we can prune the node. Otherwise we obtain another feasible solution. Unfortunately, this delays the branch-and-cut process, and the pruning effects are rather small and do not compensate turning off the dynamic search of CPLEX.

- *MaxSAT heuristic*: With a similar approach as in the feasibility case, finding an optimal periodic timetable can be translated into a weighted partial MaxSAT problem (Großmann (2016)). Since the number of clauses and variables is rather high, even a

[3] `http://www.labri.fr/perso/lsimon/glucose`

fast MaxSAT solver, like e.g. *Open-WBO* (Martins et. al. (2014)), cannot help finding good quality solutions fast. However, selecting only a small portion of arcs for the optimization is sometimes superior to a MIP approach (Roth (2019)).

4 PESPlib Instances

The library *PESPlib* serves as a benchmark set for our solver. It currently comprises 20 periodic event scheduling instances, all of which have a period time of 60 minutes. The first 16 instances arose from the German long-distance railway network. They typically decompose into disjoint paths when removing all free arcs. The last four instances are bus timetabling instances and have a different structure, e.g., there are multiple arcs between two vertices.

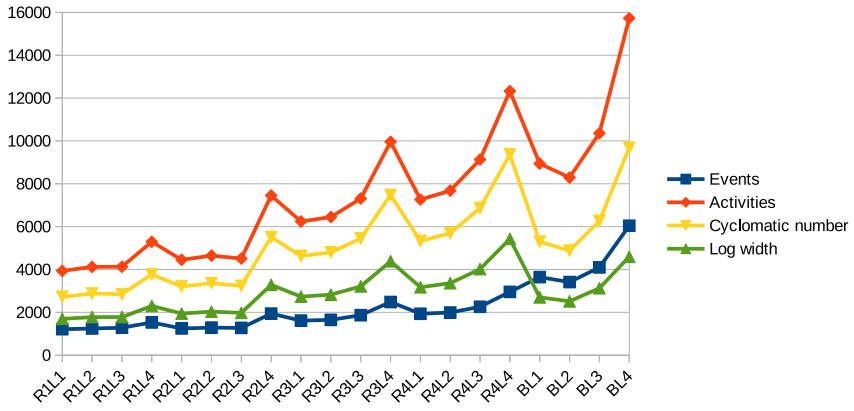


Figure 7: Sizes of PESPlib instances after heuristic preprocessing

Figure 7 compares the PESPlib instances using various measures of difficulty: number of events, number of activities, cyclomatic number and log width. Here, log width means the decadic logarithm of the number of possible values for the periodic offset vector p in the incidence matrix MIP formulation. The data refers to the networks after heuristic preprocessing. Furthermore, we transformed the instances BL1-BL4 to simple graphs, i.e., without multiple arcs. The smallest PESPlib instance R1L1 has a cyclomatic number of 2 722, and solving to proven optimality is currently out of reach. For example, a PESP instance with the rather tiny cyclomatic number 294 has found its way into the *MIPLIB2003* collection under the name *timtab2* (Liebchen and Möhring (2003)), and was solved to optimality with a pure MIP strategy within 6 432 seconds in 2016 – using the commercial MIP solver Xpress on 6 144 cores in parallel. On a single standard computer, a solving time of 22 hours with the help of special cuts is reported[4].

The effect of preprocessing is shown in Figure 8. For the first 16 instances, exact preprocessing reduces the number of events to roughly two thirds, and heuristic preprocessing

[4] http://miplib2010.zib.de/miplib2003/miplib2003/timtab2.php

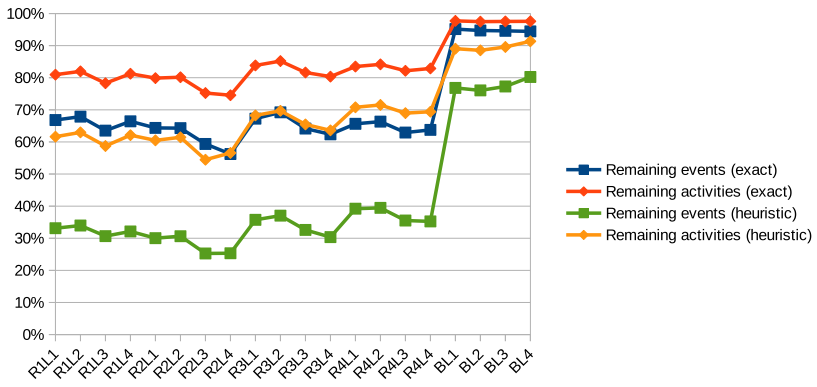


Figure 8: Size reduction by preprocessing

even to one third. The BL instances show different behavior: After heuristic preprocessing, more than 75% of all events remain.

Finally, Table 1 shows the currently best known incumbents, lower bounds, and optimality gaps on the weighted slack of all PESPlib instances. The dual bounds have received special attention only for the first instance so far.

Instance	primal	dual	gap [%]
R1L1	30 780 097	16 897 987	45.10
R1L2	31 682 263	4 975 398	84.30
R1L3	30 307 719	6 498 424	78.56
R1L4	27 326 571	6 297 850	76.95
R2L1	42 502 069	9 507 113	77.63
R2L2	41 534 563	7 768 806	81.30
R2L3	39 942 656	8 224 882	79.41
R2L4	33 063 475	5 217 025	84.22
R3L1	44 396 635	7 906 870	82.19
R3L2	46 048 483	7 432 716	83.86
R3L3	42 833 223	6 628 317	84.53
R3L4	34 694 043	5 623 632	83.79
R4L1	51 650 471	10 089 083	80.47
R4L2	49 579 843	7 975 150	83.91
R4L3	45 881 499	7 477 035	83.70
R4L4	38 836 756	5 147 195	86.75
BL1	7 387 963	1 477 565	80.00
BL2	8 143 507	1 730 247	78.75
BL3	7 826 762	1 205 501	84.60
BL4	7 359 779	1 004 303	86.35

Table 1: PESPlib incumbents as of October 24, 2018. The average optimality gap is 80.32%.

No.	Concurrent phase	Final phase	Repetitions
1	20 min	2 min	10
2	60 min	5 min	10
3	4 h	–	1
4	8 h	–	1

Table 2: Primal bound experiments

5 Computational Results

In this section, we report on the progress that our concurrent PESP solver achieved on the PESPlib instances. The computations were carried out on two machines: A 3.4 GHz Intel Xeon E3-1245 CPU and a 3.7 GHz Intel Xeon E3-1290 V2, both equipped with 32 GB RAM and allowing 8 threads running in parallel.

5.1 Primal Bound Experiments

We ran four consecutive experiments to improve the primal bounds of the PESPlib instances, see Table 2. As initial solution, the first experiment uses the timetable returned by the SAT solver (see §3.7). In particular, we construct an initial solution from scratch and do not use any input timetable from PESPlib. The first experiment spends 20 minutes in the concurrent phase and is repeated 10 times with different parameter settings for the ignore problems and the modulo network simplex quality-first strategy. The best out of this 10 solutions is taken as input for Experiment 2, which is also run with 10 different configurations. Again, the best of these solutions is taken over to Experiment 3, which is executed only once and in turn delivers its solution to Experiment 4. The experiments 3 and 4 do not enter the final phase, as the heuristic preprocessing for the master problem is switched off. In all experiments, we also do not make use of SAT methods beyond finding an initial solution.

As a MIP solver, we use CPLEX 12.8 with feasibility emphasis and dynamic search applied to the incidence matrix formulation. We add cycle inequalities before starting, but no further cuts. Experiments 1-3 use 7 threads as described in §3.1. In particular, CPLEX is run on one thread for each master and ignore problem. In Experiment 4, the ignore problem is turned off, and CPLEX uses 4 internal threads for the master problem.

The results of the experiments are summarized in Table 3. Already after Experiment 1 with 20 minutes in the concurrent phase, 10 out of 20 instances end up with a better objective value than in the PESPlib. After Experiment 2, all but the two instances R2L1 and R4L4 have been improved. In Experiments 3 and 4, the improvements become smaller, however we were able to find better incumbents for the two remaining instances as well. As the BL instances seem to have received less attention in the past, we can even improve their primal bounds by more than 10%.

Figure 9 shows the relative improvement of the objective value of the master problem for each of the heuristics ignore, MIP, MNS, max-cut, and reflow. 100% denote the total gain in objective value during the best run of the experiment. While the ignore heuristic, i.e., expanding timetables from the ignore problem, is the dominant source for new incumbents in the early stage, it has almost no effect in later phases. On the other hand, the “global” strategies such as MIP and max-cut become more important. This is why we decided to

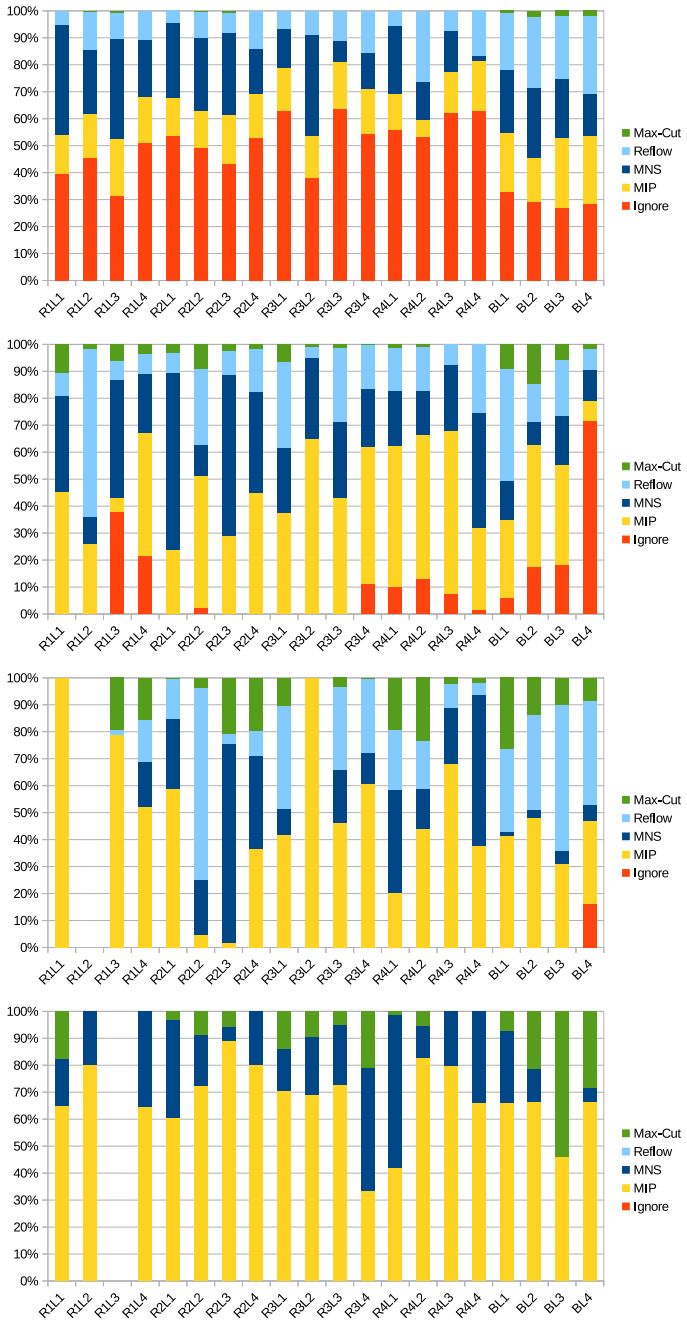


Figure 9: Relative improvement by heuristic for Experiments 1-4 from top to bottom.

Inst.	SAT start*	Exp. 1	Exp. 2	Exp. 3	Exp. 4	**
R1L1	74 234 870	30 861 021	30 501 068	30 493 800	30 463 638	1.03%
R1L2	72 731 210	30 891 284	30 516 991	30 516 991	30 507 180	3.71%
R1L3	71 682 438	30 348 596	29 335 021	29 319 593	29 319 593	3.26%
R1L4	67 395 169	27 635 070	26 738 840	26 690 573	26 516 727	2.96%
R2L1	97 230 766	42 863 646	42 598 548	42 463 738	42 422 038	0.19%
R2L2	95 898 935	42 024 414	41 149 768	40 876 575	40 642 186	2.15%
R2L3	93 800 082	39 054 513	38 924 083	38 881 659	38 558 371	3.47%
R2L4	84 605 216	33 256 602	32 707 981	32 548 415	32 483 894	1.75%
R3L1	92 939 173	44 216 552	43 521 250	43 460 397	43 271 824	2.53%
R3L2	91 336 260	45 829 180	45 442 171	45 401 718	45 220 083	1.80%
R3L3	89 741 119	42 112 858	41 103 062	41 005 379	40 849 585	4.63%
R3L4	74 142 083	34 589 170	34 018 560	33 454 773	33 335 852	3.91%
R4L1	98 276 297	50 638 727	49 970 330	49 582 677	49 426 919	4.30%
R4L2	101 135 698	50 514 805	49 379 256	49 018 380	48 764 793	1.64%
R4L3	96 629 751	46 406 365	45 656 395	45 530 113	45 493 081	0.85%
R4L4	80 446 905	40 706 349	38 884 544	38 695 188	38 381 922	1.17%
BL1	15 367 998	7 299 228	6 394 914	6 375 778	6 333 641	14.27%
BL2	16 046 736	7 378 468	6 837 447	6 819 856	6 799 331	16.51%
BL3	14 850 854	7 512 685	7 065 270	7 011 324	6 999 313	10.57%
BL4	15 618 608	7 997 783	7 330 393	6 738 582	6 562 147	10.84%

Table 3: Objective values after Experiments 1-4. Green objectives are better than in the PESPlib. All solutions to Experiment 4 are locally optimal for max-cut. *The objective value of the initial solution provided by SAT is to be interpreted on the heuristically preprocessed instance. **This is the relative improvement compared to PESPlib.

switch the ignore problem off for Experiment 4, so that 3 threads become available, which are again invested in CPLEX on the master problem.

5.2 Dual Bound Experiments

Our set-up for the primal bound experiments does not provide strong dual bounds, mostly due to the feasibility emphasis parameter setting of CPLEX. Moreover, the incidence matrix MIP formulation seems to be better for finding good primal solutions fast, but weaker concerning lower bounds. When computing a minimum-weight cycle basis and plugging in the corresponding cycle basis, we empirically observed stronger dual bounds.

For our dual bound experiment, we run CPLEX on 6 threads for 8 hours with best bound emphasis. We also invoke our heuristic cutting plane separator for cycle and change-cycle inequalities (§3.4). All other primal heuristics, e.g., MNS and max-cut, are not started. To simplify the original PESP instance I even more, we switch the master problem off and perform the computations only on the ignore problem given by the ignore-0.01 instance $I_{0.01}$ (§3.3). Since $\mathrm{OPT}(I_{0.01}) \leq \mathrm{OPT}(I)$ by Lemma 2, lower bounds on $\mathrm{OPT}(I_{0.01})$ are also valid lower bounds on $\mathrm{OPT}(I)$. Shrinking the incumbent timetable from Experiment 4 to the ignore problem provides a MIP start.

Table 4 contains the results of the dual bound experiment. We could improve all dual

bounds significantly, often by a factor of 2. As a consequence, we can reduce the average optimality gap over all instances from 80.32% to 48.36%.

Instance	Dual bound	PESPlib improvement	Optimality gap
R1L1	19 878 200	17.64%	34.75%
R1L2	19 414 800	290.22%	36.36%
R1L3	18 786 300	189.09%	35.93%
R1L4	16 822 200	167.11%	36.56%
R2L1	25 082 000	163.82%	40.88%
R2L2	24 867 400	220.09%	38.81%
R2L3	23 152 300	181.49%	39.96%
R2L4	18 941 500	263.07%	41.69%
R3L1	25 077 800	217.16%	42.05%
R3L2	25 272 600	240.02%	44.11%
R3L3	21 642 500	226.52%	47.02%
R3L4	16 479 500	193.04%	50.57%
R4L1	27 243 900	170.03%	44.88%
R4L2	26 368 200	230.63%	45.93%
R4L3	22 701 400	203.62%	50.10%
R4L4	15 840 600	207.75%	58.73%
BL1	3 668 148	148.26%	42.08%
BL2	3 943 811	127.93%	42.00%
BL3	3 571 976	196.31%	48.97%
BL4	3 131 491	211.81%	52.28%

Table 4: Dual bound experiment: Best lower bound, applied (change-)cycle cuts, relative improvement compared to the PESPlib bound, optimality gap w.r.t. the primal solution of Experiment 4. Average optimality gap over all instances: 48.36%.

6 Summary

We described a powerful framework for solving periodic event scheduling problems, providing better solutions faster, and on relatively large instances. Moreover, our approach combines many of the currently best known strategies for periodic timetabling in a single program, and it is able to compare the impact of the different methods.

Combining many state-of-the-art methods in a concurrent manner provides a significant speedup: For example, we are able to compute a new best solution to 10 out of 20 PESPlib instances in as little as 20 minutes, starting from scratch and not using any input timetable. Given the fact that many previous incumbent solutions were computed with a sequential approach using MIP and MNS for 8 hours (Goerigk and Liebchen (2017)), we achieved a speedup factor which is bigger than the number of parallel threads our solver uses.

Our cutting plane separation approach is tailor-made for improving the lower bounds on the objective values. Given that not much progress is expected in the primal bound on the PESPlib instances, we believe that the key to solve PESPlib to optimality lies in better strategies for the dual side, where even our heuristic separator is able to reduce the optimality gap significantly.

Acknowledgements
We thank C. Liebchen for sharing his PESP expertise. Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – The Berlin Mathematics Research Center MATH+ (EXC-2046/1, project ID: 390685689).

References

Borndörfer, R., Hoppmann, H., Karbstein, M., Lindner, N., 2018. "Separation of Cycle Inequalities in Periodic Timetabling", ZIB-Report 18-16, Zuse Institute Berlin.

Gleixner, A. et al., 2018. "The SCIP Optimization Suite 6.0", Technical Report, Optimization Online, July 2018.

Goerigk, M., Schöbel, A., 2013. "Improving the modulo simplex algorithm for large-scale periodic timetabling", *Computers & Operations Research*, vol. 40, no. 5, pp. 1363-1370.

Goerigk, M., Liebchen, C., 2017. "An improved algorithm for the periodic timetabling problem", In: *17th Workshop on Algorithmic Approaches for Transportation Modelling, Optimization, and Systems (ATMOS 2017)*, vol. 59, pp. 12:1-12:14.

Großmann, P., Hölldobler, S., Manthey, N., Nachtigall, K., Opitz, J., Steinke, P., 2012. "Solving Periodic Event Scheduling Problems with SAT", In: *Advanced Research in Applied Artificial Intelligence*, Springer Berlin Heidelberg, pp. 166-175.

Großmann, P., 2016. "Satisfiability and Optimization inPeriodic Traffic Flow Problems", PhD thesis, TU Dresden.

Horton, J., 1987. "A polynomial–time algorithm to find the shortest cycle basis of a graph", *SIAM J. Comput.*, vol. 16, pp. 359-366.

Kümmling, M., Großmann, P., Nachtigall, K., Opitz, J., Weiß, R., 2015. "A state-of-the-art realization of cyclic railway timetable computation", *Public Transport*, vol. 7, no. 3, pp. 281-293.

Liebchen, C., Möhring, R., 2006. "Information on MIPLIB's timetab-instances", Technical Report No. 2003/49, Technische Universität Berlin.

Liebchen, C., 2006. "Periodic timetable optimization in public transport", PhD thesis, Technische Universtität Berlin.

Liebchen, C., 2008. "The first optimized railway timetable in practice", *Transportation Science*, vol. 42, no. 4, pp. 420-435.

Liebchen, C., Peeters, L., 2009. "Integral cycle bases for cyclic timetabling", *Discrete Optimization*, vol. 6, no. 1, pp. 98-109.

Martins, R., Manquinho, V., Lynce, I., 2014. "Open-WBO: A Modular MaxSAT Solver", SAT 2014, pp. 438-445.

Nachtigall, K., 1998. "Periodic Network Optimization and Fixed Interval Timetables", Habilitation thesis, Universtität Hildesheim.

Nachtigall, K., Opitz, J., 2008. "Solving periodic timetable optimisation problems by modulo simplex calculations", In: *8th Workshop on Algorithmic Approaches for Transportation Modeling, Optimization, and Systems (ATMOS '08)*, vol. 9.

Odijk, M., 1994. "Construction of periodic timetables, part 1: A cutting plane algorithm.", Technical Report 94-61, TU Delft.

Roth, S., 2019. "SAT heuristics for periodic timetabling", Master's thesis (in preparation), Freie Universität Berlin.

Serafini, P., Ukovich, W., 1989. "A mathematical model for periodic scheduling problems", *SIAM Journal on Discrete Mathematics*, vol. 2, no. 4, pp. 550-581.

Re-optimizing ICE Rotations after a Tunnel Breakdown near Rastatt

Ralf Borndörfer[a], Boris Grimm [a,1], Thomas Schlechte[b]
[a] Zuse-Institute Berlin
Takustr.7, 14195 Berlin, Germany
[1] E-mail: grimm@zib.de, Phone: +49 (0) 30 84185 202
[b] LBW Optimization GmbH
Obwaldener Zeile 19, 12205 Berlin, Germany

Keywords
case study, railway rolling stock optimization, mixed integer programming, re-optimization

Abstract
This paper deals with the situation of re-scheduling railway rolling stock rotations after a major disruption. Disruptions are an all day life problem when operating a railway system. Nevertheless, an unusal example for a disruption is the collapse of a tunnel ceiling near Rastatt in southern Germany due to construction works related to the enlargement of the tracks between Karlsruhe and Basel. As a result the main railway connections Amsterdam-Basel and Berlin-Basel, located on top of the tunnel, had to be closed from August 12th to October 2nd 2017. This had a major impact on the railway network in southern Germany. Hence, all rotation plans and train schedules for both passenger and cargo traffic had to be revised. Since the disruption was very long lasting the revision of the rotation plans was done on a tactical rather than on an operational level which would usually be the case. In this paper we focus on a case study for this situation and compute new rotation plans via mixed integer programming for the ICE high speed fleet of DB Fernverkehr AG one of the largest passenger railway companies in Europe. In our approach we take care of some side constraints to ensure a smooth continuation of the rotation plans after the disruption has ended.

1 Introduction

Planning rolling stock movements in industrial passenger railway applications is a long-term process based on timetables which are also often valid for long periods of time. For these timetables and periods rotation plans, i.e., plans of railway vehicle movements are constructed as templates for these periods. Those rotation plans will gain accuracy the closer the day of operation comes. During operation the rotation plans are affected by all kinds of unplanned events such as natural disasters, technical problems, or man-made impediments. An example for the latter case is the collapse of a tunnel ceiling between Rastatt and Baden-Baden in southern Germany due to construction works related to the enlargement of the tracks between Karlsruhe and Basel. In Figure1 the ICE highspeed train line network is shown with a red mark where the tunnel collapsed close to Baden-Baden on the pink line in the south-west part of the map. As a result the main railway connections Amsterdam-

Basel and Berlin-Basel, located on top of the tunnel, had to be closed from August 12th to October 2nd 2017. This had a major impact on the railway network in southern Germany since this is the direct electrified railway corridor connecting the Netherlands via Germany with Switzerland and Italy. Hence, all rotation plans and train schedules for both passenger and cargo traffic had to be revised. As a side effect many other industry branches suffered from a lack of materials that could not be delivered in time.

In this paper we focus on this concrete case and compute new rotation plans for the ICE high speed fleet of DB Fernverkehr AG one of the largest passenger railway companies in Europe. To bring these rotations into practice the two following conditions had to be considered:

1. Passenger trains operating in southern Baden-Wuertemberg or Switzerland were only operated till Rastatt.

2. The 3rd of October is a national holiday in Germany.

As a result only a limited offer on railway connections exists for this day comparable to the weekend traffic. Nevertheless a seamless connection between the rotation plan for the period covering the construction works, the holiday, and the relaunched regular timetable should be guaranteed. Thus the less differences between the different parts of the rotation plans exist the better.

Constructing new or revised tours of rolling stock vehicles through the timetable after disruptions is a well studied topic in the literature, see Cacchiani et al. (2014) for an overview. Usually, a rescheduling based on a timetable update is done, followed by the construction of new rotations that reward the recovery of parts of the obsolete rotations. We consider a different, more integrated approach with side constraints on the start and end states of the vehicles and a system to reward preserved or similar operated train connections in both periods. The approach is based on the mixed integer programming approach presented in Reuther (2017). The goal is to minimize the operating costs while preparing as best as possible for the relaunch of the regular timetable afterwards. In contrast to the situation this paper deals with most of the research in the literature considers 'ad hoc' rescheduling approaches. For example in Lusby et al. (2017) vehicle rotations have to be revised (almost immediately) for some suddenly occurring reasons. In the case this paper focuses on the disruption is long lasting and therefore changes to the rotations are more on a tactical or strategical level than on an operational.

2 Rolling Stock Rotation Problem with Side Constraints

In this section we consider the *Rolling Stock Rotation Problem* (RSRP) and extend a hypergraph-based integer programming formulation to suit our setting. We focus on the main modeling ideas and refer the reader to the paper Borndörfer et al. (2016) for technical details.

In our computations we distinguish between a cyclical planning horizon of one week and an acyclic planning horizon of two weeks. In the latter case the exact period is from September 27th to October 10th. Let T denote the set of timetabled passenger trips. Let V be a set of *nodes* representing timetabled departures and arrivals of vehicles operating passenger trips of T. In the acyclic case there are additional nodes for start and end positions of vehicles at beginning and at end of the planning horizon. The sets of start and end positions are denoted by $S \subset V$ and $E \subset V$, respectively. Trips that could be operated

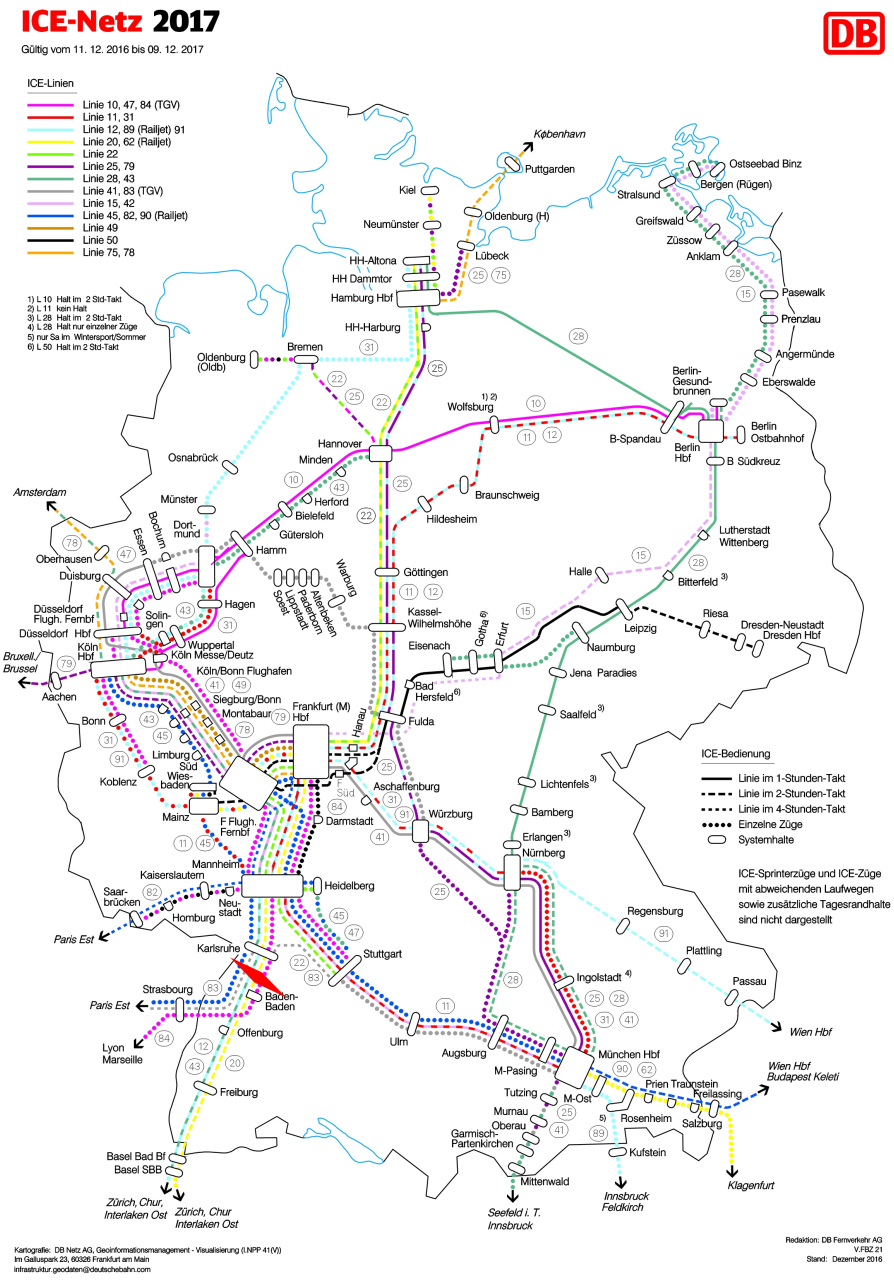


Figure 1: ICE highspeed train line network.

with two or more vehicles have the appropriate number of arrival and departure nodes. Let further $A \subseteq V \times V$ be a set of directed standard arcs, and $H \subseteq 2^A$ a set of *hyperarcs*. Thus, a hyperarc $h \in H$ is a set of standard arcs and includes always an equal number of tail and head nodes, i.e., arrival and departure nodes. A hyperarc $h \in H$ *covers* $t \in T$ if each standard arc $a \in h$ represents an arc between the departure and arrival of t. Each of the standard arcs a represents a vehicle that is required to operate t. We define the set of all hyperarcs that cover $t \in T$ by $H(t) \subseteq H$. By defining hyperarcs appropriately, vehicle composition rules and regularity aspects can be directly handled by the model. Hyperarcs that contain arrival and departure nodes of different trips are used to model deadhead trips between the operation of two trips (or even more if couplings are involved). The RSRP *hypergraph* is denoted by $G = (V, A, H)$. We define sets of hyperarcs coming into and going out of $v \in V$ in the RSRP hypergraph G as $H(v)^{\text{in}} := \{h \in H \mid \exists\, a \in h \,:\, a = (u, v)\}$ and $H(v)^{\text{out}} := \{h \in H \mid \exists\, a \in h \,:\, a = (v, w)\}$, respectively. We call a set $R \subseteq H$ a reference solution if R is a set of hyperarcs that define a set of s-e-(hyper)paths in G such that each $t \in T$, $s \in S$, and $e \in E$ is covered by a single hyperarc. Let $c_R : H \mapsto \mathbb{Q}^+$ denote the cost function of G associated with reference solution R respectively the vehicle movements behind it. Though by $c_R(h)$ all costs for vehicle usage, deadhead trip costs, and energy consumption are given as a weighted sum of the different parameter. Additional a penalty for choosing a different vehicle movement compared to the reference solution is included. In more detail there are penalties for choosing different vehicle types, configurations, or orientations for a trip or choosing different connections between two trips where one succeeds the other. Another important aspect in rolling stock planning and optimization is vehicle maintenance. At DB Fernverkehr AG there are several different maintenance rules for the different ICE fleet that all have to be considered. In this paper we focus on a single maintenance rule that is based on the accumulated kilometers a vehicle is operated between two maintenance services. We denote its upper bound on the total mileage between two maintenance services by U. Maintenance services could only be performed at special maintenances locations $m \in M$. The kilometers a vehicle is moved during an operation modelled by a chosen hyperarc is given by a function $r : V \times H \mapsto [0, U]$. This includes necessary deadhead trips to reach maintenance facilities or turn around trips to change the orientation of the vehicle. To model maintenance services in the RSRP *hypergraph* additional maintenance service hyperarcs were added for each pair of trips if it is possible to visit a maintenance facility and perform the service between the operation of the two trips. The cost for the additional deadhead trip and the cost for the maintenance service is added to the cost of the hyperarc. In this sense a s-e-(hyper)path or a cycle in G is called maintenance feasible, if and only if the accumulated kilometers of all trips and deadhead trips along this path or cycle between each two hyperarcs with a mainenance service is smaller than U. The *Rolling Stock Rotation Problem with Side Constraints* (RSRPSC) is to find a cost minimal, maintenance feasible set of hyperarcs $H_0 \subseteq H$ such that H_0 is a collection of cycles in the cyclic or a collection of s-e-(hyper)paths in the acyclic case and all nodes in V are covered by a hyperarc $h \in H_0$.

Using a binary decision variable x_h for each hyperarc the RSRPSC can be stated as an integer program as follows:

$$\min \sum_{h \in H} c_h x_h, \tag{1}$$

$$\sum_{h \in H(t)} x_h = 1 \qquad \forall t \in T, \tag{2}$$

$$\sum_{h \in H(s)^{\text{out}}} x_h = 1 \qquad \forall s \in S, \tag{3}$$

$$\sum_{h \in H(d)^{\text{in}}} x_h = 1 \qquad \forall d \in D, \tag{4}$$

$$\sum_{h \in H(v)^{\text{in}}} x_h = \sum_{h \in H(v)^{\text{out}}} x_h \qquad \forall\, v \in V \setminus \{S \cup D\}, \tag{5}$$

$$w_a \leq \sum_{h \in H(a)} U\, x_h \qquad \forall a \in A, \tag{6}$$

$$\sum_{a \in A(v)^{\text{out}}} w_a - \sum_{a \in A(v)^{\text{in}}} w_a = \sum_{h \in H(v)^{\text{out}}} r(v,h)\, x_h \quad \forall v \in V \setminus \{D\}, \tag{7}$$

$$\sum_{a \in A(m)^{\text{out}}} w_a = \sum_{h \in H(m)^{\text{out}}} r(s,h)\, x_h \quad \forall m \in M, \tag{8}$$

$$w_a \in [0, U] \subset \mathbb{Q}_+ \qquad \forall a \in A, \tag{9}$$

$$x_h \in \{0,1\} \qquad \forall\, h \in H. \tag{10}$$

The objective function 1 minimizes the sum of the operational cost of all chosen hyperarcs. This includes all cost for operating a trip, deadhead trips, performing maintenances, and costs to penalize irregularities. The first three sets of constraints 2, 3, 4 ensure the covering of each trip, start position or end position. Equations 5 take care about the (hyper)flow conservation. The following four sets of constraints deal with the vehicle maintenance. First, the maintenance variables w were coupled to the hyperarc variables allowing only those to be used for which a hyperarc was chosen. Followed by equations 7 which ensure the correct aggregation of the maintenance resource consumption. The constraints 8 state the possibility to reset the resource flow at maintenance service locations. Finally, the variable domains are given by 9 and 10.

3 Case Study for October 2017

We test our algorithmic approach on real world scenarios of the DB Fernverkehr AG covering the different rotation plans for different sets of ICE-1 vehicles. To do that we use our rolling stock rotation optimizer ROTOR with the modification to the model from section 2. For further information on the implementation and algorithms of ROTOR see Borndörfer et al. (2016); Reuther (2017). All computations were performed on an Intel® Xeon(R) E3-1245 v5 @ 3.50GHz CPU with eight cores and Gurobi 8.1 as LP and sub-MIP solver.

We consider two different scenarios related to the tunnel breakdown. First, we focus on the period between August 12th and October 2nd, i.e., the period where the tracks between Rastatt and Baden-Baden were closed. As this period lasts for roughly eight weeks it was considered as some kind of a regular period comparable to a (planned) longer maintenance

period. Therefore new or re-optimized cyclic vehicle rotation plans were computed for each affected vehicle type that were valid from August 12th and October 2nd. We call these instances *period* instances. The second set of instances deals with the situation after the tracks were reopened. Thus, there should be a smooth transition between the rotation plans for the maintenance period and the normal, i.e., undisturbed timetable. Thus, we considered scenarios that optimize the vehicle rotations between September 27th and October 10th. We call them *transition* instances.

3.1 Period Scenarios

As mentioned earlier these scenarios consider a cyclic time horizon of one week, such that the rotations plans could be repeated as long as the tunnel is closed. Thus, we drop all start and end constraints in our model to compute solutions for these scenarios. Additionally, these scenarios arise from the normal timetable that was offered by DB Fernverkehr AG by removing all passenger trips that were operated in Switzerland or south of Rastatt, i.e., all ICE trains going from Karlsruhe to Basel were stopped at Rastatt. Therefore a (maybe non-optimal) solution for these scenarios exists by taking the obsolete vehicle rotation removing all train movements south of Rastatt and connecting the last movement before a removed one to the first one after a removed one. As these scenarios contain less passenger trips, respectively less operated kilometers (roughly 90% of the accumulated trip kilometer of an undisturbed week), it is not clear how good these solutions are. Moreover, there is a maintenance facility near Basel for the ICE fleet which is as a consequence of the tunnel breakdown disconnected from the remaining network where the trains are running. Thus, it is not clear if the heuristically constructed solution is at least feasible. Additionally, planned maintenances at this facility had to be compensated by other facilities.

Table 1: Results for the *period* scenarios.

ID	$\|T\|$	$\|H\|$	Reg. Veh.	DM	Rev. Veh.	Heuristic Cost($\times 10^x$)	Optimized Cost($\times 10^x$)	Imp.(%)	CPU(s)	Gap(%)
1	37	0.01m	3	0	3	0.138	0.138	0.0	0.81	0.34
2	258	0.46m	16	1	16	0.918	0.908	1.0	246	0.16
3	582	1.88m	32	1	32	2.285	2.280	0.3	1958	0.64
4	889	18.20m	52	2	50	3.443	3.352	2.7	15053	0.50

Table 1 shows the computational results of the period instances for different sets of the ICE-1 train fleet. The first two columns identify the instance itself by an index and its number of included passenger trips. After that the total number of hyperarcs required to model the respective scenario with our approach is given in column '$|H|$'. The columns 'Reg. Veh.' and 'Rev. Veh.' show the number of required vehicles to cover the regular and the revised timetable in an optimal way. Numbers in column 'DM' mark the number of maintenances of the reference rotation that could not be reached anymore. The two 'Cost' columns list the operational cost of the heuristically constructed solution where trips passing Rastatt were shortened and the best solution found by our approach. The improvement of the latter over the first solution is given in the 'Imp.' column. Finally, the required computation

time and LP-IP gap are given by the last two columns. Focusing on the number of vehicles required for this period it is possible to operate the disturbed timetable by the same amount of vehicles. This is not trivial because of the disconnected maintenance facility, but also more expected than surprising. It makes sense to consider the optimized rotations especially if they contain less vehicles as for instance 4, since that frees vehicles for other purposes, even if the heuristically constructed solution is maintenance feasible. The numbers for required vehicles and the cost of instance 1 shows exactly the case where you could not do anything better than shortening the trips passing Rastatt and use the heuristically constructed solution. Keeping in mind that this period lasts for roughly eight weeks, operating a rotation with 1.0 or 2.7% decreased operational cost is highly desired and leads to noticeable total cost savings. The last two columns show that the approach can come up with near optimal solutions in reasonable short computation times. Thus, with an automated approach to recompute rotations for disturbed scenarios planners have a strong tool at hand to react to the new situation and to qualify their solutions.

3.2 Transition Scenarios

These scenarios model the exact situation between September 27th and October 10th. Thus, we consider an acyclic time horizon of two weeks, with start and end conditions for the vehicles of the fleet. The timetable considered in these scenarios is composed of the disturbed timetable of the disruption period for the first week and the regular timetable for the last week. Additionally, the 3rd of October is a public holiday in Germany. In 2017 this was a Tuesday. Therefore, DB Fernverkehr AG offered a limited number of trips on Tuesday and on Monday, due to a limited demand for train rides on these days. This leads to a somehow irregular period within the timetable. Nevertheless, regularity is always a very desired property in the railway industry. It holds for timetables in the sense of regular, i.e., periodically repeating connections or arrival and departure times. The same holds for rotation plans. To optimize towards regular rotation plans, we consider reference solutions for each of our scenarios. It is composed of two parts the first part is the optimized rotation plan for the respective *period* scenario and the second part is an optimized rotation plan for undisturbed timetable that should be operated again after the 3rd of October. Thus the vehicle locations at the beginning of September 27th and the end of October 10th with respect to the solutions are the start and end conditions for our model. Additionally, there is a penalty for each deviation from any vehicle movement included in the reference solution.

Table 2: Results for the *transition* scenarios.

ID	$\|T\|$	$\|H\|$	Veh.	M	Dev.	Dev M.	Cost($\times 10^x$)	CPU(s)	Gap(%)
1	95	0.07m	3	7	6	1	0.314	26	0.00
2	545	1.5m	16	44	52	1	2.053	871	0.84
3	1173	7.1m	32	103	137	2	4.741	7166	1.00
4	1854	74.7m	52	159	175	4	7.223	49311	0.55

Table 2 shows the computational results of the transition instances for different sets of the ICE-1 train fleet. Analogue to the previous table the first three columns identify

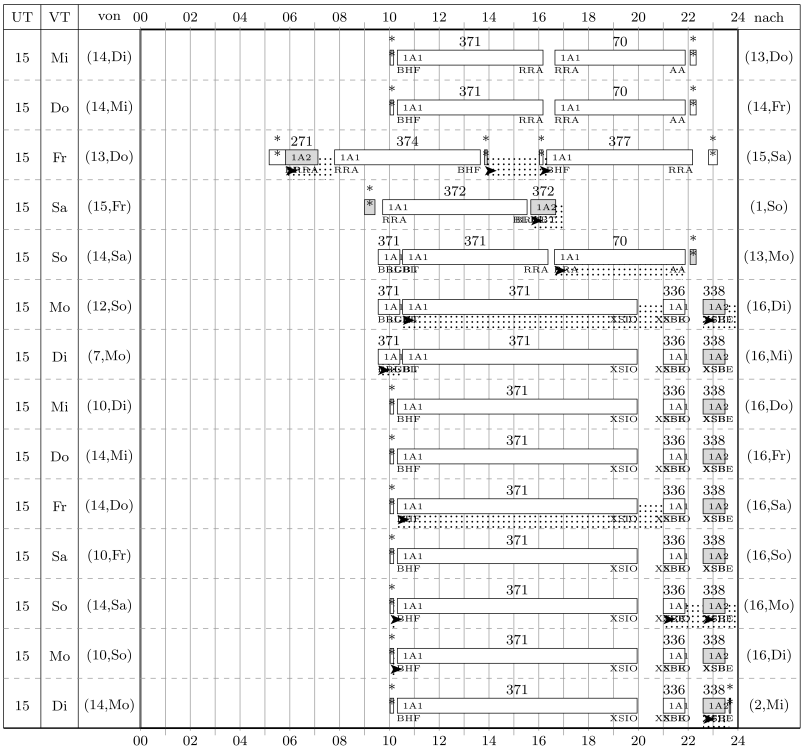


Figure 2: Visualization for a single vehicle contained in the rotation plan of the ICE-1 train fleet.

the respective instance and give the number of included trips as well as the number as required hyperarcs to model the problem. Numbers in columns 'Veh.' and 'M' show the number of required vehicles respectively planned vehicle maintenances for the scenario. Column 'Dev.' marks how many trips of the reference solution have different succeeding trips in the optimized and in the reference solution. A quite similar number is given by 'Dev. M.' as it is the number of maintenances of the optimized solution that deviate from maintenances of the reference solution. The last three columns show again operational cost of the solution, runtime of the approach, and the LP-IP gap. The first observations for the transition instances is that all instances could be solved nearly to optimality. Focusing on the number of trip and maintenance deviations, the optimized solutions keep on average over 90% of all trip connections and over 95% of the planned maintenances of the respective reference solutions. This leads to the conclusion that the vehicles movements during the time horizon of the scenarios are very similar to the movements before and afterwards.

Figure 2 shows a subset of the optimized solution for the ICE-1 train fleet containing trip sequences of a day of operation associated with train 371. In more detail all boxes headlined by a number show a trip with this number. Boxes in the same line succeed or precede the other starting at midnight on the left at ending at the following midnight on the right. The color of a box gives the orientation of the used vehicle, i.e., a white bow shows

a vehicle with its first class in front of the second class with respect to the driving direction and gray boxes for the opposite orientation. The abbreviation on the bottom of a box, e.g., 'BHF' (for Berlin Ostbahnhof, in the north-east of Figure 1) gives the arrival and departure stations of the trip. Focusing on trip '371' one can see that the same orientation and vehicle configuration was chosen to operate it over the complete planning horizon although it has different arrival stations in the beginning ('RRA' for Rastatt) and end ('XSIO' for Interlaken Ost, in the south-west corner of Figure 1) of the planning horizon. The vizualization of the optimized solution was done with the techniques discribed in Borndörfer et al. (2019).

4 Conclusions

We presented a case study of a real world scenario of a heavy disruption of the German railway network. This long lasting disruption had a major influence on many parts of the railway system. We presented an approach how to deal with this kind of disruption and how to compensate it. The approach is capable to deal with the size of a nation wide scenario and leads to near optimal solutions in reasonable time.

Acknowledgements
This work has been supported by the Research Campus MODAL Mathematical Optimization and Data Analysis Laboratories funded by the Federal Ministry of Education and Research (BMBF Grant 05M14ZAM). All responsibility for the content of this publication is assumed by the authors.

References

Borndörfer, R., Reuther, M., Schlechte, T., and Weider, S., 2016. "Integrated optimization of rolling stock rotations for intercity railways". *Transportation Science*, vol. 50., pp. 863-877.

Borndörfer, R., Grimm, B., Reuther, M., and Schlechte, T., 2019. "Optimization of Handouts for Rolling Stock Rotations Visualization", *Journal of Rail Transport Planning & Management*.

Cacchiani, V., Huisman, D., Kidd, M., Kroon, L.G., Toth, P., Veelenturf, L., and Wagenaar, J. C., 2014. "An overview of recovery models and algorithms for real-time railway rescheduling", *Transportation Research Part B: Methodological*, vol. 63, pp. 15-37.

Lusby, R. M., Haahr, J. T., Larsen, J., and Pisinger, D., 2017. "A Branch-and-Price algorithm for railway rolling stock rescheduling", *Transportation Research. Part B: Methodological*, vol. 99, pp. 228-250.

Reuther, M., 2017. *Mathematical Optimization of Rolling Stock Rotations*, PhD thesis, Technische Universität Berlin.

Understanding the Impact of Driving Styles on Reactionary Subthreshold Delays on a Fixed Block Signalling System

Oliver Bratton [a], Giorgio Medeossi [b],
[a] MTR Corporation Ltd
[b] trenolab

Abstract
Train punctuality in the UK is focussed on measuring the time trains are booked to pass a fixed point and when that event occurs. What is not considered in this measurement of performance is whether the capacity of the system is being optimised. It is posited in this paper that performance needs to consider how closely the delivered train service matches the minimum time signals should be red for that pattern of train services. Any changes to the operation of the system that cause the signals to be red for longer than necessary will decrease system capacity and this will have a detrimental effect on *delay per incident*.
This paper compares on-train data recorders (OTDR) from 2002 and 2018. It shows that average braking rates have declined from 4%g to 3.5%g. This will typically add 4 seconds per stop. Train lengths in the UK have also increased in this time, with a typical train length increase being from 8 to 10 cars. If the slower braking curves and longer trains are combined, and a hypothetical block joint positioned 300m from a stopping point, it can be shown that the signal in rear will take 8 seconds longer to clear on average today than in 2002.
While the impact of these changes on time at destination can be easily demonstrated using distance/time graphs, the effect on the signalling system is more complex. The simulation system *trenissimo* has been used to show that the effect on a system of longer trains and slower braking curves is [x], with the system responding in a non-linear way to very small changes in train operations. It is posited that this is key reason for the increase in delay per incident currently being seen in the UK.

Keywords
Braking, Capacity, Performance

1 Introduction

A railway is a balance of journey time, intensity of service (given infrastructure constraints) and service reliability. While the general principles of how these interplay are axiomatic to operations management, the precise mechanism by which one affects the other is less well understood. In the UK, the primary focus has been 'PPM' - the per cent of trains that arrive at destination within 5 minutes of booked time (10 minutes for long distance operators) having called at all booked stops.

PPM does not measure how effectively a train uses the available capacity of the system. One way to improve the PPM metric is to increase the planning time between stations; this increases the probability that a single train will arrive at the timing point 'on-time', even if

it has been subject to a delay en-route. While this works for individual trains, it does not consider the interaction of train movements with a fixed-block signalling system. A fixed-block signalling system is designed to operate with trains at a given speed; the greater the difference of actual train speed against this optimised speed, the longer signals will display a red aspect.

When the network is considered as a system (rather than as a collection of individual trains moving against the timetable), performance becomes not only the ability of trains to cover a distance between two points in a given time, but the interaction of trains. This is becomes increasingly important as the system approaches capacity, or when there is an incident and trains operate to the signalling system rather than to a timetable.

UK railways have seen an increase over the past 7 years of two metrics - 'Delay per Incident' (DPI) and 'sub-threshold delays'. This means that delays affecting passengers have increased, even though the number of incidents has declined. This has largely been attributed to increase in the number of services in operation and an increase in passenger numbers (affecting dwell times). There are parts of the UK system, however, where there have not been significant timetable changes and where passenger numbers have decreased in recent times—yet performance has still declined, even with fewer infrastructure incidents. This suggests that the increase in DPI and sub-threshold delay is not only caused by passenger numbers and/or increased numbers of trains but that other factors are affecting how the system performs.

1.1 Background to train operations in Wessex

The Wessex route in the UK (operated by First-MTR South Western Railway) connects London with the south west of England, as shown in figure 1. The terminal at London Waterloo is the busiest station in the UK; 8 tracks connect it with Clapham Junction, from where four continue to Reading, and four — the South Western Main Line — towards Southampton, Bournemouth and Weymouth. Additional branches connect the line with other centres in the region, such as Portsmouth and Exeter.

This intercity traffic uses two of the four tracks available between Woking and London, with the other two dedicated to local services: in peak hours the same fast track is used by 24 trains per direction. Despite the high density of block sections, with such high traffic density the probability of unplanned braking actions due to restrictive signal aspects is quite high; the lower speed cause a further increase of the blocking time, which propagates to the following trains.

The timetable on the Wessex route in the UK has been largely unchanged since 2004. Although the rolling stock in use is principally the same, the formations have changed with trains being typically increased from 8 cars to 10 cars on suburban routes. The infrastructure layout has remained the same with the exception of Waterloo which saw some remodelling in 2017. Despite considerable efforts by Network Rail and the train operator, performance has declined steadily since 2010 (figure 2).

Passenger numbers increased on Wessex have increased steadily up to 2017 but recent data from the Office for Rail and Road shows a decrease in journey numbers by 7.9% in the past year to 212 million per annum (the lowest level since 2012/13) yet there has been no corresponding improvement in performance.

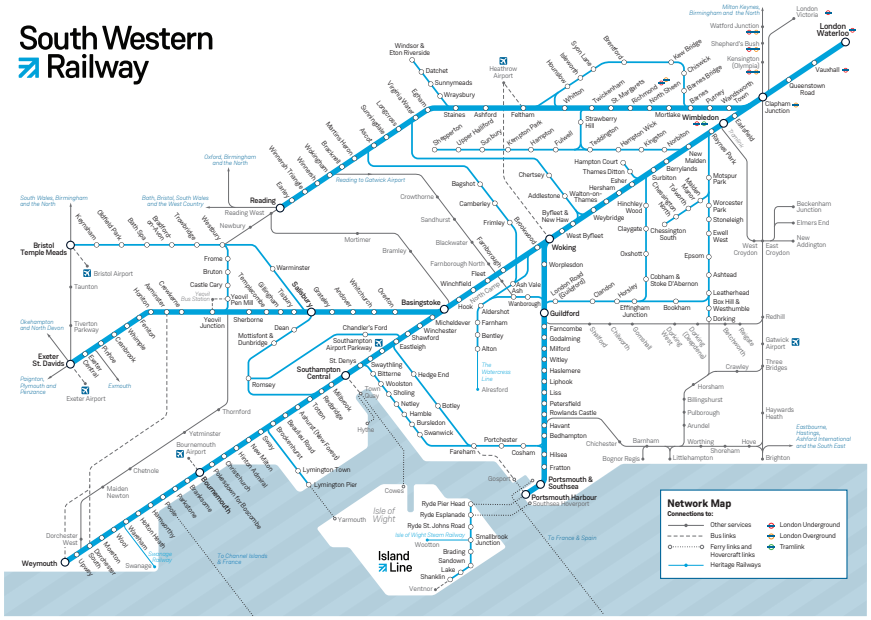


Figure 1: South Western Railway routes

1.2 Identifying changes to operations

The changes in performance in Wessex have been subject to numerous recent reports but these focus on the reliability of the infrastructure [ref to Holden report] rather than on how the train operator is delivering the service. This is partly because asset failures are readily identifiable (typically these cause delays that the system can easily identify) whereas train operations and station delays are often much smaller and harder to identify, despite being more prevalent. While there is a more work taking place to identify where these 'sub-threshold' delays are occurring, there is very little historical data to show if these have changed.

In this study, Hasler TELOC on-train data recorders have been analysed from files extracted in 2002 and 2018 and the braking curves from the data sets compared. These changes have then been simulated to quantify how much of the decline in recent performance can be attributed to the changes in how a train operated uses the available capacity of the network. The focus of the study is on how drivers brake to a halt. This is because this variable is wholly within the control of the train operator and is an action repeated continuously on a train's journey. Even a small change in braking style is likely to manifest itself when repeated for every station stop or restricted aspect. Furthermore, the braking of rolling stock conforms to standards and therefore a step one application (i.e. 3%g) in 2002 will be consistent with a step one application in 2018.

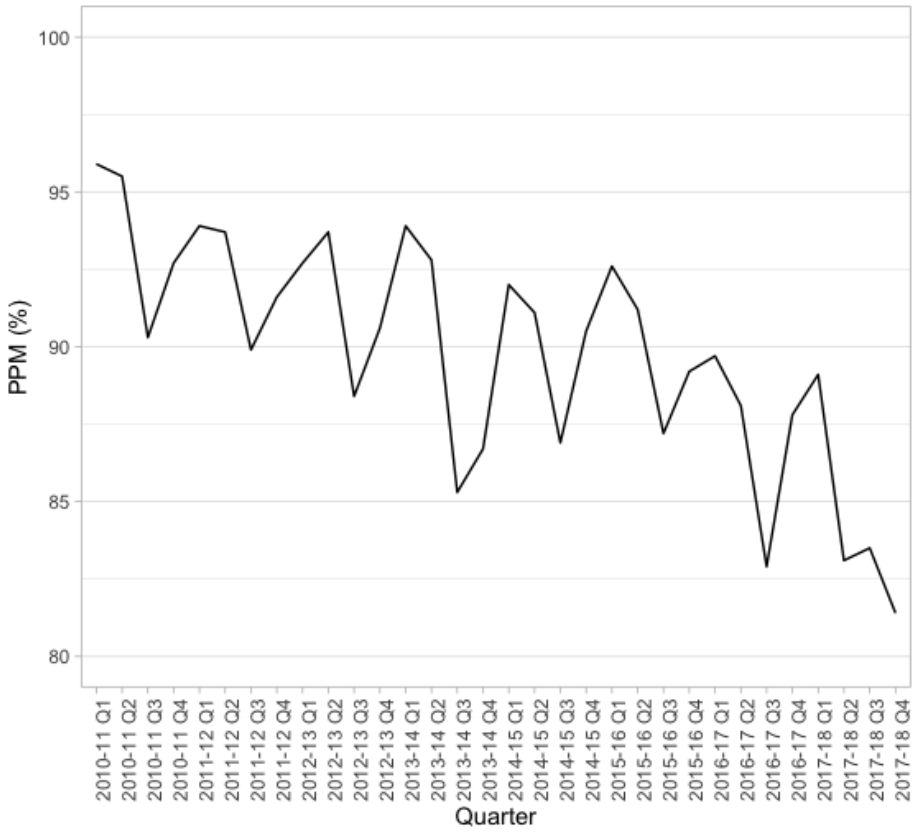


Figure 2: South Western Railway performance since 2010

2 Methodology

2.1 Sources of 2002 data

Routine down-loading of OTDR has only become common in recent years. In the early days of OTDR, it was necessary to physically connect a laptop to the train to obtain a file, and then save the data to a floppy disk, making data collection onerous. Most files were therefore obtained to investigate an operational incident or a unit defect. No policy existed to preserve these files. The author, however, as part of a previous role in 2002 working between the then Infrastructure Manager (Railtrack) and Transport Undertaking (South West Trains), had access to some files obtained as part of safety investigations. The files came from class 458 (Alstom Juniper) units introduced in Wessex around 2000 and one of the earliest new fleets fitted with OTDR as part of the build specification.

The files were uncovered as part of a review of archived data by the authors in 2018.

2.2 Processing of 2002 data

The recovered data archive included 6 OTDR files and a copy of Hasler TELOC 2.0 software. The files were processed in TELOC and saved as .LTM files (a native TELOC text-based format). These files were the processed in Unix using a bash script to remove extraneous lines of text in order that the files could then be saved as .CSV files. Hasler files were then limited to approximately 65,000 rows; this tended to equate to about 24 hours' worth of data per file.

2.3 Sources of 2018 data

MTR is the parent company of MTR Crossrail. MTR Crossrail currently operates the TfL Rail services in London on behalf Rail for London. The TfL Rail services will become the Elizabeth Line on the opening of the Crossrail tunnels through central London.

MTR Crossrail operates class 345 (Bombardier Aventura) units. These are equipped with Hasler OTDR and TELOC software. MTR Crossrail provided data from the class 345s from February 2018. The files were processed through TELOC and exported to Microsoft Excel in order that the outputs for speed and time could be converted to .CSV format. Due to a limitation of the software, it was only practical to parse a subset of the total available class 345 data to .CSV format.

Although 2018 and 2002 data come from different rolling stock and routes, it should be noted that they operate on similar suburban railways. Station distances and line speeds are comparable for the two railways.

2.4 Extraction of braking curves

The R data.table package was used to process the .CSV files to extract the braking curves with ggplot2 being used to produce graphics. The start of the braking curve has been defined as being the maximum speed in the last 90 seconds prior to stopping. Curves were only included where the maximum speed was between 50 and 65 mph. [The scripts are attached as an appendix?]

There is no geographical context to the class 458 braking curves. The OTDR has a distance column but there is no means of identifying precisely to where those distances apply. The class 458s were used on the Waterloo - Reading and the Waterloo - Alton services but there were multiple variations in stopping patterns that make it hard to interpolate where the train is from the data available. Each curve could therefore be subject to variables such as railhead adhesion, gradient and track curvature but these have not been factored into the work.

Location information is available from the class 345 OTDRs but, for consistency with the work on the class 458s, this has not been included.

Identifying the train stopping point

It is extremely difficult to identify the exact moment a train wheel stops rotating from either GPS or the OTDR. In the class 458 data in particular, the algorithm appears to hold the last known rotational velocity of the wheel rather than replace the value with a zero. Since all other metrics (such as distance and acceleration) are derived from the wheel rotation, the train will often never appear to become stationary. This makes it difficult to trust the data

below 2.5 mph (1.1176 ms^{-1}); the train speed often stays above 0 mph even when it is known that the train is stationary.

A stopping point has therefore been defined as being (for the class 458 data) 3 seconds after the train speed has decreased below 3mph. For the 345s a combination of the position of the power brake and train speed has been used since the odometer readings show higher accuracy (but still not clearly reaching an absolute 0 mph), with the stopping point being assumed as being 1 second after the train speed decreases to 1 mph. GPS data also shows limitations in identifying the exact stopping and so has not been used for this work.

2.5 Changes to driver training policy

There have been changes in the driver training policies in the UK over the past 20 years in response to incidents involving drivers failing to stop at red aspects [ref to Ladbroke Grove, Southall,], and to accompany the fitment of Train Protection Warning System. The mitigation for preventing SPaDs including braking on sight of restricted aspects, not entering a platform at more than 30 mph, not exceeding 20 mph at 200 yards (approximately 200m) from the signal; and stopping 20 yards short of the signal. Furthermore, drivers were discouraged from using step three braking (i.e. 9%g) and taught to only brake in steps 1 & 2 (3%g and 6%g respectively). Whereas, on suburban systems, drivers used to drive at line speed on double yellows (it being possible to brake to a stand from line speed from sighting the single yellow), drivers are now required to start braking on sight of a double yellow. Despite the changes in driving styles, there has been no corresponding changes to signalling design specification to account for these changes.

2.6 Simulation of the network

Simulation tool

The *trenissimo* simulation programme (see 3) has been used to simulate the small changes in driver style. The tool (de Fabris et al., 2018) has been developed by trenolab. It is a synchronous, microscopic simulation tool, aimed at reproducing railway operations in the most accurate way, with a special focus on the representation of stochastic factors, such as the dwell times and the variability of running times. *trenissimo* is a Java application natively compatible with all operating systems. It was developed using the Netbeans Platform, a framework designed to create a very flexible and user-friendly environment.

The tool reproduces railway operations in a mixed continuous-discrete approach: it calculates the solution well-known motion equation (Wende, 2003) of trains in a continuous way, considering the discrete processes of signal states. At present, *trenissimo* features the Italian, French, British and Norwegian signalling systems, as well as the ETCS Level 1 and 2.

One of key strengths of *trenissimo* is that the dispatcher is simulated: as in the real world, while automatic block signals are automatically set to green, a dispatcher oversees the operation, opening the home and exit signals based on the planned and actual positions of trains. As a result, and similar to real operations, the dispatcher always controls operations: he can take decisions based on simple rules, or more complex algorithms. This principle allows implementing robust deadlock-prevention algorithms, as well as testing the effective impact of different dispatching strategies. Based on a set of rules, the dispatcher is also able to cancel train services, or short-turn them to reduce the propagation of delays.

Following the principles explained in (Medeossi et al, 2018), the key stochastic inputs for an accurate simulation are the initial delays, the dwell times, and the variability of running times. *trenissimo* implements the combination of stochastic dwell times and departure inaccuracies proposed in (Longo, 2012) to accurately represent the dwell times of the early- and late-arriving trains. The dwell time is considered as the stop time related with the exchange of passengers and this is applied to all trains stopping at a station, while the departure inaccuracy represents the departure variability of trains that arrive early at the stop, but do not depart on time due to an overlong departure process, or to passengers arriving at the last second.

In previous work (Medeossi et al, 2011) it was demonstrated that to represent accurately the running time a set of parameters is required, each representing the way drivers drive during one of its phases: acceleration, cruising, braking and coasting. Additionally, ideally braking at stops, signals and speed restrictions would be considered separately though this is not currently implemented. The work also proposed and tested a method to estimate the distribution of these parameters based on GPS or OTDR data, which has been used in this study as an input for the simulation.

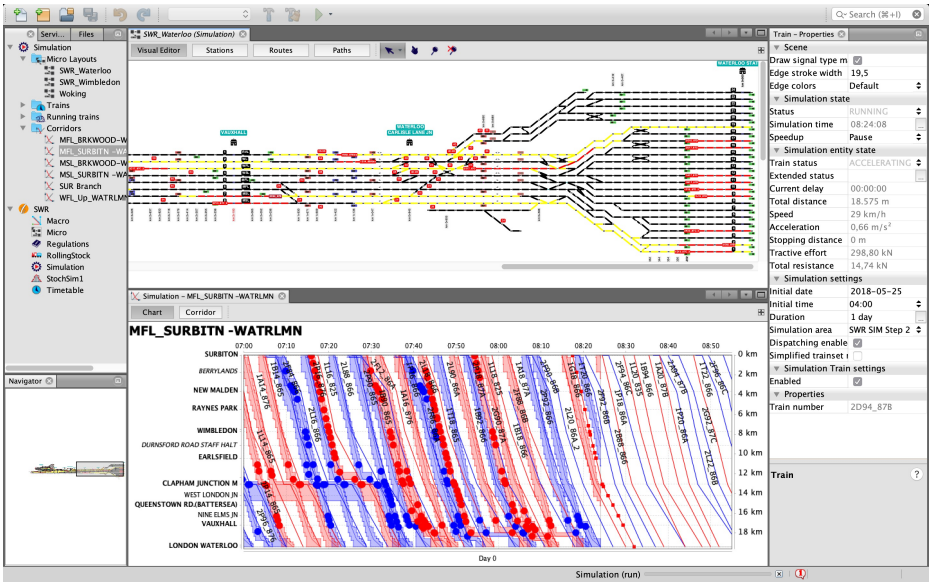


Figure 3: Screen shot of trenissimo system

2.7 Simulation of the system

The May 2018 morning peak hour timetable of the Woking - London Waterloo section of the South Western Main Line was simulated using the real distributions of input delays, dwell times and departure inaccuracy derived from track-circuit logs. The input delays, which are the distribution of departures from the first ocp within the simulation area, were filtered to remove secondary delays (Medeossi et al., 2011), while the distributions of departure

inaccuracy and dwell times were respectively obtained considering only record of late- and early-arriving services, plus the variability of driving styles obtained from the analysis of OTDR data.

Three scenarios were simulated. The first represents 2018 operations, in particular considering the 2002 braking behavior and shorter train formations. The second scenario instead considers the 2018 braking behavior, while the third one combines it with the 2018 train formations.

The simulation of each scenario was repeated for 200 times using a Monte Carlo approach; the occupation time of selected block sections during the peak hour (08:00 – 09:00) and delay indicators (mean delay and punctuality at 5') at arrival at Waterloo in the were used as KPIs of each scenario.

3 Results

3.1 Comparison of braking curves

Figure 4 shows the spread of speeds of braking curves at three second intervals to a halt from the class 458s in 2002 and the class 345s in 2018.

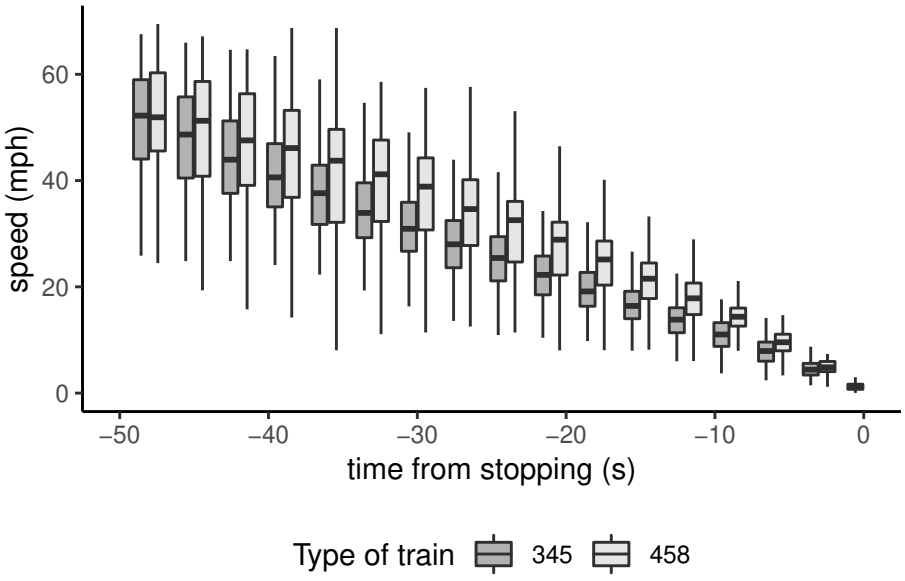


Figure 4: Braking curves from 2002 (class 458s) and 2018 (class 345s)

At 50s prior to stopping there is very little difference between the spread of results for the two data sets but the class 345s start braking sooner and, at -30s to stopping, the median for the 345s is 8mph slower than the 458s. When comparing the median for the two data sets, the class 345s take 4 seconds longer than the class 458s to halt.

3.2 Implications of changing train length

The class 458s were bought as four-car units. For peak services they would be coupled together to make 8 car trains, totalling 164m in length. In 2016 the trains were lengthened to five-car sets to provide increased peak capacity. When coupled, these are now 204m long.

It is easy to calculate the increase in time it takes for the longer trains to clear a given track circuit, and the commensurate time the signal in rear stays at red. This needs, however, to be combined with the changes in braking rate. If it is assumed that there is a block joint 300m from the stopping point, the slower braking curve and the longer train means that the protecting signal in rear will be red for 8 seconds longer than it would have been in 2002.

For a suburban railway with a planning headway of 150 seconds, this represents a near 6% loss in capacity.

3.3 Network Rail modelling assumptions

Network Rail has a standard for the modelling assumptions to be used in RailSys simulations. Braking of Electric Multiple Unit trains is required to be modelled at 0.588 ms^{-1} (8%g). It is apparent that this was not met in 2002, and the discrepancy has increased since then.

Figure 5 shows the braking curves of the class 345s in 2018 compared with the expected trajectories from step 1 braking (3%g) and step 2 braking (6%g). This shows that, even though the class 345s have a continuous brake, the drivers continue to operate within the range that they are used to driving stepped brakes. This shows the extent to which drivers *feel* braking.

The Network Rail assumption for RailSys of 8%g is met in only one instance. This is unsurprising given that drivers are taught to avoid step 3 braking. Since braking typically involves amending the braking curve at some point before stopping, and drivers must be assumed to be using less that step 3 braking, it is extremely unlikely that a driver will be able to average 6%g since that would apply using the same step 2 brake throughout the whole braking curve. Instead, we see the outcome of a mix of step 1 and step two braking, averaging considerably less than the 8%g assumed by Network Rail.

3.4 Impact of changing train length and braking curves on performance

textbf[Note that these simulations are to be re-run. Currently the simulation is a doubling in train length from 4 to 8, not from 8 to 10 car] Figure 6 shows the mean simulated arrival lateness at Waterloo for the up fast lines. These graphs show the cumulative effects of a loss in capacity caused by the longer trains and slower braking styles, and the exponential rate at which delay accumulates once trains start to interact. It also shows that the system is extremely senstive to changes in train length. Of particular note is the rapid increase in lateness around 08:30 in the morning; the peak time for arrivals into Waterloo. The delays then increase exponentially, only recovering at the end of the peak as the services start to thin.

Figure 7 shows the total delay on route for each train for each of the scenarios. It can be seen how the combination of driving styles and longer trains are in themselves alone sufficient to fundamentally alter the performance of a route. Since fast trains to do not stop at many stations, the impact of the driver styles is much lower than that of the train length.

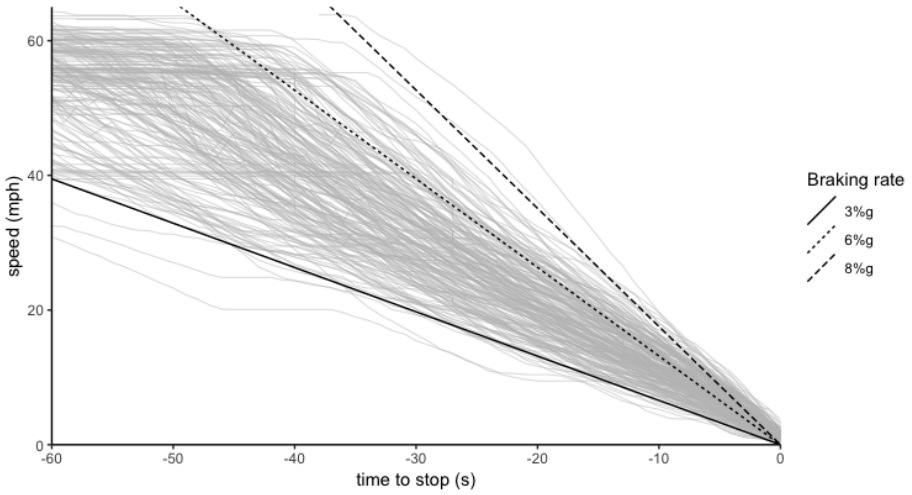


Figure 5: Comparison of 2018 braking curves with braking steps and NR modelling assumptions

The impact of driver styles is, however, almost 50% greater on long trains than it is on short trains.

The increases in overall lateness are shown in table 1 for each of the changes.

Increase from driving style	5.9%
Increase from longer trains	21.7%
Increase from longer trains and driving style	37.1%

Table 1: Changes in delay from 2002 to 2018

4 Discussion

4.1 Validity of findings

[Note of 1 Feb 2019 - these results are of the Up Fast only. The Up Slow results also need to be incorporated. Second, these results are based on doubling of train lengths where as the train formation changes on the fast are less substantial than on the main.]

A number of assumptions have been used for the purposes of the study. The consistency between the sets suggests that these assumptions do not invalidate the findings. The braking curve rates of 2018 are lower than 2002 from similar starting speeds; both data sets are lower than the modelling rates assumed by Network Rail for capacity and performance modelling; and the simulations show that even small changes in braking rates on a network with similar operating constraints leads to an increase in delay per incident and decline in overall performance of the system. It is acknowledged, however, that there are weaknesses of the study caused by not having the same base assumptions. In practice, however, there

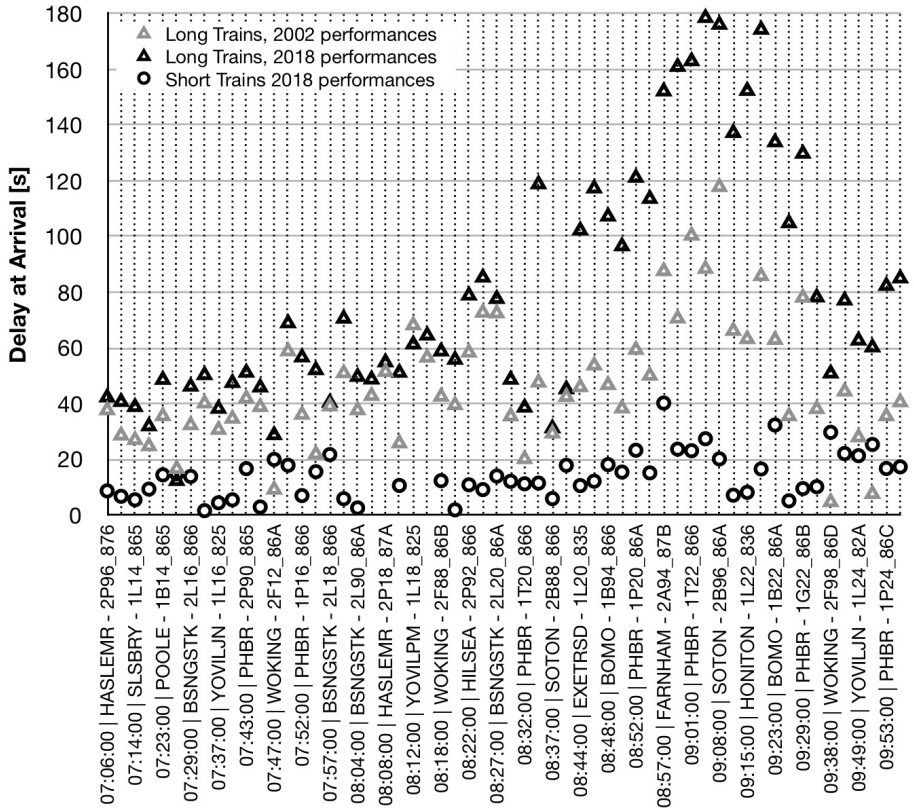


Figure 6: Simulated average lateness at Waterloo for changes to train length and braking style

is very limited data available from 15 years ago and, even where this does exist, it is very unlikely that the same units are operating on the route, or that the route has remained entirely unchanged.

The system shows considerable sensitivity to changes in train length. On an urban system, with relatively short track circuits, this is perhaps unsurprising since it takes the train longer to clear the track circuit at slower speeds.

4.2 Implications

The train operator has the capability to determine the rate at which delay dissipates across the network by varying from the optimum profile at which, for a given train service, signals revert from red to a proceed aspect. This effect is greatest when trains are constrained by the signalling system, rather than by the timetable. This typically occurs in congested routes, at busy junctions, and during disruption. Two factors that have implications for performance but which (in the UK at least) have not been assessed prior to their introduction are the

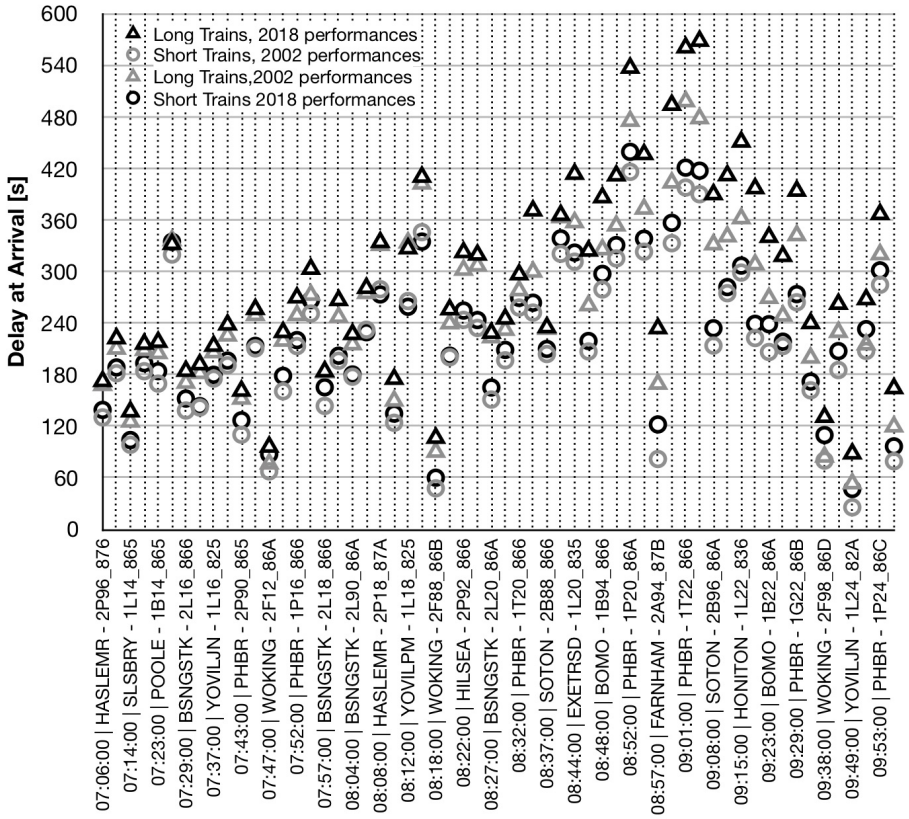


Figure 7: Total delay on route for each scenario

length of trains and speed at which they are driven. Both of these by themselves increase the time in section, but a combination of both greatly increases the delay with slower driving styles and longer trains being 50% worse than slower driving and short trains.

5 Summary of findings

The change in driver styles between 2002 and 2018 have resulted in it taking four seconds longer today to halt a train from 60mph than in 2002. When combined with longer trains, the signalling system takes longer to clear. The *trenissimo* simulations show the sensitivity of the network to these changes. On the up fast, the combined changes account for [x]% of trains being more than [x] minutes late at destination.

References to be completed

References

Albrecht T., Goverde R.M.P., Weeda V.A., Van Luipen J., 2006. "Reconstruction of train trajectories from track occupation data to determine the effects of a Driver Information System", In: Allan J., Brebbia C.A., Rumsey A.F., Sciutto G., Sone S., (eds.), *Computers in Railways X*, WIT Press, Southampton.

Besinovic, N.; Quaglietta, E.; Goverde, R.M.P., 2013. "A Simulation-based Optimization Approach for the Calibration of Dynamic Train Speed Profiles", *Journal of Rail Transport Planning & Management 3*, (2013).

Cui Y., Martin U., Zhao W., 2016. "Calibration of disturbance parameters in railway operational simulation based on reinforcement learning", *Journal of Rail Transport Planning & Management*, vol 6, issue 1.

de Fabris S., Medeossi G., and Montanaro G., 2018. "trenissimo: Improving the microscopic simulation of railway networks", In: *Proceedings of the 16th International Conference on Railway Engineering Design & Operation July 2-4 2018*, Lisbon.

Medeossi, G., Longo, G. and de Fabris, S., 2011. "A method for using stochastic blocking times to improve timetable planning", *Journal of Rail Transport Planning & Management 1*, (2011), pp. 1-13.

Powell, J., Palacin R. (2015). "A comparison of modelled and real-life driving profiles for the simulation of railway vehicle operation", *Transportation Planning and Technology*. 38. 10.1080/03081060.2014.976984.

Pre-planned Disruption Management in Commuter Railway Transportation: Algorithms for (partial) Automation of passenger-oriented Design and Evaluation

Anna-Katharina Brauner [a, 1], Andreas Oetting [a]
[a] Chair of Railway Engineering, Technische Universität Darmstadt
Otto-Berndt Straße 2, 64287 Darmstadt, Germany
[1] E-mail: brauner@verkehr.tu-darmstadt.de, Phone: +49 (0) 6151 16 65916

Abstract
A pre-planned disruption management helps to amend disruption operations. Pre-planned train dispatching instructions, unified in a so-called disruption program for a typical disruption situation facilitate the work of the dispatchers. Those instructions are mostly manually designed and often focus solely on the train runs. The proposed approach aims to improve the quality of disruption programs concerning operations and especially concerning the reduced passenger mobility. For this purpose, the algorithms to be presented evaluate the operating concept on its functionality and transition capability in a solely train operations focused way. A stable and fast transitioning disruption program is already enhancing the passenger mobility in a disruption, but this is not enough to call it passenger-friendly. The goals of a fast transitioning, realized by a low number of train runs and the quality of the passenger's mobility are strongly conflicting. For this purpose, the algorithms design a transportation concept including passenger guidance measures and comprise a final evaluation of the disruption program in a passenger-oriented way.

Keywords
disruption management, disruption programs, passenger-oriented, operating concept, transportation concept

1 Motivation

Major disruptions in commuter railway transportation alter railway operations significantly and affect passenger mobility tremendously. Pre-planned disruption programs (DRP) contain a substantial amount of train dispatching decisions for the event of a disruption. They facilitate the work of the train operating company's dispatchers since the operating concept for the disruption is known. Therefore, measures can be taken and communicated quickly.

Currently, highly experienced employees are responsible for drawing up the DRP. They design the operating concepts manually and use their dispatching experience to foresee effects and interactions of the potential disruptions and especially the operational measures. Reaching a stable - smooth, punctual and reliable - operation is the first main aim in disruption operations. The second aim is to reach this state of operations as fast as possible. So far, this can only be estimated by an employee on the basis of his own experience. Therefore, it is not thoroughly ensured that the DRP consists of an actually viable set of operational measures for the typical disruption.

The manual design of an operating concept and the experience-based consideration of interdependencies is highly time-consuming; therefore, they mostly leave passenger

guidance measures aside. A fast transitioning, realized by a low number of train runs, and the quality of the passenger's mobility are strongly conflicting goals. This conflict is not yet solved in the creation process, as the focus and the expertise are mainly on the dispatching side.

Next to the high effort of designing DRPs, there is another reason why they are lacking a thoroughly prepared passenger guidance. The experience based on rough estimation whether the concept could be working, should not be the basis for a transportation concept. To enhance DRPs covering also passenger guidance measures, the operational functionality itself needs to be ensured first, because constant ad-hoc dispatching interventions might generate deviations conflicting with planned guidance measures. This would then result in an unstable disruption situation for the passengers with unreliable information.

2 Related Work

Real-time traffic management and support tools targeting dispatching assistance are discussed in Corman and D'Ariano (2011), Ochiai et al. (2016) and Törnquist (2012) for example. Toletti (2018) discusses algorithms for dispatching support concerning dynamic capacity increase. However, dispatching support tools for train operating companies (TOC) mostly focus merely on parts of their dispatching processes, like connection dispatching in Stelzer (2016) and Schütz and Stelzer (2015), as most operational dispatching decisions of a TOC have to be accepted by the rail infrastructure company first. Next to that, TOCs only have restricted access to the infrastructure data. For those two reasons, it is difficult for TOCs to constitute an overall real-time dispatching tool including occupation conflicts, whereas a pre-planned operating concept can be pre-coordinated with the rail infrastructure company. Therefore, DRPs imply an approach with high practical relevance and actual application especially for the Swiss and German commuter railway networks.

The manual drawing up of a disruption program by applying a well-defined procedure including related operational measures, relevant dependencies to consider and which stakeholder to include was presented by Chu et al. (2012), enabling decision-makers to work in a structured way using the proposed flowcharts. They also introduce different phases of a disruption, as illustrated in Figure 1, including key characteristics according to DRP usage. Subsequently, the causes of delays and the importance of the transition phase are determined in Chu et al. (2013). A steady disruption operation has three key characteristics: all train runs are on the DRP planned tracks, the number of trains is reduced according to the DRP and all train runs operate at the typical level of punctuality as without disruption. Therefore, capacity must not be exceeded. (Chu 2014)

Oetting and Chu (2013) analyzed the transition phase and performed a case study on operational data of two big German urban railway networks, where they identified the main influences on the duration of the transition phase. One main reason for delays is the queuing of trains at and in front of stations, especially at the turning stations in front of the disruption. The generated congestion influences the duration of the transition phase primarily. The

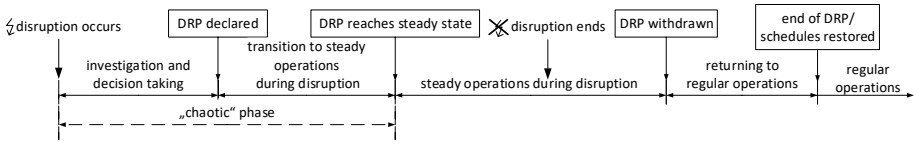


Figure 1: phases of a disruption using DRPs (Chu and Oetting 2013)

initial approaches to perform an operational evaluation of capacity in stations during the transition phase are presented in Chu and Oetting (2013) and Chu (2014), but the transition phase is not yet modelled in a predictable way.

As disruptions lead to limited mobility options, the negative impacts on the passenger's travel routine are important when assessing the dispatching of a disruption. Josyula and Törnquist Krasemann (2017) studied alternative strategies to utilize passenger flow data in re-scheduling. They observed, that re-scheduling models which include passenger flow data, mostly did not choose the metrics according to the changed passenger needs in a disruption. Passenger guidance measures combined into an applicable concept as part of a DRP do not play a significant role the DRP related publications yet.

The literature shows that DRPs are very useful, however the manual creation is time-consuming, and the functional testing is rough and based on experience. Passenger guidance plays mostly a marginal note. A (partial) automation of design and evaluation would be necessary to enhance the benefits of DRPs. The logical and initial mathematical relations for delays, capacity and implemented measures are not yet sufficient as models for (partial) automation. The aim of the research described in this paper is to develop algorithms that ensure the development of operationally functional concepts, which have a reliable passenger guidance and information established on top, to handle the conflict of goals mentioned in chapter 1. The approach to be presented aims to support the design of customer-oriented disruption programs with general validity for commuter rail transport.

3 Research Contribution and Methodology

A disruption program with its applied measures and the following information is the more reliable the faster it works as a whole. To this end, the time required identifying the cause of the disruption, to decide on a bundle of measures and to implement it initially, must be minimized. (Chu and Oetting 2013)

This can become challenging if a disruption program is applied, because the actual advantage of a disruption program can also be a disadvantage: its concreteness. It must fit as exact as possible to the present disruption, because the concreteness enables to process train dispatching quickly, but makes this even more difficult if the DRP does not fit exactly and adjustments have to be made by the dispatcher. These adjustments always have to be made during the transition phase as the actual operating situation cannot be planned as it appears different every minute. Crespo (2018) is dealing with the automation of the non-pre-plannable decisions in the transition phase. This does not counteract with the application of a DRP and is part of the idea including a transition phase. However, the planned stable phase should be applicable as planned.

Therefore, it might make sense to have a large number of disruption programs for many different variants of a disruption. As already mentioned, the effort required to create a disruption program is very time-consuming for now. Therefore, usually DRPs are either very generic or concentrate purely on the most critical points within the network. In addition, the DRPs are limited to the dispatching of trains and focus on the operational events, thus the effects and measures for the passenger remain rather unnoticed. The aim of the concept presented is to achieve a reduction in effort and a quality increase in the creation and evaluation of disruption programs through (partial) automation.

(Partial) automation aims for the evaluation of the operating concept and the creation and evaluation of a transportation concept. Many operating circumstances cannot yet be reliably described with data, like common daily problems or are not available to the TOC as data sets, e.g. freight trains. Freight trains can be very relevant for the disruption

dispatching in German commuter railway networks, as there are many mixed traffic railroads in commuter rail networks existing. Especially when deviating is applied as a measure for the commuter trains, they might use mixed traffic railroads like the Güterumgehungsbahn in Hannover. However, an experienced employee has this background information due to his many years of experience and can incorporate it into a DRP. Therefore, the design of the operating concept remains a task, which has to be carried out by an experienced employee of the TOC, who is familiar with the network. The creation process is ought to be considerably simplified and the quality of the results shall be increased by the underlying automated test algorithms. The DRP designer‘s work is supposed to be supported by a software implementing the presented concept with evaluating his operational planning and creating a transportation concept based on it automatically. Besides the DRP designers, there is also the perspective of the DRP users.

As a user of a DRP, the dispatcher expects the results to ensure operational functionality. Therefore, DRPs need to be operational feasible. Since not all operational data is available for TOCs, relevant assumptions must be included in the developed algorithms. The result must then be edited in an easy understandable way for use of dispatchers, train drivers, the railway infrastructure undertaking (RIU) and others, so that a simple and uncomplicated operational implementation is guaranteed. Misunderstandings and frequent inquiries should be avoided in terms of the workflow during the disruption.

Next to the operational flow in the disruption, should the disruption program be in the passenger’s interest. The passengers also want to fulfill their mobility needs in the event of a disruption. Therefore, a mobility preservation is aimed at. The design and evaluation of DRPs should be carried out from a passenger’s point of view.

The passengers must be informed about the occurrence of the disruption and the subsequent applied operational measures. They eventually have to change their planned travel behavior but despite the disruption, passengers should have to make as few changes as possible to their usual mobility behavior. Therefore, a DRP should not intervene in the connections that are still functioning in the event of a disruption.

A DRP affects many train journeys and many more passengers. In order to implement (partial) automation, the complexity of the interrelationships between operational and transportational events must be simplified to a manageable extent. The algorithms should therefore be clear and efficiently convertible into software. This also includes the supply of data available that is at a TOCs disposal.

As a first step, the evaluation algorithms intend to enable the creation of validated, operationally functional DRPs (solely focusing on the train operations). These so called operating concepts are manually designed and can then be evaluated automatically by the algorithms. In a second step, a functional operating concept can be complemented with

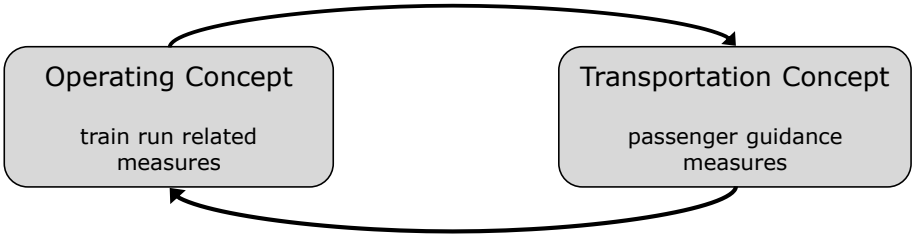


Figure 2: DRPs as a combination of operating and transportation concept

passenger guidance and related communication measures. This so called transportation concept is created and evaluated automatically. Both concepts, as to be seen in Figure 2, are mutually dependent as explained in the following.

Disruptions result in reduced availability of infrastructure and therefore operational measures have to address the availability of infrastructure by omitting the disrupted area. Turning, diverting and parking selected trains reduces a potential capacity over-use. A passenger-friendly disruption management needs as little deviations from regular operations as possible, but if the original timetable is fully preserved, the DRP cannot function in a stable way. A functioning operating concept is the basis for the transportation concept. The evaluation of the transportation concept is used to detect, whether the operating concept is passenger-friendly or whether it needs to be designed differently. A DRP in the future is supposed to be the combination of an operating and a transportation concept to enable a both reliable and passenger-friendly disruption management.

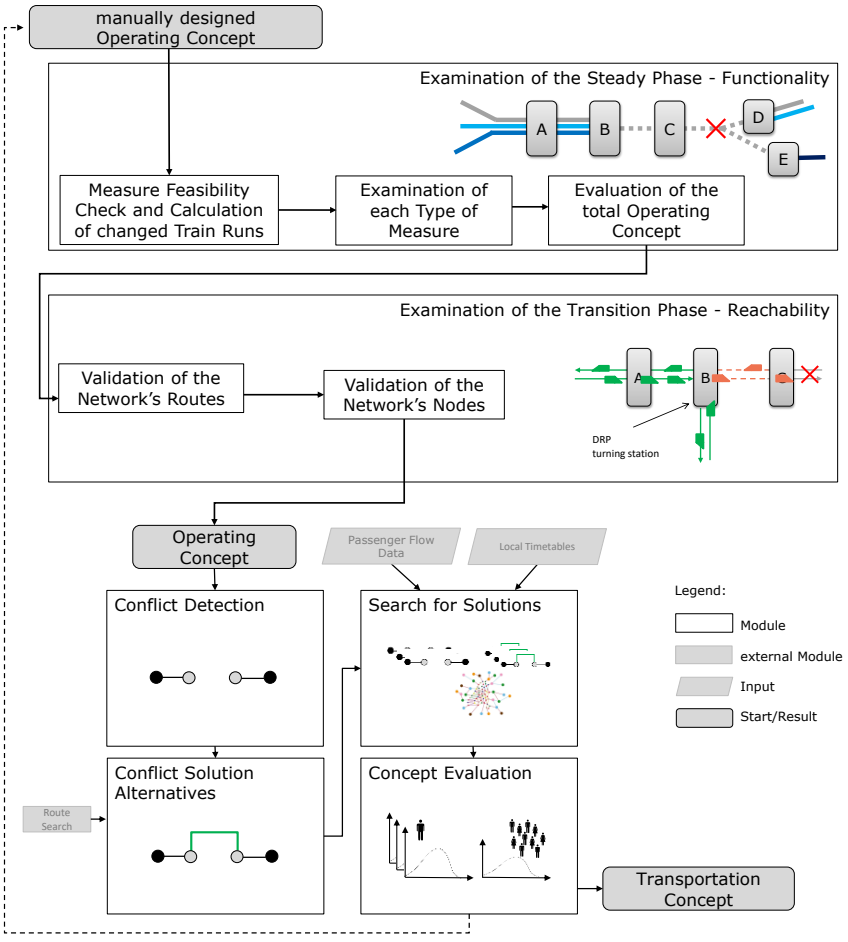


Figure 3: modular structure of the DRP creation and evaluation

To sum up, the evaluation of disruption programs consists of the two superordinate modules: the operating concept and the transportation concept. These two modules can interact with each other cyclically. The transportation concept is built based on the operating concept. In case the transportation concept is evaluated being not sufficient for the passengers, adaptions in the operating concept have to be made and a new transportation concept is built on that. The modular concept and all of its submodules are already shown in Figure 3. In the following, the subordinated modules for these two concepts are derived and the algorithms, which work in these modules and their output are described.

4 Operating Concept

A disruption program pursues the goal of ensuring operational stability. The operating concept ensures that initial effects of the disruption are spatially limited and that delays do not propagate by a rapid application of measures. The operating concept must be able to achieve a stable, delay-free condition: this is called the transition capability. Subsequently, the operating concept should enable to keep up a stable, delay-free, although still disturbed operation: this is called the concept functionality. To this end, the operational measures planned must be capable of being implemented individually and in combination. Thus, the operating concept can be divided into the transition phase and the steady phase. Functionality in the steady phase is a prerequisite for a successful transition. If the planned measures do not enable stable operations, the previous transition phase is also not feasible. It makes sense to first check, whether a steady phase can exist and to model the transition phase afterwards. Therefore, the algorithm does not work along the chronology of the disruption when using a DRP.

4.1 Operational Measures in a Disruption Program

In order to check the functionality in the steady phase, the following measures have to be evaluated during their application:

- Partial cancellation: The line carries out a turnaround at a deviating turning station. This turning station is called a "DRP turning station". The rest of the route is no longer served. This means that only one section of the line and one original terminal station are served.
- Partial cancellation with replacement: The line carries out two turnarounds at two new DRP turning stations - one on each side of the disruption. Two route sections of the line are served. The disrupted course connecting the two sections is not used. Both original terminal stations are served and two additional DRP turning stations are declared. Replacement therefore means that another train serves part of the line on the other side of the disruption. Therefore, the partial cancellation of that line is located in between those two operating trains.
- Total cancellation: The selected train run of the line or the whole line is cancelled completely and the vehicle is either parked or stops at a platform.
- Diversion: The route of the selected train run of the line or the whole line is directed through a section of the scheduled route and/or on a completely different route. The two original terminal stations remain intact. This can be done under simplified conditions if the rules according to Ril 408.1431 (DB Netz AG 2011) are complied.
- Diversion with replacement: The route of the selected train run of the line or the whole line is run on a section of the scheduled route and/or on a completely different

route. In addition, one turnaround at a terminal station is carried out at a different station. The remaining section of the route is not operated. This means that only one original terminal station is served and the other one is replaced with a different terminal station.

These measures have various effects: on the one hand, they can reduce the number of trains in the network or influence the characteristics of a train journey (running length, route, turning station, journey time, etc.). Both can lead to a relief of the infrastructure usage and thus make it possible to reduce and avoid the occurrence and transmission of delays.

First of all, a measure needs to be feasible independently to its application being checked in module 1. However, that is not sufficient. The functionality of all measures together within the surrounding operational situation in the network is important. Thus, each measure has to be examined when scheduled, whether there are restrictions that make it unusable for the typical planned application. However, after the scheduling of all measures, the whole network is examined because the feasibility of each measure individually cannot be equated with the functionality of the operating concept. This is mainly a search for arising delays due to overload or other reasons that lead to deviations from the timetable. Effects of the changes in operations can be detected on the routes or at nodes. Furthermore, depending on the usage of a route or node, occupancy conflicts possibly arising on routes or in nodes are not always part of the consideration that can be made as a TOC.

On the routes, the measures generate changes in the minimum headways, the occupancy times, the waiting times and therefore have an influence on its occupancy level. This means, that those routes have to be checked towards the changes in the occupancy that are caused by the operating concept. Nodal changes occur in stations that show additional usage. Additional time requirements for turnarounds and changed arrival and departure times influence the occupancy rate of these stations. Thus, module 2 examines the routes and module 3 examines the stations, for each that has a different use due to the operating concept. Lines and stations, which are exclusively relieved by the measures do not have to be checked.

If an operating concept has successfully completed modules 1 - 3, it is functional in the steady phase, and can now be examined for its ability to achieve this stable state. The ability to transition is given, if the network is able to have all train runs on the lines planned for in the disruption program, in the planned number and with the punctuality of regular operations. This condition indicates a steady state. It is now to be determined, whether any exclusion criteria exist, that would prevent a successful transition. However, some exclusions for a transition cannot be determined individually, but result from interdependencies and thus generate a constant development of delays, which would inhibit a transition or extend its duration extremely. Therefore, the transition phase and its duration must be modelled in a further module. Thus, module 4 checks for exclusion criteria such as the non-achievement of the planned number of trains (in an appropriate duration) and constant congestion and module 5 determines the transition quality on the basis of duration. After applying modules 4 and 5, the operating concept created and tested is considered functional for the corresponding disruption.

4.2 Evaluation of the Steady Phase

A stable disruption operation is characterized by the fact that all trains are on the routes according to the disruption program, in the planned number and with the punctuality of regular operations. Neither the recommended limit values for capacity utilization nor the regular capacity utilization (of the regular routes and stations) is exceeded.

In a steady disruption state, dispatching actions are predictable and information (internal and external) is reliable. In the interest of the passengers, as much traffic as possible should still be maintained unchanged and as few deviations from regular operation as possible are to be planned. However, if too much regular operation is preserved, a DRP cannot function stable. It is therefore necessary to typically check, whether a DRP can be functional for all measures of the pre-planned disruption.

4.3 Feasibility Check and Calculation of the Modified Train Runs

The feasibility of each planned measure must be examined along the following three dimensions: technology, operations and transportation. A distinction is made between absolute and soft exclusion criteria for the selected train or line. Absolute exclusion criteria do not allow the measure for the typical application. Soft exclusion criteria allow an application after adjustments by the creator.

The *technical* feasibility means that the infrastructure intended for the train in question is available. Absolute exclusion criteria can be, for example, the track gauge or the need for an overhead line. Soft exclusion criteria may englobe the non-existence of switches, signals, tracks or appropriate train protection system. With soft exclusion criteria, the measure cannot be applied as planned, but can possibly be implemented using operational rules like written commands. It is suggested that no operational measures such as operating with written commands are to be used in a disruption program, as this would restrict the workflow largely. The data basis available to a TOC for setting up the infrastructure is usually not sufficient to carry out an automated check at this level of detail. Since a release test of DRPs by the RIU is mandatory, the technical test is not carried out within the framework of (partial) automation at the TOC. The creator should apply his local knowledge and use routes that are expected to be feasible.

The *operational* feasibility means that train runs are feasible on the existing infrastructure. Absolute exclusion criteria can be, for example, operational parameters such as the clearance gauge or the line category or operational regulations such as route knowledge or local guidelines. The regulations allow (approved) deviations to a limited extent in some cases. However, if these limits are exceeded, the train run is not permitted. Therefore, there are no soft exclusion criteria. For the examination of the operational measures, it is not possible to use data records, as these rules and regulations cannot yet be read automatically. In addition, there are still required skills of the driving personnel whose allocation and abilities (e.g. route knowledge) are not consistent.

The *transportation* feasibility is the accomplishment of the planned (adapted) transportation service by the TOC. In this examination, only soft exclusion criteria are initially considered, as they do not imply safety issues. The creator can define them as absolute exclusion criteria if required, e.g. depending on the requirements of the transportation association. This includes, on the one hand, the handling of a train's traffic performance: service frequency, punctuality, required train length, stops and on the other hand, the passengers' access to the platform and access to the train there (distance between platform edge and vehicle entrance).

The feasibility check can only partially be automated and the creator is required to plan reasonable measures that can be carried out according to his level of knowledge.

In this module, the (partial) automation mainly serves the recalculation of the time requirements of the trains, which change due to the application of a measure.

This includes:

- determination of delays at initial departure due to creation of a timetable message
- calculation of the duration for threading and unthreading into a diversion
- determination of the stopping time for new stops on deviations
- calculation of the minimum turning time at the selected DRP turning station
- determination of the turning buffer as a function of the selected frequency to which the turn is applied
- calculation of the actual turning time
- driving and occupancy time calculation
- determination of possible effects of a driving time extension on the return train
- preparation times for parking

After applying module 1, it is known whether the selected measure is feasible and the temporal changes in the train run become clear.

4.4 Validation of the Network's Routes

From module 2 onwards, the planned measures are no longer considered independently, but always in their entirety and in relation to the resulting effects in the network. As deduced in chapter 4.1, an evaluation of the network's routes becomes necessary if a change in use occurs while not being exclusively a reduction in capacity utilization.

The occupancy changes because of operational measures, but it must not get a height, which generates delays. The diversion measure leads to an increased occupancy of the routes that are additionally used. For each of these routes, it is therefore necessary to check whether the capacity is sufficient. Known characteristics of the route are the timetables of the existing TOC own trains and their number. The TOC is not aware of the timetables of the other services and the timetables for its own trains, which run on this route DRP exclusively, as these timetables have to be created by the RIU. For both, only assumptions can be made, which means that all timetable-dependent and exact methods are omitted. The massive change on the route results from the additional assignment of trains and their running times, which now have to be handled on it. These can be incorporated by rough information and empirical values concerning the missing characteristics, whereby the occupancy rate is suitable as a rating.

It is a rough rating method, which does not supply exact quality limit values and does not recognize occupancy conflicts. However, neither is within the scope of the possibilities of a TOC. It is important to use an evaluation method that offers a possibility to check the planned measures for the disruption, avoiding unrealistic planning and thereby limiting the revision effort for TOCs and the reconciliation effort with the RIU.

The occupancy of routes is also diverting if a single track using both directions is created out of a regular double track. For this purpose, the mean minimum headway including the number of trains must be recalculated.

The UIC recommended limit values must not be exceeded. A stable condition would not be ensured if the limit values were exceeded, and congestions could cause disruptions and delays.

However, this does not mean that a train must not arrive delayed at the terminal station. In addition to unscheduled waiting times caused by an excessively high occupancy rate, scheduled waiting times and extended driving times can also lead to a delayed arrival in disruption operations compared to the regular operations schedule.

If a route causes a delay, e.g. due to additional travel time on a deviation route, it is acceptable, if the return service can start on time. This prevents delays from continuously

1. calculation of the scheduled waiting times for threading and unthreading		
$t_{Wm} = \frac{1}{2} \cdot t_{Bnm} \cdot \frac{\rho}{(1-\rho)} \cdot (1 + V_b^2)$ (Fischer and Hertel 1990)	with	
	t_{Wm}	waiting time for threading/unthreading
	t_{Bnm}	mean operating duration
	ρ	occupancy rate
	V_b	coefficient of variation

2. estimation of the driving time expected in the event of a disruption		
$t_{journey,DRP} = t_{journey,dev} + a_{stops} \cdot t_{stop} + t_{Wm,threading} + t_{Wm,unthreading}$	with	
	$t_{journey,DRP}$	driving time when applying a DRP
	$t_{journey,dev}$	driving time on the deviation
	a_{stops}	number of stops
	t_{stop}	duration of one stop

3. departure time at the last node before leaving the standard route		
$CET_{departure,start}$	with	
	$CET_{departure,start}$	time of departure at last standard node

4. determination of the new arrival time at the destination		
$CET_{arrival,end} = CET_{departure,start} + t_{journey,DRP}$	with	
	$CET_{arrival,end}$	time of arrival at terminal station

5. determination of the feasible departure time for the return		
$CET_{mindeparture,return} = CET_{arrival,end} + t_{turn,DRP}$	with	
	$CET_{mindeparture,return}$	feasible time of departure for return
	$t_{turn,DRP}$	turning time at DRP turning station

Comparison with the actually planned departure time for the return	
$CET_{mindeparture,return} \leq CET_{plandepartue,return}$	

Figure 4: flowchart for evaluating a planned deviation

establishing and spreading uncontrollably. If the train arrives only slightly delayed, it might be compensated by the turning buffer. Therefore, the departure time of the return service must be checked for adherence using the algorithm shown in Figure 4.

To sum up, a critical influence on the route utilization is caused by diversion, diversion with replacement and a single instead of double track operation. Routes that show these measures, have to be evaluated using this module. Unchanged routes with lower or regular occupancy and routes that are cleared by module 2 are functional in a disruption.

4.5 Validation of the Network's Nodes

The transfer of the delay at the terminal station as discussed in module 2 can originate not only on the line, but also in the nodes. Since occupancy conflicts cannot be determined, not all stations and operating points with changes are considered. The additional rides and stops in between are not examined. In this module, the nodes at which the DRP measures are applied and thereby generate far-reaching effects for the station, are validated. Deviations due to measures occur at:

- DRP turning stations
- stations, where trains with total cancellation are parked on the platform
- original turning stations (not relevant as there is no additional use, reduction leads to free capacities)

Therefore, the question arises, as to how high the utilization at the application points in the DRP will be. All application points are checked also using the occupancy rate calculation.

The following validation must be carried out for all operating points that have been declared as DRP turning stations:

The first step is to determine which driving relationships are possible. Based on this, driving types can be determined based on an adaption of the method of Chu (2014). For the modelling of the infrastructure use of the station, the number of tracks i to be considered has to be determined. The possible driving relationships are determined as follows:

Find all combinations of entry from *previous station* to *station track* and exit from *station track* to *next station*
if *previous station* = *next station*, then categorize as turnaround
if *previous station* ≠ *next station*, then categorize as continuation

All trains existing in the timetable for the period under review are determined and journeys for long-distance and freight transportation are supplemented. Subsequently, the train movements f are assigned to possible driving types j within the station. A type is, for example, j_1 from A to 1 with turnaround or j_2 from A to 1 being a continuation. This means that the example pictured in Figure 5 has eight types j. For each track i, the total occupation time $t_{B,f}$ must be calculated. For this purpose, the occupation time shall be calculated for each train f for all journey types j using the track i concerned.

Occupancy time $t_{B,f}$ by train f with $j_{continuation}$ $$t_{B,f,continuation} = t_{Sp} + t_H$$

Occupancy time $t_{B,f}$ by train f with $j_{turnaround}$ $$t_{B,f,turn} = t_{Sp} - t_H + t_{turn,DRP}$$

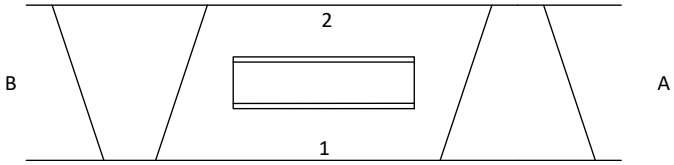


Figure 5: example station for deducing driving types j

with

t_{Sp}	blocking time
t_H	stopping time
$t_{turn,DRP}$	turnaround time at chosen DRP turning station

The total occupation time of a track $t_{B,i}$ is therefore the sum of all runs f on the track i under consideration.

$$t_{B,i} = \sum_{f=1}^{n} t_{B,f}$$

After determining the total occupancy times of the individual tracks $t_{B,i}$, the occupancy rate ρ_i is now calculated for each track i and compared with the recommended occupancy rate according to UIC (2013): check all ρ_i for the following condition: $\rho_i < \rho_{max}$.

The same calculation procedure can be applied in stations where one track is occupied by a parking train. For this, the allocation of the trains that would have used the occupied track must be transferred to other tracks. The calculation procedure can also easily depict the combination of measures at one station.

After applying module 3, it is known whether the measures and their effects are feasible at the individual application points.

4.6 Evaluation of the Transition Phase and Examination of the Transition Capability

The section "reachability" verifies the transition capability. A stable disruption condition is characterized by the fact that all trains runs are on their planned DRP lines, in DRP planned quantity and with the punctuality of regular operations. This needs to be reached within the transition phase.

When evaluating the transition phase, the first step is not to state whether the transition is feasible, but whether there are exclusion criteria that can prevent a successful transition. These are the non-achievement of the planned number of trains (in an appropriate duration) and a constant congestion of the infrastructure in the area under consideration.

Reaching the planned number of trains is achieved by the measure total cancellation being applied only in the transition phase. It leads (temporarily) to an increased occupancy of the stations at which the parking is to be carried out. For each of these stations it must be checked whether the capacity is sufficient. The validation of the measure by considering the capacity of the station tracks, is analogous to the calculation of the capacity of the DRP turning stations in chapter 4.5. However, the following adjustments must be made:

For trains to be parked, the preparation time t_{Vb} instead of the stopping time is to be used to calculate the occupancy time. When driving into the parking area after turning the blocking time is increased as the block is used twice: the signal viewing time and the travel

time in the block are set twice. It is assumed that the first route release time and the setting of the second route take place during the preparation time. In this case, there is no driving time for the approach signal distance and clearing time for the entrance.

Occupancy time $t_{B,cancel,noturn}$ due to one train parking without turnaround	$t_{B,cancel,noturn} = t_{Sp} + t_{Vb,noturn}$
Occupancy time $t_{B,cancle,turn}$ due to one train parking without turnaround	$t_{B,cancle,turn} = t_{Sp,turn} + t_{Vb,turn}$

The reduced number of trains in the system is reached before punctuality of the system can be reached. Therefore, the average duration of the cancellations should only take up a part of the desired transition time.

The following model determines the average duration of the cancellations. If the disruption interrupts a line, there may be trains to be parked on both sides of the lines being interrupted by the disruption. Both sides of must be examined.

The observation period $T_{U,AD}$ corresponds to the frequency of the line to be observed. The reference point for determining the average duration of the cancellation is the parking station. This model has to anticipate the different situations (location of the train in relation to the location of the parking station) that can be present in the disruption.

a) train drives in the direction of the railway station where the train is parked
b) train drives in the opposite direction to the holding station

If only one holding station has been declared for each side of the disruption, cases a) and b) must be taken into account. If a holding station is declared on each side, at each end of the remaining route, only case a) must be considered. If there are two stations but they are not at the end of the route, a) and b) must be taken into account.

For each trip f of the line L in question, it must be determined for every minute where it is located for the respective timetable minute m. Subsequently, it must be determined how much of the travel time on a) or b) has already been driven for the respective timetable minute m and how much of the travel time remains.

For each timetable minute m, the maximum value of all trips is selected and defined as the relevant value M_m for this timetable minute. This results in the following for the determination of the average duration of the parking of all trains.

$$t_{B,i} = \sum_{f=1}^{n} t_{B,f}$$

To test for a constant queue, the occupancy rate

$$\rho = \frac{\lambda}{\mu}$$

with

λ	arrival rate
μ	operating rate

is checked for being greater than 1 during the entire transition phase in the direct disruption influence area, as seen in Figure 6. If it is greater than 1, the DRP cannot transition because of a constant queuing and the resulting waiting times.

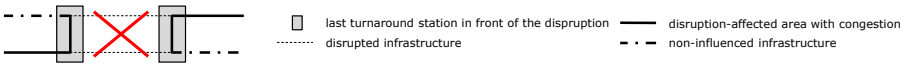


Figure 6: area under consideration in the transition phase

ρ can be reduced during the transition phase by parking trains. If the measure total cancellations affects the area under consideration, the occupancy rate can be calculated before and after completion of the parking.

After applying module 4, it is known whether there are serious exclusion criteria, which prevent the transition to be completed. However, it is not yet certain if the transition is conceivable or possible in the desired time.

4.7 Determination of the Transition Duration

A disruption program is stable if the delays in the system correspond to the delays in regular operations. Timetable conformity is assumed for both, therefore punctual means scheduled and the considered delayed trains must undergo a delay reduction down to 0 minutes.

As shown in Figure 6 the congestion and the resulting delays are mainly on the two sections of the route with occurrence exclusively in the direction of the disruption. There are no delays in the opposite direction, as there are no operational restrictions during the investigation and decision phase.

The duration is modelled in three phases: detection of disruption induced delays before applying the DRP, detection of delays arising from congestions in front of DRP turning stations and the calculation of the probable time, which is most likely needed to fully reduce the delay.

Phase 1: Detection of Disruption induced Delays before applying the DRP

The vehicles comes to a stop when the disruption occurs or at the latest when they reach the last turning point and can then only continue with the start of the DRP and the decision on how to proceed. The vehicle closest to the last possible turn before the disruption is considered at first.

If this vehicle is in the turnaround station at the time of the disruption, the resulting delay $t_{w,first}$ complies with the duration of the investigation and decision phase. Every minute that the vehicle can still drive to that station reduces the delay by one minute.

Since the time of the disruption is purely random, each line can be the foremost vehicle in the queue.

Phase 2: Detection of Delays arising from Congestions in front of DRP Turning Stations

After the DRP has started, all other following vehicles must first wait to be operated in the turning station and can only move up one after the other. For the following vehicles, the evaluation takes place on the basis of the waiting times, which are determined based on the queue. The examination of the vehicles waiting in the queue begins with the expiry of the planned arrival distance $t_{An,plan}$ after the foremost vehicle.

The queue length L_W is determined for every minute. The following vehicles are included in the calculation:

- vehicles operating in the stable DRP on this route and
- vehicles operating on this route during transition before they are parked or deviated

This queue results in a waiting time $t_{W,rear}$ for each observation minute r, which occurs before entering the turning station. It should be noted that the input values can change, as there can be lines running on the route which are cancelled or deviated during transition. The mean arrival distance t_{Anm} and the mean operating time t_{Bnm} will change then.

The development of the initial delay in the queue is now being investigated. For this purpose, each delay caused by the waiting queue is to be calculated for each observation minute r and each line L. At the DRP turnaround stations, any delays can be reduced or nulled. For further consideration, the delay with which the vehicle leaves the turning station must be determined.

For this the following two rules apply, with which the delay after the turn can be determined.

for $t_{turn,DRP,tstation} - t_{W,E,tstation} < t_{minturnRil,tstation}$

applies $t_{turn,DRP,is,tstation} = t_{minturnRil,tstation}$

then $t_{W,A,tstation} = t_{W,E,tstation} - t_{turn,DRP,tstation} + t_{minturn,Ril,tstation}$

for $t_{turn,DRP,tstation} - t_{W,E,tstation} \geq t_{minturnRil,tstation}$

applies $t_{turn,DRP,is,tstation} = t_{turn,DRP,tstation} - t_{W,E,tstation}$

then $t_{W,A,tstation} = 0$

with

$t_{turn,DRP,tstation}$	planned turning time at the DRP turning station
$t_{W,E,tstation}$	delay when entering the DRP turning station
$t_{minturnRil,tstation}$	minimum turning-time needed at the DRP turning station
$t_{turn,DRP,is,tstation}$	realized turning time at the DRP turning station
$t_{W,A,tstation}$	delay when leaving the DRP turning station

After the turnaround at the DRP turning station, the line goes back to the other terminal station. At the turnaround there, delays may also be reduced or nulled. After the turnaround the train drives back in the direction of the DRP turnaround station under consideration. It has to be determined at which observation minute the train will be at the DRP turning station again.

The system then checks whether there is still a queue at that time. If this is the case, the waiting time caused by the new queue is added to the previous delay. This results in a new delay with entry into the DRP turning station. The previous steps are then repeated until no further delay is caused by a new queue.

Phase 3: Calculation of the probable Time, which is most likely needed to fully reduce the Delay

If there is no additional delay created by a queue, phase 3 follows with the reduction of the delays by turning buffers at the turnarounds until there are none left. The transition time results from the duration of the delay reduction plus the duration of the DRP in which a delay development occurred.

The average duration of the transition $\overline{t_{ED,y}}$ for a vehicle y depends on which line L represents the foremost vehicle in the queue. Thus, all cases of the foremost vehicle are to be mapped and calculated.

$$\overline{t_{ED,y}} = \frac{t_{ED,1} + t_{ED,2} + t_{ED,3} + \cdots + t_{ED,r}}{n_r}$$

with

$\overline{t_{ED,y}}$	mean transition duration for vehicle y
$t_{ED,r}$	transition duration at the minute r under consideration
n_r	number of minutes r under consideration

The number of cases corresponds to the number of lines L that are part of the transition phase. For each case (different vehicles being the foremost in the queue), the maximum transition duration $t_{ED,V}$ of each vehicle and line must now be determined. They are averaged over their probability of occurrence P according to the corresponding line frequency. This results in the average duration of transition on the considered side of the disruption.

All t_{ED} are to be compared with each other and the largest mean transition duration $t_{ED,max}$ must be used.

If a reduction of the delays to zero minutes can be achieved, a DRP is capable of transitioning. Whether it is suitable for practical use, however, depends decisively on its duration. It is recommended to classify DRPs, which do not settle within the observation period T_U = 4 hours as not transitional, since a usage longer than this period is unlikely. However, this is not to be equated with a desired duration of transition, which should be significantly lower in order to give a large share of the DRP to the steady phase.

After applying module 5, it is not only certain whether the operating concept can transition from chaotic to stable, but it is possible to appraise its quality.

The approach enables the assessment of the operating concept based on its operational functionality and the quality of its transition phase. If the assessment is validated positively, the algorithm starts with the design and evaluation of a fitting transportation concept.

5 Transportation Concept

Based on the functional operating concept, four modules develop and evaluate passenger guidance and information measures for a corresponding *transportation concept*, also seen in Figure 3. The concept allows a customer-oriented creation and assessment of a transportation concept and therefore an indirect evaluation of the underlying operating concept.

Module 1 *Conflict Detection* searches for conflicts that imply perceptible restrictions for passengers like the non-availability of a regularly scheduled connection. Train runs that are influenced by the operating concept are determined and the resulting train relation conflicts are transferred from the operational basis into travel connection conflicts, which are perceived in the passenger's travel routine.

Module 2 searches for possible *Conflict Solution Alternatives*. Every individual conflict is provided with alternative travel connections being a feasible and acceptable solution for an individual passenger. The algorithm works with a hierarchical search behavior as seen in Figure 7. It favors the diversion of passengers in the regarded system (S-Bahn) as level 1 over the diversion of passengers in the entire public local transportation network (level 2). Releasing other trains for use like long-distance trains is the third level. Additional transportation capacities like bus distress traffic is not an option in this module but in

module 3 “Search for Solutions”, if necessary. To decide whether a level offers an acceptable solution, the connection alternatives are checked on their impact concerning feasibility of the alternative for the passenger, acceptable height of delays and transfers.

Module 3 *searches for Solutions* relating to the whole network and not only to individual conflicts. The transportation concept shall be universally valid for the typical disruption so that communication measures can be applied based on it. By allocating passenger flows, general travel connection corridors for the disruption are to be found. They are created for important connections e. g. linking both sides of the disruption. The best corridor for an important connection shows the lowest resistance increase for the related passenger flow. These optimal solutions need to be evaluated with a bottleneck analysis to check whether capacity problems at stations or in trains occur because to many corridors plan to use the same infrastructure. The overall aim is to get as many passengers to their destination as possible. A bottleneck is solved by aspiring the lowest resistance increase throughout the whole network.

Module 4 *evaluates the Passenger Guidance Concept* resulting from the conflict solution in module 3, considering the overall destination attainment of the affected passengers. The quality of the offered transportation services is reviewed from the subjective passenger's point of view. Using the method of resistance alteration modelling, the changes in passenger travel comfort, especially concerning delays and transfers are evaluated. Therefore, these characteristic values of this evaluation process are already part of the modules two and three. Every disruption that creates a conflict for the passenger leads to a resistance increase because of the necessary adaption to the situation and the deviation from the usual travel routine. The algorithm of this module identifies the changes in passenger travel comfort concerning delays and transfers by calculating the resistance alteration and evaluates the concept in context with the overall destination attainment of the affected passengers. If the transportation concept is validated positively, the DRP is completed, otherwise the operating concept needs some revisions or quality losses would have to be accepted for the transportation concept.

The algorithms give feedback on the strong weaknesses of the transportation concept like open conflicts or poorly solved conflicts. Those can be displayed as problem areas in the operating concept, so that those can be rechecked by the creator for further improvement in terms of the passengers.

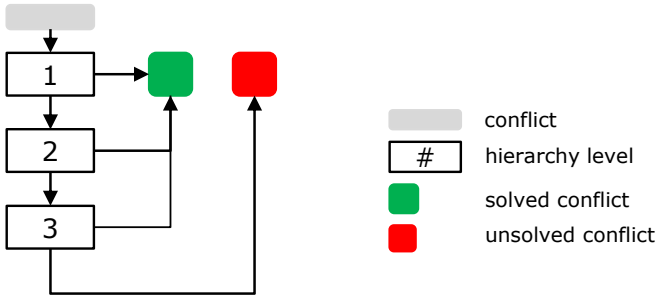


Figure 7: hierarchical search for a conflict solution alternative

6 Discussion

The presented approach can be summed up in Figure 3 and it enhances disruption programs covering both stable operations as well as passenger-friendly solutions including passenger guidance measures, which can now be reasonably designed based on a functionality checked operating concept. The algorithms enable an automated evaluation of the disruption programs in commuter railway transportation in a customer-oriented way by including and evaluating the resulting travel changes for the passengers.

The objective of the evaluation algorithms presented is to support the creation of an operationally functional and at the same time customer-oriented disruption program.

A manually created operating concept is checked for operational functionality and evaluated on the basis of the automatically calculated transition time into stable disruption operation. One the one hand, the algorithms ensure that an attainable disruption operation has been planned.

On the other hand, based on the functional operating concept, the extension of the disruption program by passenger guidance is an aim. The algorithms, which are based on a conflict search and solution approach, determine a customer-oriented transportation concept with the available travel connections and passenger routing options, taking into account passenger flows and possible infrastructural bottlenecks induced by operational measures.

Some parts of the modules principally need data sets that are currently not available for TOCs. This implies the use of sound assumptions by experts. Experienced staff is still needed for the design of DRPs but once implemented their work will be simplified, results might be of higher quality and more DRPs can be created or adapted to network discrepancies due to construction works, for example. The next step is to implement the presented approach into a software and to evaluate the algorithms with experts and test scenarios so that the approach ensures an evaluation of disruption programs based on their transition and their transportation quality for passengers. Future research on how to adapt a pre-planned DRP when in use to a deviating operating and infrastructural situation is the next step to ease disruption dispatching, that has already begun.

References

Chu, F., Fornauf, L., Wolters, A., Böhme, A., 2012. „Methode zur Erarbeitung von Störfallprogrammen“, *Deine Bahn*, 2012(07), pp. 45-48.

Chu, F., Wolters, A., Böhme, A., 2013. „Störfallprogramme betrieblich umsetzen“, *Deine Bahn*, 2013(06), pp. 20-25.

Chu, F., Oetting, A., 2013. “Modeling Capacity Consumption Considering Disruption Program Characteristics and the Transition Phase to Steady Operations During Disruptions”, *Journal of Rail Transport Planning & Management*, 3(3), pp. 54 - 67.

Chu, F., 2014. *Beurteilung von Störfallprogrammen anhand ihres Einschwingverhaltens - am Beispiel des Schienenpersonennahverkehrs*. Darmstadt, TUprints.

Crespo, A., 2018. “Dynamic Disruption Management in Commuter Railway Networks”, In: *Eisenbahntechnisches Kolloquium 2018*, Darmstadt, Germany.

Corman, F., D'Ariano, A., Hansen, I. A., Pacciarelli, D., Pranzo M., 2011. “Dispatching trains during seriously disrupted traffic situations”, In: P*roceedings of 8th IEEE International Conference on Network, Sensors and Control*, Delft.

DB Netz AG, 2011. “Züge fahren. Umleiten unter erleichterten Bedingungen.“ In: *Ril 408.1431 Abschnitt 2 Absatz 2*, Germany.

Fischer, K., Hertel, G., 1990: *„Bedienungsprozesse im Transportwesen - Grundlagen der Bedienungstheorie“*, Berlin, Germany.

Josyula, S., Törnquist Krasemann, J., 2017. „Passenger-oriented Railway Traffic Rescheduling: A Review of Alternative Strategies utilizing Passenger Flow Data”, In: *Proceedings of RailLille - the 7th International Conference on Railway Operations Modelling and Analysis*, Lille, France.

Ochiai, H., Nakamura, T.,, Kitai, M., 2016. ““Traffic Operator Assistance System” aiming at the Next Generation of Integrated Traffic Management System”, In: 1*1th World Congress on Railway Research (WCRR 2016)*, Milan, Italy.

Oetting, A., Chu, F., 2013. “Disruption Programs in Passenger Rail Transport — Ensuring Steady Operations During Disruptions”, In: *13th World Congress on Transportation Research (WCTR 2013)*, Rio de Janeiro, Brazil.

Schütz, I., Stelzer, A., 2015. “Field Evaluation of a New Railway Dispatching Software”, In: *Proceedings of ACHI 2015, The Eighth International Conference on Advances in Computer-Human Interactions*, pp. 63–68.

Stelzer, A., 2016. *„Automatisierte Konfliktbewertung und -lösung für die Anschlussdisposition im (Schienen-)Personenverkehr.“* Darmstadt, TU prints.

Toletti, A., 2018. *“Automated railway traffic rescheduling and customer information”*, Zürich, IVT Schriftenreihe, 179.

Törnquist Krasemann, J., 2012. “Design of an effective algorithm for fast response to the re-scheduling of railway traffic during disturbances”, *Transportation Research Part C*, 2012, 20, pp. 62-78.

The Disruption at Rastatt and its Effects on the Swiss Railway System

Beda Büchel[a], Timothy Partl[a], Francesco Corman [a,1]
[a] Institute for Transport Planning and Systems, ETH Zurich,
Stefano-Franscini-Platz 5, 8093 Zurich, Switzerland
[1] E-mail: francesco.corman@ivt.baug.ethz.ch, Phone: +41 44 633 33 50

Abstract
A railway track near Rastatt, Germany, lowered on 12 August 2017 and caused a complete blockage of a sector of a major rail corridor, which lasted until 1 October 2017. This track closure had severe effects on the railway freight and passenger transport. This work investigates the effects on the Swiss railroad network, using openly available realized operation data. The behavior of the delays before, during and after the disruption is investigated on three different levels. First, the delay of arriving trains to Basel SBB, as it can be seen as the input delay into the Swiss railway system. Secondly, it is investigated how the delay evolves on the Swiss intercity and interregional lines in short distance (i.e. first stop) and thirdly how this delay evolves over the course of the lines. The results display a consistent improvement of punctuality during the disruption period, which however decreases when considering stations farther away from Basel SBB. This can be explained by the fact that during the disruption period, trains arriving from Germany at Basel SBB exhibit, due to the shorter running distance, significantly lower delays than during other periods. The improved punctuality is therefore a result of a reduced delay propagation of the trains arriving from Germany. The effects of this severe and long lasting disruption can be quantified even in some spatial and temporal distance. It can be used as an example to test theoretical delay forecasting models, or examine train network complexity and interconnectivity.

Keywords
Rastatt disruption, delay analysis, open data, delay patterns

Type of Submission
Type A: Research paper.

1 Introduction

The goal of this work is to investigate potential effects of a major and long lasting disruption on a railway system, and as such to assess the interconnectivity and phenomena of delay propagation in the system under this exceptional circumstance. The disruption studied is the Rastatt disruption, taking place during one and a half month on an important freight and passenger transport bottleneck.

Railways have the challenge to operate under high capacity to achieve good economic performance, and on the other hand to reduce their vulnerability to unplanned situations. While relatively frequent, small delays can be countered by robust timetabling or traffic management systems. Large disruptions on the other side are very rare, and their impact can be much larger. Thus, they underline the vulnerability degree of networks. In this work, though, we do not study the effects of the disruption per se, but its secondary effects, in the sense that we are able to estimate at system level effects of changes to the circulation and

their effect on delay propagation phenomena.

The Swiss railway network is known for its high punctuality, reliability and robustness, and not least the capability to successfully cope with many extreme weather conditions. The actual delay of trains in the Swiss networks is typically one of the lowest worldwide, despite a very high occupation of resources. Figure 1 reports a global ranking of many countries in terms of punctuality and train km per track km, i.e. density of services on the network. This includes also the neighboring countries of Switzerland, namely Austria, Italy, and Germany. We put focus on delays originating in the latter.

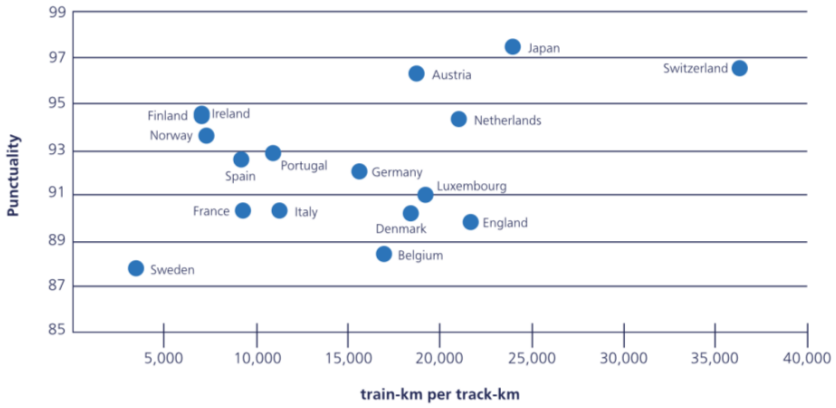


Figure 1 Punctuality (percentage of trains within 5 minutes delay) in relation to network load (train km per track km) (NS, 2017)

We perform a detailed descriptive analytics on train runs in Switzerland, where we focus on delays as key variables to quantify these effects, since they are easily accessible. The investigations focus on daily, weekly and yearly patterns of delays during the pre-/post-disruption period, and during the interrupted period.

Differently from most works in the literature, we aim to study the effects of a disruption onto a network by looking at operational data. We focus on data that span multiple months before and after the disruption, to take into account for seasonality. We are unable to identify all processes related at this point, but we can identify macroscopic effects of punctuality change at network level, as function of the distance from the disrupted area. We find that this change is in good agreement with accepted delay propagation theories.

The rest of this paper is structured as follows. Chapter 2 presents the disruption considered. Chapter 3 reports on literature on delay propagation and delay analysis. Chapter 4 reports on our methodology to study the network wide influence of the disruption. Chapter 5 reports on the analysis and the results. The last two chapter 6 and 7 respectively discuss the main findings, and conclude the paper.

2 The 2017 Disruption at Rastatt

On 12 August 2017, a track settlement occurred between Baden-Baden and Rastatt (Germany) due to the construction of a new tunnel. The affected section is part of the Rhine-Alpine Corridor (or Rhine Valley corridor) stretching from the north sea ports (Rotterdam) to Italy (Genoa), two of the most important harbors in Europe. The Deutsche Bahn (DB)

had to take out of service a track section of around 20 kilometers. The normal operations resumed on 2 October 2017.

Such a disruption has been one of the economically most relevant disruption in the last years. Most severely affected was the freight transport. The quantifiable costs in industrial and manufacturing terms amount to more than 2 billions euro. From those, about 12 million euros per week are related to freight companies' losses, according to the European Railways Network (ERFA, 2018; BLS cargo, 2018).

Diversions had to be put in place for the 200 freight trains, from different operators, that travel every day on the Rhine valley corridor. A DB Cargo usually schedules about 80 trains on the Rhine valley corridor every day. In order that as many trains as possible could use the diversion route, DB Cargo deployed additional diesel and electric locomotives and 70 train drivers; Special agreements were put in place to allow operations from vehicles and drivers, which would not normally be involved in the freight transport on the Rhine corridor (Deutsche Bahn Group, 2017).

Overall, it has been estimated that most freight trains were able to run, via a set of very complex diversion routes, as the most direct ones were affect by maintenance works, or with different power systems. The monitoring of freight at the alpine crossing (Gotthard and Lötschberg tunnels) estimates about 1500 trains being cancelled, and 400 being rerouted. Other statistics would suggest that two thirds of the expected volume of freight traffic was actually running on the alpine crossing (UVEK, 2018). The precise estimation is of course difficult as freight trains took diversions; and some freight trains were not directed towards the other side of the Alps (HTC, 2018).

Figure 2 The Rhine valley corridor between the north sea ports and Italy (Source: https://www.corridor-rhine-alpine.eu/downloads.html, adapted)

Also the passenger transport was affected. In fact, it is very difficult if not impossible to quantify now the compensation costs for passengers, modal shift, time lost for passengers, and extra money required for extra work to restore services as soon as possible. Some trains were unable to reach their maintenance workshop, which was on the other side of the disruption (Deutsche Bahn Group, 2017). Passengers travelling from Switzerland to Germany faced travel time increases from one to two hours, possible transfers, and extra bus services. To organize the replacement services for about 30.00 passengers, 450 shuttle buses runs have been organized per day, across the main stations (Deutsch Bahn Inside, 2017).

3 Related Scientific Work

The study of disruptions in transport networks and their vulnerability has been described by many researchers in the last years. Most studies refer to real-life situations, but are able to perform quantitative studies only under hypothetical situations, which are simulated in a calibrated environment, where some of the variables can be controlled.

Key concepts are and the reaction in terms of resilience, reliability, robustness, friability. While there is not complete agreement on those terms, the most common interpretations, which are also considered in this paper, are as follow (see Corman et al, 2018; Janic, 2015; Jenelius, 2007). Robustness is considered the ability of a transport system to perform its functions when it is under perturbed conditions. Reliability is related to a transport service, which deviates in a limited manner from a prescribed time plan. Resilience is the ability to recover to a normal state after having been disturbed, i.e. to neutralize the impacts of disruptive events, after their occurrence. Friability is related a reduction on a network resilience due to removing particular nodes or links, and consequently cancelling some services. Vulnerability might be related to the susceptibility to extreme strains on a dynamic system (Reggiani et al, 2015).

Resilience has been studied for road networks and more recently for public transport networks, based on simulated conditions, and with a direct filter by the demand, i.e. the users of the system, which are exposed to a different abnormal situation (see for instance, Malandri et al, 2018). When dealing with case-studies, real (quantitative and qualitative) data are used, and the focus is on understanding probability and impact of shocks and sudden change in states. Differently from all studies reviewed in (Reggiani et al, 2015) and (Mattson and Jenelius, 2015), we focus on railway system, and on the analysis of a real-life situation.

The connectivity of the network either in a purely topological sense, or in a more service oriented manner, has been identified often to play a major role. Different connectivity structures would enable different exposure and impacts of the same disruption (such an approach has been for instance studied in Malandri et al, 2018). A disruption in a heavily connected part of the network has larger impacts than a disruption in a less-connected part of the network. Moreover, a connected network enables a higher resilience to disruptions, i.e. mitigating its exposure or impact, if focusing on users or operations, respectively.

Different connectivity in network structure would also put different strain on the network components and may lead to different disruption probability and or resilience. This concept is related to friability, as a change in resilience after removing some of the network links/nodes. In fact, network with different levels of interconnections can exhibit different dynamics when exposed to abnormal conditions (see for instance Corman and D'Ariano, 2012). In the light of the friability analysis performed in (Janic, 2015), railway systems face similar corrective actions under disruptions, namely cancelling and rerouting services. Similarly to the analysis in (Janic, 2015), we aim to quantify how the resilience (or more

properly, the performance) of the network reached a higher value in some parts of the network, during the disruption thanks to the mitigating actions implemented. From a different perspective, our analysis aims to understand the "unknown connectivity" (Reggiani et al, 2015) which is underlying a large scale railway network. In particular, we target to identify some implications on network reliability based on a disruption, which has no direct physical connections by means of services or links, but only to indirect effects of service quality.

Railway systems are typically built with very small reserve capacity, and the effect of traffic to service quality is relatively strong (see Figure 1 above). In fact, delays and irregularities in operations, which can be related to disruption or many other events, propagate in heavily used networks as knock on delays. Many works studied formulas to either simulate delay propagation from a deterministic perspective or from a stochastic perspective, or to recognize delay propagation phenomena in operations (see for instance Goverde, 2010). The study of network operations, pertaining stability, reliability and robustness, and also their interconnection is typically addressed from a theoretical point of view, simplified networks, ideal conditions, or simulation studies. Instead, we refer to only operational data, where multiple factors have been recorded and aggregated and the precise root cause of all phenomena cannot be clearly separated. Delay propagation or delay prediction approaches using those models are often used in small perturbations, while little evidence is used that similar prediction models can perform good in presence of very serious changes to operations, such as disruptions (Corman and Kecman, 2018).

How to react to a disruption typically involves a series of actions, like cancelling trains, rerouting them at a global network scale, or introducing additional stops or turnaround points, depending on the severity of the disruption and the expected length (Ghaemi, 2018). The effect of short turning, and shuttling during disruptions has been also investigated in Corman and D'Ariano, (2012), who evaluated it from a large amount of possible performance indicators. Nevertheless, both of those approaches refer to academic situations, and not recorded operations. In those cases, simulation, optimization models and what-if scenarios might deliver useful data, as far as they are fed with correct data. This is typically a challenge in the frenetic aftermath of disruptions and during the strong efforts to bring situation to normality.

Summarizing, with regards to the literature, we focus on realized operations during a real-life disruption, which include a large amount of uncontrollable and unmeasurable phenomena; we tackle railway networks of particularly limited available capacity; we study the impact of disruption and mitigating actions over a large network, where mostly indirect effects of delay propagation can be seen.

4 Data and Methodology

4.1 Data

An effort that started some years ago is the publishing of open data about realized operations. This has been established since a few years in many countries, including Norway, Netherlands, UK, and also Switzerland. In particular, the Swiss Federal Railways (SBB) publishes actual arrival and departure data of train, bus, tram and boat rides in Switzerland since December 2016 on their Open Data Platform. This paper bases on the timetable years 2017 and 2018, which start and end in mid-December respectively. The recorded and published data in this time window result in a size of 120 GB, which were used as raw material for this investigation. For graphical representations, the data of the full available period was used. For statistical analysis, however, we focus on the timetable year 2017, starting on 11 December 2016 and ending on 12 December 2017, to avoid any

systematic effect.

4.2 Methodology

We consider delays of the higher product level of train service. Therefore, we consider the international passenger trains running through Basel SBB, namely EuroCity (EC), Intercity-Express (ICE), and TGV. Furthermore, we consider national services namely InterCity (IC) service, which connect major cities within Switzerland, and InterRegio (IR) service, which connect regions within Switzerland and typically stop in cities and mid-size towns only.

To be able to do reasonable evaluations of change of delays, delays have to be aggregated. We aggregate delays in three spatial levels. Firstly, since Basel SBB is the first major entrance station in Switzerland for services of the Rhine Valley Railway, we investigate the arrival delays of trains coming from Germany to Basel SBB. This delay is considered as the initial delay in the Swiss Railway Network. In a second step, stations with direct (non-stop) connection to Basel SBB are considered. Finally, we also investigate the delays at all stops of direct lines running from Basel SBB. Always arrival delays of trains from Basel SBB are considered.

Stations with a direct connections to Basel SBB are Liestal, Rheinfelden, Olten and Zürich HB. For the analysis, delays in Liestal and Rheinfelden as well delays in Olten and Zürich HB are considered together (see Table 1). This, as Liestal and Rheinfelden are quite close to Basel and are subordinate stations, where also IR trains offer direct connections, whereas Olten and Zürich HB are further away and are superordinate stations of the Swiss railway network.

Table 1: Considered Stations with direct connections from Basel SBB

	superordinate stations		subordinate stations	
Stations	Zürich HB	Olten	Rheinfelden	Liestal
Direct connecting services	TGV, ICE, IC	TGV, EC, IC	IR	ICE, EC, IC, IR
Travel time from Basel SBB	53 min	24 min	12 min	9 min

To have a comparison, we also investigate delays at stations, which are most likely not or only very limited affected by the Rastatt disruption. The chosen stations are located in the south western part of Switzerland and have direct (non-stop) connections from Lausanne, and are not connected by a service to Basel SBB (see Table 2). Also considering these stations, we can distinguish between superordinate stations (Yverdon-les-Bains, and Fribourg / Freiburg) and subordinate stations (Morges, Palézieux, and Vevey). A geographical depiction of lines and considered station is reported in Figure 3.

Table 2: Considered Stations with direct connections from Lausanne

	superordinate stations		subordinate stations		
Stations	Yverdon-les-Bains	Fribourg / Freiburg	Morges	Palézieux	Vevey
Direct connecting services	IC	IC	IR	IR	IR
Travel time from Lausanne	24 min	43 min	10 min	15 min	13 min

We not only aggregate the delays on a spatial level, but also on a temporal level. Therefore, we consider percentile delays on a daily base. When considering e.g. the 20th percentile, the percentile value is the delay value below which 20% of the daily delay observations may be found.

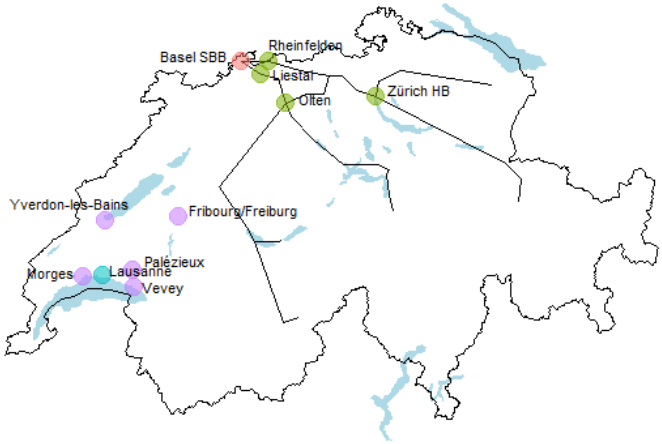


Figure 3: Map of the investigated stations. The black railway lines are routes that IC and IR trains take from Basel SBB (red). These lines stop first in Rheinfelden, Liestal, Olten or Zürich HB (green). The comparison stations in southwestern Switzerland (Yverdon-les-Bains, Fribourg / Freiburg, Morges, Palézieux, and Vevey (violet) are first stops from Lausanne (blue).

The characteristics of the time series of arrival delays were analysed during preliminary tests. The time series are highly variable; there is no clear trend nor seasonality. However, the series show significant auto correlation for lags 1, and 7. This means that the delay at a given day and at the next day, as well as the delay of the same day a week later, are correlated. There is no distinct weekly pattern throughout the year, as can be seen in Figure 4, where exemplary the median daily arrival delay to Basel SBB is shown split into days and weeks. Therefore, we do not investigate the weekdays separately.

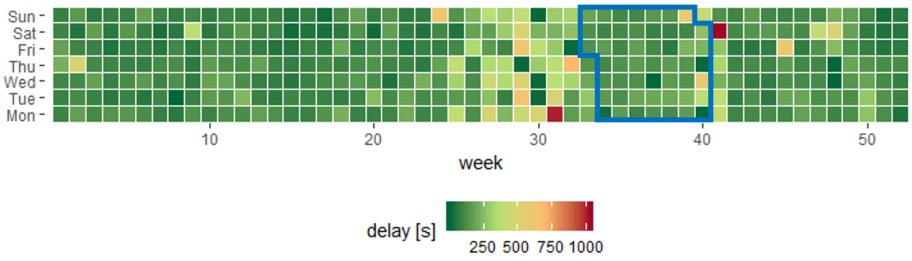


Figure 4 Median daily arrival delay to Basel SBB over the course of the year 2017. During the blue marked period, the disruption took place.

Given the highly variable daily pattern of delays, it is difficult to prove a distinct difference in the time series of daily delays during the disruption period in comparison with daily delays of the period before and after the disruption. Furthermore, we don't have a baseline time series of the disrupted period for comparison. Therefore, a difference cannot be proven but only indicated. To do so we perform three different tests.

First, we define an indicator that addresses the fact that the disruption might lead to a remarkable change in a short term to the time series (i.e. a shock). Therefore, we calculate the difference of the mean value of the percentile delays of the seven days before the disruption and the respective value of the first seven days in the disrupted period (d_b). Analogously, we calculate the difference of the mean value of the percentile delays of the seven days before the end of the disrupted period with the mean value of the seven days after the disruption (d_e). In Figure 5 these differences are explained visually. For comparing those differences in delay, we compute the difference of the means of any seven consecutive days (d_i). We then build the indicators I_1 and I_2 for the differences, given by

$$I_1 = P\left[\min(|d_b|, |d_e|) > \min(|d_{i,1}|, |d_{i,2}|)\right], \quad (1)$$

$$I_2 = P\left[(|d_b| + |d_e|) > (|d_{i,1}| + |d_{i,2}|)\right]. \quad (2)$$

Where $d_{i,1}$ and $d_{i,2}$ are randomly chosen samples of weekly mean changes. I_1 compares the minimal difference in delay at the begin or at the end of the disruption with the minimal difference of two random dates. I_2 considers, in opposition to I_1, the sum of the two differences. The indicators can take values between 0 and 1. The higher the value is, the more infrequently such a distinct change happens. In reverse, this means that the dates of the disruption are more special compared with two random dates.

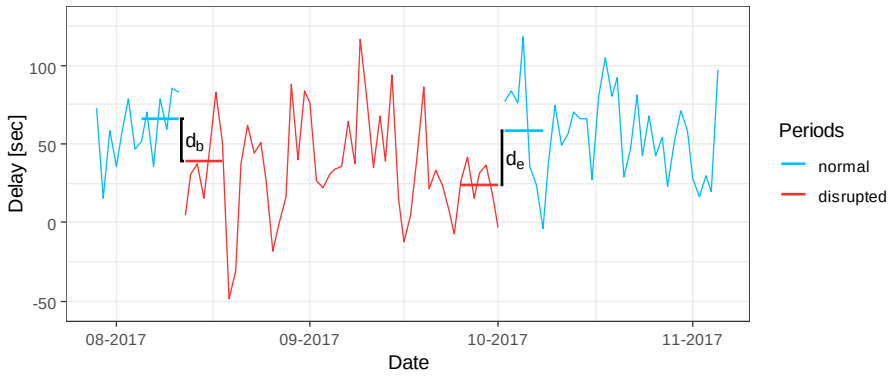


Figure 5 Visual explanation of the calculation of the differences of the weekly averages

In a second test, we compare the delay distribution during the disruption period with the delay distribution of the preceding and subsequent thirty days as Figure 6 shows. Note that by doing so, the dependencies of delays of consecutive days is removed. We assume that the delay distribution during 30 days before and after the disruption is a good proxy for the hypothetical delay distribution during the disruption period. Therefore, these two

distributions are compared by the aid of a two-sample Kolmogorov-Smirnov (KS) test. By this test, we check if the two samples are probable to come from the same distribution. The test is performed under the Null Hypothesis H_0 that the data comes from the same distribution. H_0 is rejected if the p-value is lower than the level of significance $\alpha = 0.01$.

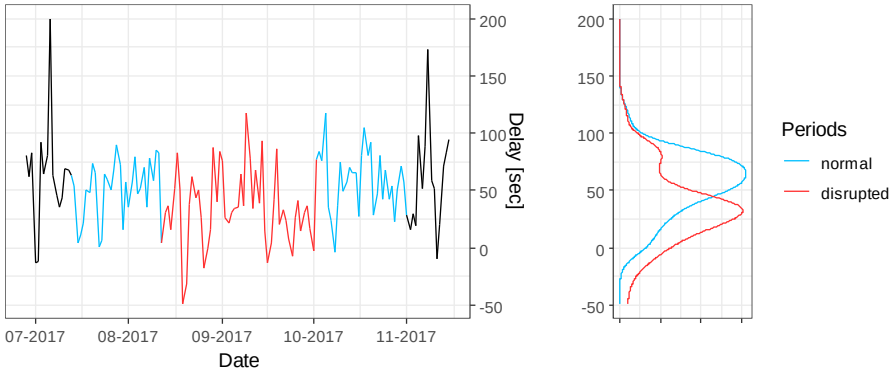


Figure 6 Visual explanation of the two compared distributions

Third, to take into account the temporal dependencies of the time series, we determine the best fitting ARIMA model, based on the AIC, for the data of 2017, while excluding the disruption period. The mean value of the baseline time series during the disruption period, is estimated by Kalman smoothing on the state space representation of the ARIMA model. This is reported to be a powerful method for filling gaps in time series (Moritz et al., 2015). In the following, we compare the baseline time series in the disruption period with the real measured values. We conduct a t-test under the assumption of equal variances on a significance level $\alpha = 0.01$. If the p-value is smaller than the significance level, we reject the H_0, which proposes that there is no difference between the mean of the baseline time series and the observed time series.

5 Results

5.1 Arriving Trains from Germany to Basel SBB

The daily median delay of trains arriving to Basel SBB is shown in Figure 7. The red line shows the delays during the disruption period, the blue lines show delays when the disruption was not present. The change in timetable years, which is in December, is indicated by a slight change of the blue color. Furthermore, a black line, representing the simple moving average with a period of 7 days (average over the course of a week) is introduced.

The pattern is quite distinct; the highest delays of this observation period of two years are reached just before and after the disruption, presumably due to the construction works in southern Germany. Then, during the disruption, when extra trains where running, the delay dropped remarkably.

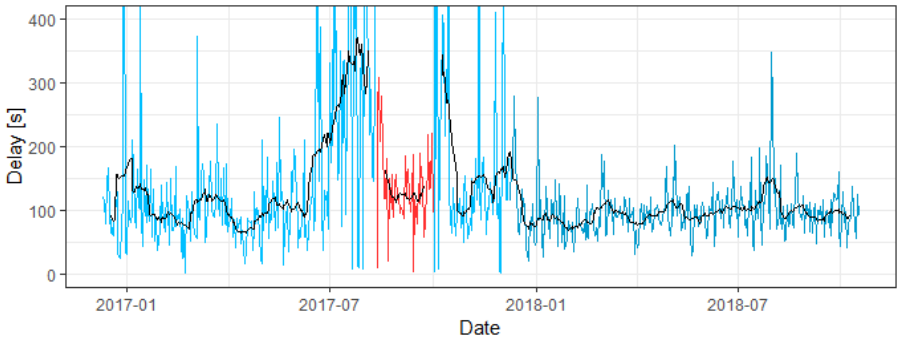


Figure 7 Daily median delays of trains arriving from Germany to Basel SBB

Figure 8 shows the time series of further percentile values of the weekly moving average of the daily delay distributions. It shows the 20th, 40th, 60th, and 80th percentile values of those values over the course of the timetable years 2017 and 2018. The blue curves show moving averages of delays at dates, which were not influenced by the disruption, the red curves show moving averages of disrupted dates. The grey values represent the transition phase, or in other words, they are computed with delays of dates that were affected by the disruption and such that were not.
The different percentile values show a similar course. All percentile time series have distinctly higher delays before and after the disruption period, than during the period. Furthermore, the variation of the delays is remarkably smaller during the disruption period.

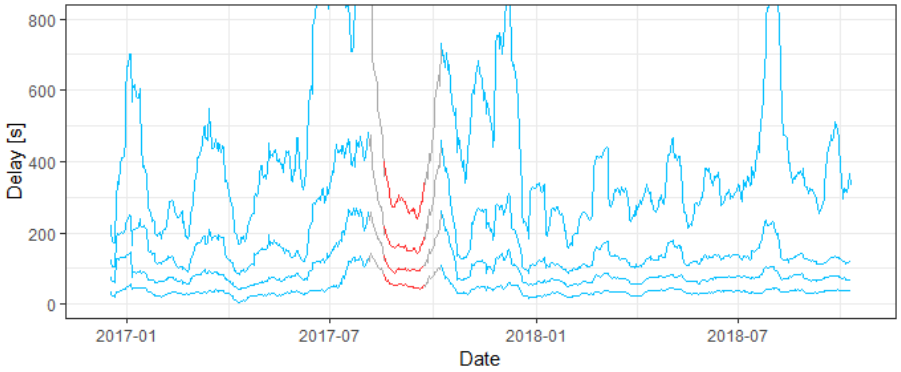


Figure 8 Moving average (period of 7 days) of the percentile values (p = 20, 40, 60, 80) of arrival delays at Basel SBB

This change can be underlined in a statistical way. In Table 3 the indicators I_1 and I_2, as well as the result of the KS-test and the t-test for the 20th, 40th, 50th, 60th, and 80th delay percentiles are shown. The I_1- and I_2-values are rather high (often 0.8 and more). This indicates that the change during this period is remarkable and comparatively high. Also, the KS-test and the t-test clearly indicate a significant change in the time series. The KS-test

states, that the distribution changes while the disruption is present and the t-test indicated that the time series have different means.

Table 3: Indicators I_1 and I_2 and test results of KS- and t-tests for Basel SBB

Basel SBB				
p	I_1	I_2	p-value KS-test	p-value t-test
0.2	0.86	0.80	7.2×10^{-3}	8.9×10^{-16}
0.4	0.80	0.93	2.4×10^{-6}	1.8×10^{-21}
0.5	0.81	0.93	1.2×10^{-6}	6.5×10^{-25}
0.6	0.79	0.95	3.6×10^{-7}	7.7×10^{-25}
0.8	0.64	0.70	2.0×10^{-9}	1.3×10^{-24}

This distinct change is most likely due to the fact that trains were running along a much smaller network, i.e. short turning at Baden-Baden (1.5 hours away from Basel) instead of Hamburg (7 hours away from Basel). This caused that these trains did not arrive in Switzerland with their potentially accumulated delays as they would if there were no interruption.

5.2 Delay Pattern One Stop Away

After assessing the entrance delay at Basel, we look at the delay propagation in the Swiss Railway network. We look at the first stop of direct trains from Basel. Figures 9 – 12 show, under the same styling convention as Figure 7, moving average of the percentile values of daily delays. The investigated percentiles are 20^{th}, 40^{th}, 60^{th}, and 80^{th} percentile.

For the two groups of stations with direct train connections from Basel, namely Zürich HB and Olten, as well as Liestal and Rheinfelden a relatively clear trend of lower delays during the disruption period can be recognized. The average delays during the disruption is as low as the minimum delay recorded throughout the year. The variability is actually much smaller during the disruption period, than throughout the rest of the year.

For comparing these observations, a placebo test was conducted with the train station that have direct connections from Lausanne. Neither the farther away located major stations, as Fribourg / Freiburg and Yverdon-les-Bains, nor the nearer and less important stations Vevey, Morges, and Palézieux show a clear influence of the Rastatt disruption. These stations are far enough away from the disruption there the effects of the disruptions cannot be quantified anymore.

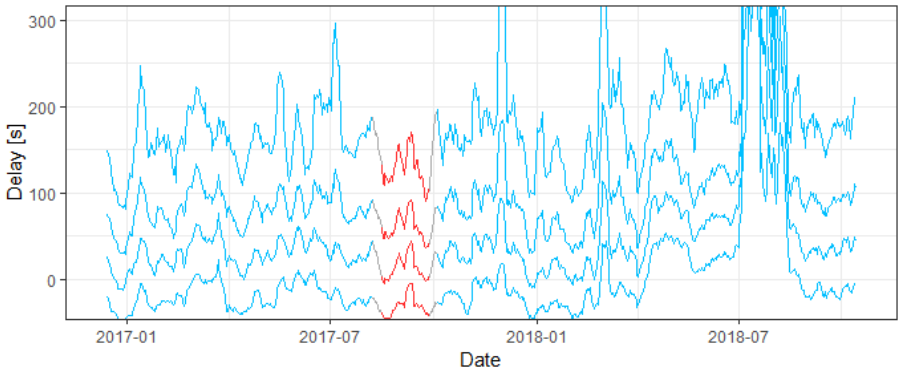


Figure 9 Moving average (period of 7 days) of the percentile values (p = 20, 40, 60, 80) of arrival delays at Zürich HB and Olten

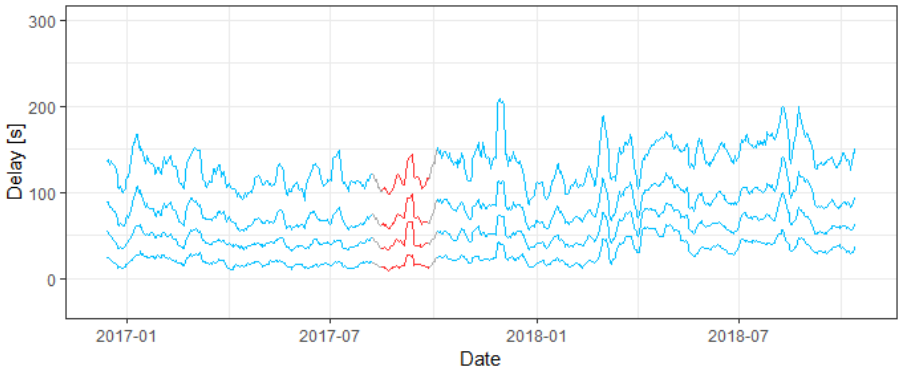


Figure 10 Moving average (period of 7 days) of the percentile values (p = 20, 40, 60, 80) of arrival delays at Liestal and Rheinfelden

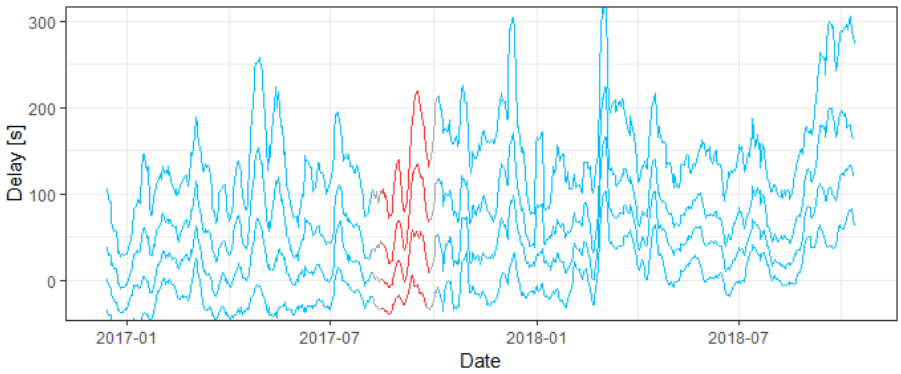


Figure 11 Moving average (period of 7 days) of the percentile values (p = 20, 40, 60, 80) of arrival delays at Yverdon-les-Bains and Fribourg / Freiburg

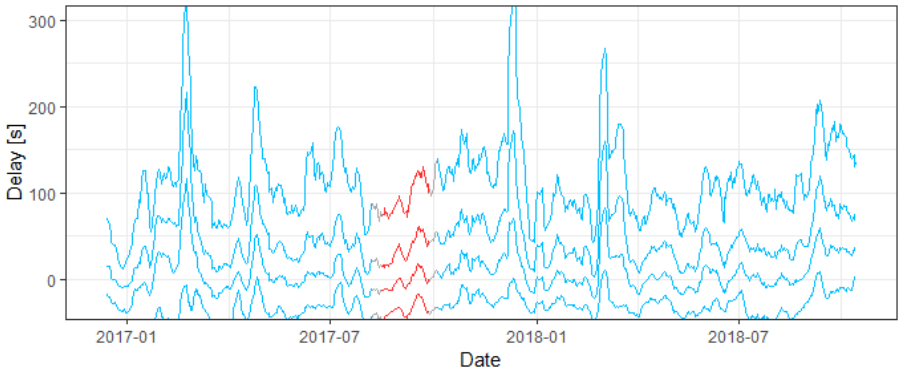


Figure 12 Moving average (period of 7 days) of the percentile values (p = 20, 40, 60, 80) of arrival delays at Vevey, Morges, and Palézieux

In Table 4 the indicators I_1 and I_2, as well as the result of the KS-test and the t-test for the 20th, 40th, 60th, and 80th delay percentile are shown for the stations with direct services to Basel SBB (left) and the stations far away (right). The values in the table are highlighted with color. A green color is an indicator for an influence of the disruption, red is an indicator that the disruption had no influence, yellow is in between.
The groups of stations close to Basel SBB exhibit clearly higher I_1- and I_2- values for all investigated percentile values compared with the stations close to Lausanne. Also considering the results of the KS-tests and t-test the difference between the groups is evident. While the groups close to Basel SBB show almost always significant differences between the disrupted and non-disrupted periods, for the groups close to Lausanne this is rarely the case.

We don't find evidence, that the stations near to Lausanne were influenced by the Rastatt disruption, whereas we find strong indication that stations near to Basel SBB felt an effect. The indicators and statistical tests are show a clear difference between the two groups of stations.

Table 4: Indicators I_1 and I_2 and test results of KS- and t-tests for groups close to Basel SBB (Liestal & Rheinfelden and Olten & Zürich HB) and for groups close to Lausanne (Yverdon & Fribourg / Freiburg and Morges, Vevey, & Palézieux)

Liestal & Rheinfelden				
p	I_1	I_2	p-value KS-test	p-value t-test
0.2	0.87	0.89	3.1×10^{-4}	8.7×10^{-4}
0.4	0.86	0.97	2.3×10^{-4}	4.7×10^{-2}
0.5	0.92	0.94	8.2×10^{-4}	1.0×10^{-2}
0.6	0.94	0.93	7.0×10^{-3}	3.4×10^{-4}
0.8	0.87	0.91	4.7×10^{-4}	1.0×10^{-5}

Yverdon & Fribourg / Freiburg				
p	I_1	I_2	p-value KS-test	p-value t-test
0.2	0.70	0.80	4.7×10^{-2}	2.8×10^{-3}
0.4	0.71	0.31	2.6×10^{-1}	3.4×10^{-2}
0.5	0.59	0.32	4.1×10^{-1}	1.5×10^{-1}
0.6	0.60	0.49	8.0×10^{-2}	3.5×10^{-1}
0.8	0.58	0.15	6.1×10^{-2}	1.9×10^{-2}

Olten & Zürich HB				
p	I_1	I_2	p-value KS-test	p-value t-test
0.2	0.71	0.90	2.3×10^{-2}	2.9E-03
0.4	0.93	0.99	1.6×10^{-3}	1.8E-07
0.5	0.88	0.88	6.8×10^{-5}	2.2E-07
0.6	0.84	0.74	7.1×10^{-4}	2.4E-07
0.8	0.88	0.90	4.3×10^{-3}	1.6E-09

Morges, Vevey, & Palézieux				
p	I_1	I_2	p-value KS-test	p-value t-test
0.2	0.18	0.45	1.8×10^{-1}	5.4E-01
0.4	0.23	0.20	7.5×10^{-2}	1.6E-01
0.5	0.35	0.37	2.0×10^{-1}	1.1E-02
0.6	0.22	0.37	8.9×10^{-2}	2.0E-07
0.8	0.33	0.33	2.6×10^{-2}	4.4E-04

5.3 All IC Lines Departing from Basel SBB

In a third step we look at all train lines from Basel. We compute for all stops of the lines the difference of the delays during the disruption and non-disruption period for different percentile values. This difference in delays is shown color coded in Figure 13. Additionally, the number of daily trains per station is shown by the size of the circle.

It is visible that near to Basel SBB the trains reduced their delay during the period. Farther away, the pattern is not so clear anymore. The line running to St. Gallen eastern Switzerland even performed worse in the disruption period compared to the rest of the year.

Furthermore, it can be seen that for high percentile values the gains and reduction respectively were more than for low percentile values, meaning particularly the strongly delayed trains performed better in the disruption period.

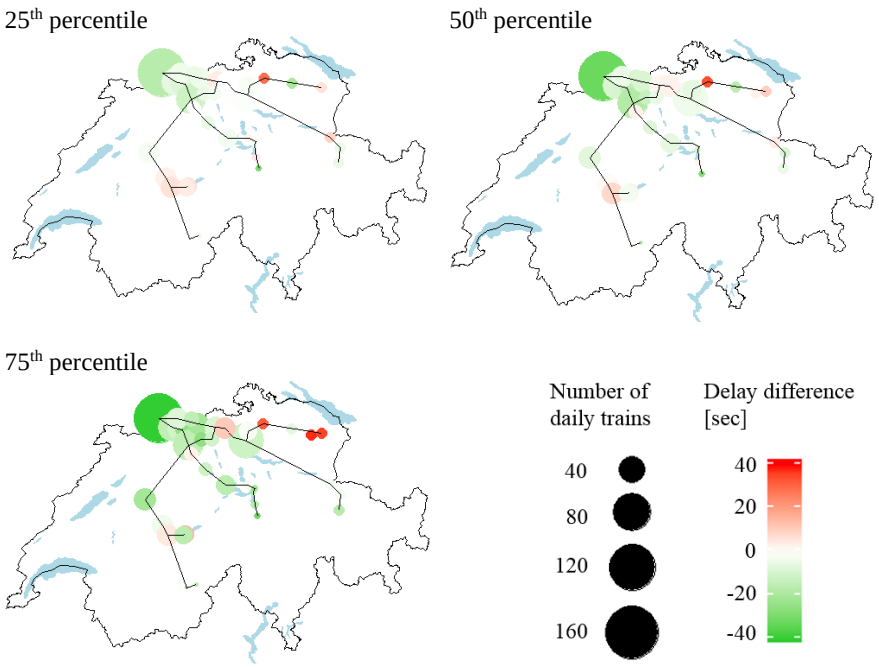


Figure 13: Evaluation of arrival delay difference given different percentile values (p = 25, p = 50, and p=75)

6 Discussion

From the analysis performed, a few points are worth being discussed. The variability of delays in real life operations is extremely high; no model of first order or second order, or with a time series analysis with one week or one day fit could explain the variance of the observed data. In fact, the realized delay is the product of so many factors, some of which are correlated to a certain extent over space, day, days, weeks (like weather; holiday seasons; maintenance actions) and some are more of a random components related to demand, some other to operational process. Further steps based on delay distributions or fitting functional relations to discover or highlight root causes in the variable performance are an interesting follow-up (see for instance Cerreto et al, 2018).

Also due to this high variability, the strength of typical statistical tests to identify the difference in samples and relate them to underlying changing in organizational pattern is quite limited. Moreover, each test can be performed at different percentile level, and maybe spurious phenomena can be pinpointed. It is difficult to clarify the philosophical dilemma between what is in the reality, what is in the data and what is in the eyes of the observer.

The study of the relation between different input conditions and the performance of the network is very crucial in reliability assessment, in economic appraisal of new projects. Most of the studies of complex network and service level based on topological structure or on service structures also do not go in detail in discussing the microscopic impact that the relation traffic-performance has to a railway network. To this end, simulated operations would need to consider an enlarged set of parameters of random processes, to result in a

performance directly comparable with the observed real life.

Another limitation or feature of the approach is the fact that every disruption is one-of-a-kind, and its impact is so large that it unavoidably changes the conditions under which the system is operating. This includes for instance running speed, planned stops, travel time, passenger demand, flow over links, and mitigation actions. The interaction of all those aspects is so intricate that is almost impossible to identify all contributions, unless a set of simulations based on some assumptions could be replicated, to isolate those components. The reasons why some mitigation actions have been chosen, what the objective was, and to which extent those mitigation actions reached their goal is something, which is very relevant for design of future contingency schemes (see for instance BLS cargo 2018). It is furthermore relevant from a process point of view, to identify bottlenecks and enable exchange of best practices, but also from a traffic planning and dispatching point of view, where there is strong need for smart decision support (see for instance Ghaemi, 2018).

7 Conclusion

From the analysis performed, the Rastatt disruption did not degrade the punctuality of the Swiss passenger trains at Basel SBB. On the contrary, it even improved it. The long train section from the Netherlands and northern Germany to southern Germany, Switzerland and Italy was split into two, what caused trains to arrive in Switzerland with lower potentially accumulated delay. A further reason for the consistently lower delays is the secondary delay, which was reduced due to much more punctual trains arriving from Germany.

Additional effects that should be investigated are the effects to passengers, in terms of additional travel time in Germany, which were related to the disruption, as the costs for the planned unreliable services in the non-disrupted situation was then felt directly by passengers as extra connection time. This analysis of disruption can be performed post-eventum only by replicating behavior of people, for instance via agent-based models, and assuming that sufficiently accurate modelling of the non-equilibrium (Malandri et al, 2018) behavior of passengers during disruptions can be replicated properly (see for instance Leng et al, 2018).

It would be very interesting to clarify the impact of freight trains, which were running in a very different pattern during the disruption, and partially cancelled or rerouted to other different parts of the network. The main limitation for this is the unavailability of sufficiently accurate data, which also includes the probabilistic chance of delay propagation by freight train under normal operating conditions, something which so far addressed large attention, but delivered few clear conclusions (Andersson et al, 2015). The possibility to fit stochastic models to the two situations, and derive parameters linking traffic, buffer time and observed delay propagation can open up a field of operational analysis of networks and their vulnerability.

References

Andersson, E. (2014). An Economic Evaluation of the Swedish Prioritisation Rule for Conflict Resolution in Train Traffic Management. In: Procedia - Social and Behavioral Sciences: Transportation. 16th Meeting of the Euro Working Group on Transportation, Porto, Portugal (pp. 634-644).

BLS cargo, (2018) The unspoken costs of rail disruptions: the consequences of Rastatt on the economy and customer confidence. Talk at general Assembly ERFA, retrieved at http://www.erfarail.eu/uploads/1-%20Stahl%20ERFA%20Meeting-1524565034.pdf

Cerreto, F, Bo F. Nielsen, O.A., Harrod, S. (2018) Application of Data Clustering to Railway Delay Pattern Recognition. Journal of Advanced Transportation, Article ID 6164534 ,pp1-18

Corman F., D'Ariano A. (2012) Assessment of advanced dispatching measures for recovering disrupted railway situations. Transportation Research Record: Journal of the Transportation Research Board.2289/2012.

Corman, F. Kecman P. (2018) Stochastic prediction of train delays in real-time using Bayesian networks. Transportation Research, Part C, 95, 599-615

Corman F, Quaglietta E., Goverde, R.M.P. (2018). Automated real-time railway traffic control: An experimental analysis of reliability, resilience and robustness. Transportation Planning and Technology 41(4), pp. 421-447doi:10.1080/03081060.2018.1453916

Deutsche Bahn Group (2017) Integrated Report. Berlin. www.deutschebahn.com

Deutsche Bahn Inside (2017) Rheintalbahn: Sperrung aufgehoben Verkehr rollt wieder. https://inside.bahn.de/tunnel-rastatt-sperrung/

European Rail Freight Association ERFA (2018) Study finds Rastatt incident to have caused losses of more than €2 billion. Retrieved at http://erfarail.eu/news/the-economic-impact-of-rastatt

HTC Hanseatic Transport Consultancy (2018) Volkswirtschaftliche Schäden aus dem Rastatt-Unterbruch-Folgenabschätzung für die schienenbasierte Supply -Chain entlang des Rhine-Alpine Corridor. Retrieved from European Rail Freight Association ERFA at http://erfarail.eu/uploads/2018_April%20Studie-1524476846.pdf

Ghaemi N.(2018) Short-turning trains during full blockages in railway disruption management, Phd Thesis. TU Delft

Goverde, R.M.P. (2010). A Delay Propagation Algorithm for Large-Scale Railway Traffic Networks. Transportation Research Part C, 18(3), 269-287.

Janić, M (2014) Modelling the resilience, friability and costs of an air transport network affected by a large-scale disruptive event. Transportation Research Part A, 71, pp 1-16

Jenelius (2007) Approaches to Road Network Vulnerability Analysis. Licentiate Thesis, KTH Stockholm.

Leng, N, De Martinis, V, Corman F (2018) Agent-based simulation approach for disruption management in rail schedule. Conference on Advanced Systems in Public Transport and TransitData (CASPT 2018), Brisbane, Australia.

Malandri, C., Fonzone, A, Cats O (2018) Recovery time and propagation effects of passenger transport disruptions. Physica A, 505, pp 7-17.

Mattson, L.G, Jenelius E (2015) Vulnerability and resilience of transport systems - A discussion of recent research, Transportation Research Part A, 81, pp 16-34.

Moritz, S., Sardá, A., Bartz-Beielstein, T., Zaefferer, M., & Stork, J. (2015). Comparison of different methods for univariate time series imputation in R.

NS Nederlandse Spoorwegen (2017) Yearly report.

Reggiani, A. Nijkamp, P, Lanzi D (2015) Transport resilience and vulnerability: The role of connectivity. Transportation Research Part A, 81 pp 4-15
UVEK Eidgenössisches Departement für Umwelt, Verkehr, Energie und Kommunikation (2018) Verlagerungsbericht 2017. Bern.

Enhancement of Blocking-time Theory to Represent Future Interlocking Architectures

Thorsten Büker [a], Thomas Graffagnino [b],
Eike Hennig [a 1], Alexander Kuckelberg [a]
[a] VIA Consulting & Development GmbH
Römerstr. 50, 52064 Aachen, Germany
[1] E-mail: e.hennig@via-con.de, Phone: +49 (241) 463 662-28
[b] Schweizerische Bundesbahnen SBB
Hilfikerstr. 3, 3000 Bern 65

Abstract
Infrastructure managers around Europe are facing two major topics within the next years: The large-scale renewal of command and control systems and the need to increase capacity. Both topics can be addressed with the further development of ETCS Level 3 and ATO in combination with the introduction of new types of interlockings.
Accompanying the further development of command and control systems there is a need to enhance the blocking-time theory by defining new time components. With this development, effects on capacity can be identified and feasible capacity gains can be evaluated. The enhanced blocking-time theory needs to be implemented into standard railway software tools, which can be quite challenging, due to a shift in paradigms compared to all current command and control systems.
Within this paper, experiences gained in previous studies regarding the necessary blocking-time theory enhancements, the implementation challenges and exemplary capacity gains are outlined. Based on this topics for further research and standardization are defined.

Keywords
Blocking-time theory, ETCS Level 3, ATO, Capacity assessment

1 Introduction

The most common means to express the capacity consumption per train movement is the blocking-time approach. It has been introduced in the 50^{ies} and has been standardised for a broader set of applicants at the beginning of the 21st century. With the emergence of ETCS, efforts have been made for an appropriate extension of the blocking-time model. Ensuring a precise representation of the capacity consumption per occupation element (usually track-clearance section) the model enables conflict-free timetabling for all types of (mixed) signalling system as well as various forms of capacity assessments and simulations studies.

Recently, railway-infrastructure managers and the supplier industries have launched major programs with the aim to revise – or even reinvent – the overall interlocking architectures plus adjacent systems and operational principles. Representatives of such initiatives are “smartrail 4.0” (Website smartrail) in Switzerland and “Digitale Schiene Deutschland” (Website DSD) in Germany. Those programs are backed-up by a set of motivations:

- Existing command and control technology is overaged or becomes outdated,
- Skills to maintain technology get lost due to demographic aging,

- Applied technology is expensive and does not allow any further capacity gain.
- System conception and architecture don't allow to make full use of actual technical capabilities

While those programs stated above imply a severe redesign of architecture in a mid-term horizon, nationwide efforts to rollout ETCS and upgrade interlockings as well as traffic-management systems (TMS) can be considered as intermediate step in a shorter horizon. For instance, the Norwegian approach (Website "Norwegian ERTMS Program") can be considered as an example. In all cases, we see a common set of actions with varying emphases:

- Replacement of (relay-) interlockings,
- Clear separation of interlocking and traffic-management layers, implying a reallocation of safe and unsafe properties,
- Revised principles to prove operational safety,
- Introduction of ETCS, usually beyond standard capabilities of Level 2,
- Usage of Automatic Train Operation (ATO) in different grades of automation,
- Partial shift from railway-specific solutions to industry standards (e. g. GSM, GPS).

Since capacity improvement is a core target of all programs, there is a severe need to express the capacity impact of the related system architecture. Such quantifications serve the broad portfolio from political decision processes ("what will be the gain?") to detailed requirement specifications ("how shall it ideally look like?").

This article contributes to the enhancement of blocking-time theory with the aim of representing the impact of the aforementioned technologies in existing principles of railway operations research (e. g. simulation, queueing). The text is setup as follows: Paragraph 2 gives a brief summary of the development of blocking-time theory so far, before paragraph 3 describes its enhancement to cope with future situations. Since related computations are usually performed by specific tools, paragraph 4 spots on implementation aspects of those necessary enhancements. Afterwards we raise attention and give insights on chances and obstacles to be taken into account when introducing future interlocking architectures in paragraph 5. In paragraph 6 conclusions are drawn and we summarize where further work needs to be done within the research community.

2 Blocking-time Theory so far

The blocking-time model has been introduced in the 50ies by HAPPEL (Happel (1959)) and in independently by ADLER a couple of years later. Implementation of the model in practical railway scheduling required several decades to pass, until computer-based scheduling systems became available. Standardisation for a broad audience took place at latest with (Hansen et al (2008)). With the emergence of ETCS, efforts have been made for an appropriate extension of the blocking-time model (Büker and Kuckelberg (2013)). Hereafter, all basics already described in (Hansen et al (2008)) are only outlined when required and the focus is laid on the evolution of the model to meet new/future architecture needs.

The blocking time is the total elapsed time a section of track, which is allocated exclusively to a train movement (and thus blocked for any other train movements). The track section may correspond to a whole block section but it may also be a subpart, for instance a route portion requested by two crossing routes.

The blocking time starts as soon as the preparations to issue a train its Movement Authority (MA) demand for exclusive occupation of a route's element. The MA must be issued before the train reaches a location at which a missing MA might cause a deviation from its scheduled train path since braking is triggered. (This corresponds to the principle of conflict-free timetabling.) The blocking time ends after the train has completely left the section and all signalling components have been reset to normal position, if needed, so that another MA with their involvement can be issued. Thus, the blocking time of a track section is usually much longer than the time the train occupies the section.

The blocking time does not embrace the time-demand to process a route request being issued by the train either via trackside equipment or by radio, since this time span does not yet require an exclusive occupation of track sections. Hindrances to train movements due to a wrong processing and provision order of route requests have to be handled outside of the blocking-time model.

Hereafter we denote such supervision systems as Automatic Train Control (ATC), which ensure continuous speed/distance supervision and provide continuous data transmission. All other supervision systems are classified as Automatic Train Protection (ATP). In this metric, ETCS Level 2/3 belongs to ATC while ETCS Level 1 (FS and LS) is an ATP.

Furthermore, we differentiate between safety distance and overlap beyond a signal. The safety distance exists physically and is bordered by an insulated train-joint or an axle counter (the danger point may be located at this border or even further away from the signal). It has to be cleared before a route to the signal can be setup. Often, but not always, the end of the safety distance corresponds to the Supervised Location (SVL). An overlap is limited by the same means but is longer than the safety distance and merely exists temporarily. Overlaps are usually installed in combination to route towards exit signals.

In the following paragraphs, we introduce aspects of the blocking time components with differentiation by signalling systems. In contrast to signalling-system specific definitions, the wording is chosen in a generic manner to allow usage for all types of usage. Whenever purposeful and known, practical implications beyond standard literature are mentioned.

Conventional Lineside Signalling

In Figure 1 the components of the blocking time in case of conventional lineside signalling are visualised. In the example, the brake initiation point to ensure standstill at the main signal is located ahead of the distant signal. In consequence, also the approaching time starts before the train passes the distant signal. Depending on the local configuration, the start of the approaching time and passage of the distant signal may also be at the same location.

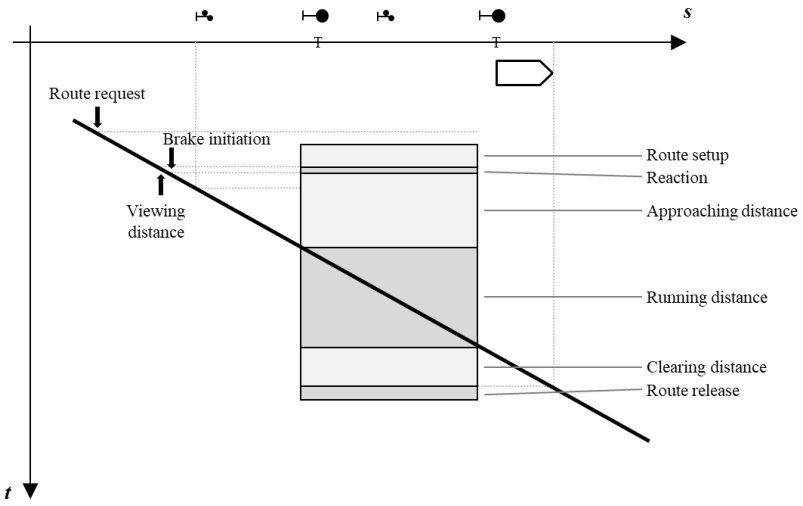


Figure 1: Blocking time for a block section

In Table 1 additional remarks on the components are stated if necessary.

Table 1: Selected time components in one/two-section signalling

Time component	**Remark**
Route setup	Preparation of the MA has to start sufficiently early that the signal aspect changes at latest with the start of the reaction time. The preparation covers moving switches, locking route elements, commanding the signal aspect. If multiple interlockings (IXL) are involved, their synchronisation cycles have to be taken into account, too.
Reaction	A reaction offset to interpret the distant signal aspect is granted to the train driver. It may be defined as a time or as a distance (in correspondence to minimum sighting distances). In case of a scheduled stop ahead of the track section the reaction time has to be replaced by the time demand for the departure process, which may usually be triggered just after the opening of the exit signal.
Approaching distance	May also start ahead of distant signal at brake initiation point

In layouts as sketched above, the approaching time is always related to the last distant signal in approach to the section. It may be either a separate distant signal (one-section signalling) or a combined main/distant signal (two-section signalling). If in two-section signalling the braking distance takes more than one block section, three-section signalling has to be applied granting two block sections for braking.

There are various principles to realise multi-section signalling as stated by (Pachl

(2005)). From the viewpoint of the blocking-time model, all have in common that a train-specific relationship between the start of the block section and the related distant signal has to be setup. Table 2 gives an overview of the impact on blocking-time theory.

Table 2: Specific time components in three-section signalling

Time component	Remark
Route setup	If signalling principle requires commanding signals to black aspect, feedback time to IXL needs to be taken into account.
Approaching distance	Starts multiple blocks ahead of the investigated track section, computation requires a logic to denote the relevant start

ATP Cab Signalling

In case of cab signalling the overall principles remain very similar. Merely the approaching time is determined by the time the train runs through the indication distance that is signalled by the cab signal system. The start of indication distance goes along with a change in the driver-machine display (DMI) indicating to expect the end of the current MA. In Figure 2 the indication distance in case of ETCS Target-Speed Monitoring (TSM) is illustrated. As soon as the train reaches the speed-distance function denoted by „I" the Movement Authority has to be extended to guarantee hindrance-free operation. The diagram is based on UNISIG Subset 026-3 but enriched by a visualisation of SBD/SBI1 principles. In the given parametrisation, SBI2 is decisive for Warning, Permitted and Indication Curve, anyway. A separate Guidance Curve replaces the Permitted Curve if not inhibited by National Values.

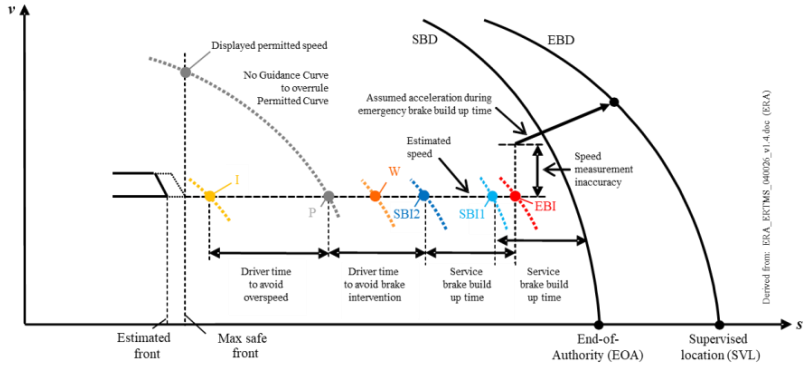


Figure 2: Indication distance in case of ETCS Full Supervision

Since ATP systems rely on dis-continuous data transmission, the MA has to be updated at latest at the transmission spot (e. g. balise group) which is closest to the start of the indication distance. Table 3 summarizes the major differences to the previously discussed principles with the nomenclature being related to ETCS. The mechanism and time components introduced above are applicable to comparable Class B systems (e. g. EBICAB, TBL2).

Table 3: Specific time components in ATP cab signalling (here ETCS Level 1)

Time component	Remark
Route setup	If cab signalling is operated in combination to dark signals, feedback time to IXL after command dark aspect needs to be taken into account.
MA creation	MA is derived either by Lineside Electronic Unit (LEU) from signal aspect or, in case of centralised LEU, from state of route elements and provided to trackside balise group (BG).
MA transmission	Usually via air-gap between BG and trainborne balise antenna
MA interpretation	By ETCS on-board unit (ETCS-OB)
Reaction	Indication distance usually covers a reaction time. Nonetheless, an additional in-advance reaction time may be granted to the train driver.
Odometer error front	Display of indication curve is triggered under consideration of the distance run since the last reset of the train-borne odometry, which requires a continuous adaption of the EBI curve.
Approaching distance	Starts at the transmission spot which is closest to the start of the indication distance

ATC Cab Signalling with Fixed Blocks

For continuous data transmission in case of ATC cab signalling, there is no need to consider the transmission spot closest to the start of the indication distance. (Ideally, most trackside balise are of passive nature.) Instead, the start of the indication distance equals the start of approaching distance. This way, balise engineering is simplified and approaching times are minimal per train movement. In contrast to the ATP case, additional blocking-time components have to be accounted for. They are listed in Table 4.

Table 4: Specific time components in ATC cab signalling (here ETCS Level 2)

Time component	Remark
MA creation	MA is derived by Radio Block Centre (RBC) from IXL data, usually status of switches and main signals.
MA request	Applicability depends on RBC architecture. While in early ETCS Level 2 implementations the RBC transmits the MA as soon as created (“push”), the RBC waits for the train’s request in newer implementations (“pull”). Requesting starts in time offset T_MAR before reaching the indication distance and then happens cyclic. In worst case, the whole cycle time has to be considered ahead MA transmission.
MA transmission	Via GSM-R (today) from RBC to ETCS-OB
Approaching distance	Starts at the indication distance

Again, the mechanisms and time components described above are transferable to comparable class B systems (e. g. LZB, TVM). In recent ETCS Level 2 implementations, e. g. at Gotthard base tunnel, the Train Position Report (TPR) is used to facilitate the release of

the safety distance – but not the block section – if certain conditions on the train speed are met. Additional blocking-time components have to be taken into account as enumerated in Table 5. Attention is drawn that an application of TPR is applicable only in combination to train integrity detection.

Table 5: Specific time components in recent ETCS Level 2 implementations

Time component	**Remark**
Clearing distance	The clearing distance (safety distance) does not necessarily be bordered by trackside equipment if TPR is applied to validate liberation of the safety distance.
Odometer error back	As for the Max Safe Front End in TSM, also the outermost (virtual) train end needs to be accounted for, once TPR is used.
Train position report	Position reports are sent out by the ETCS-OB in fixed cycles of T_CYCLOC or they are triggered periodically after passing D_CYCLOC (being related to the train front).

If TPR is in use, it may also serve the release of overlaps once the train has come to standstill (or at least underruns a threshold speed) instead of linking the release of overlaps to IXL-based section timers. If the IXL-based section timers are quite conservative, this may improve the capacity, since the overlap distance can be occupied by another route at an earlier moment.

For both applications of TPR we see the necessity to merge IXL and RBC functionality (or at least the need for a powerful bidirectional interface). Addressing future architectures in the next paragraph, the shift of responsibilities between IXL and RBC is intensified.

3 Enhancement of Blocking-time Theory for New Architectures

There are many different reasons to focus on the further development of IXL- and RBC-architectures. One reason mentioned already above is the necessity to combine IXL-, RBC- and TMS functionalities to guarantee a powerful interaction between these systems. Another reason comes from the economic view: the life-cycle costs of today's control command and signalling (CCS) systems are relatively high and therefore infrastructure managers' aim for their reduction. This shall be achieved by one main principal, namely limiting the outside CCS components with the use of ETCS Level 3. Dynamic block sectioning and ETCS Level 3 form a system where a train route can start and end anywhere with the use of cab-signalling and train integrity inspection. With these improvements neither outside signals nor outside track clearance equipment are mandatory. The related geometric interlockings (GIXL) will evaluate all safety relevant real-time data continuously. Thus, it will be possible to reduce double safety margins used today, to increase the system performance. If GIXLs are combined with a powerful TMS, many functionalities of todays' interlockings can be shifted to the TMS, so that the amount of "SIL 4" functions within the interlocking can be reduced.

One representative of the development of such architecture can be seen with the development of the ETCS-Interlocking (EI) with smartrail 4.0 in Switzerland. The idea behind an EI is to use digitalization to reduce the necessary outside CCS-components by up to 70 % (Grabowski and Schmidt (2018)). The trend to merge at least IXL and RBC can already

be seen in ongoing L2 implementations. For the enhancement of blocking time theory, we assume the existence of GIXL like EI and refer to “smartrail 4.0” (Website smartrail).

3.1 New Blocking-time Components

Together with dynamic block sectioning, ETCS Level 3 can be seen as a time-discrete system instead of a distance-discrete system. This causes that the occupation for a train run comes close to a blocking-time band, whereas with all other system designs (such as ETCS Level 2), the occupation will always look like a blocking-time stairway.

In earlier studies, ETCS Level 3 systems have been modelled as blocking-time band (Büker et al (2010)). During further investigations, especially with smartrail 4.0, the assumption has been revised, because the subsystems (GIXL, RBC, TPR) work periodically. One can say that even an ETCS Level 3 system is still discrete, but the time-steps in the blocking-time get more regular, depending on the different system cycle-times. The blocking-time band looks as displayed in Figure 3.

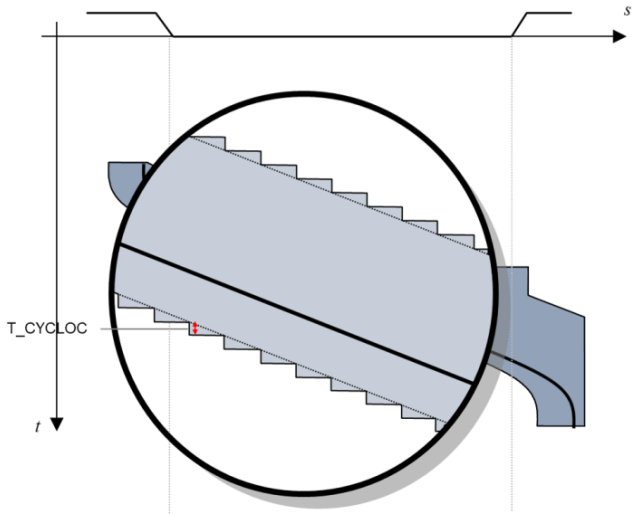


Figure 3: Discrete moving block outside fixed elements like points

In table 6 new time components of a time-discrete system are introduced.

Table 6: Specific blocking-time components in ETCS Level 3 signalling

Time component	Remark
MA creation	As above, but at least cycle time of RBC
MA preoccupation	Optional time/distance buffer to ensure smooth operation
Running distance	The track section melts down to (infinitesimal) short distance, if dynamic block sectioning is applied.
Clearing distance	Each train movements pushes its virtual safety distance ahead.

To provide the whole picture, Table 7 marks the applicable components of blocking times for the whole range of architectures (including CBTC architectures).

Table 7: Blocking-time components including recent/future architectures

Time component	System	Optical signalling, Level 1 LS	Level 1 FS	L2 „push“ with pure track sections	L2 “pull” with partial TPR	L3 „pull“
	Route setup	X	X	X	X	X
System Times Pre	**MA creation**		LEU	RBC	RBC	RBC
	MA request				X	X
	MA preoccupation					X
	MA transmission		Air gap	GSM-R	GSM-R	GSM-R / FRMCS
	MA interpretation		X	X	X	X
	Reaction	X	X	X	X	X
	Odometer error front		X	X	X	X
	Approaching distance	Distant signal	Last balise group for MA Extension	Indication Distance	Indication Distance	Indication Distance
	Running distance	X	X	X	X	
	Clearing distance	X	X	X	X	Virtual
	Odometer error back				X	X
System Times Post	**Train Position Report**				X	X
	Route release	X	X	X	X	X

3.2 Interaction with Automatic Train Operation (ATO)

In the context of the IXL architectures discussed within this paper, Automatic Train Operation (ATO) is mainly envisaged to improve the system performance and reduce energy consumption. It does not primarily contribute to an increase of safety, since it is applied in combination with an Automatic Train Control (ATC). The most popular combination, even though not yet fully standardised, is ATO-over-ETCS. Potential applications of ATO in combination to non SIL-4 systems are out of the scope of this article. Since ATO-over-ETCS according to subsets 125, 126 and 130 seems on its way to industry standard, we hereafter refer to this specific system, whenever needed.

The on-board ATO (ATO-OB) maintains a train run within a defined tolerance of its path following a particular target function (for instance energy consumption). The system

marginally adjusts operating parameters such as the ratio of power to coast when moving. Ideally, the path can be readjusted by the TMS to the current situation and transferred by the trackside ATO (ATO-TS) to the ATO-OB. This results in an outer control loop by the TMS and an inner control loop on the rolling stock, as described by (Weidmann et al (2014)), with both components following particular target functions. Communication from ATO-TS to ATO-OB is, in a simplified description, backed-up by static segment profiles and dynamic journey profiles with the latter ones covering the train-path specific timing points.

There are five Grades of Automation (GoA) of trains with GoA 0 being regular on-sight train operation. Right now, two variants are of highest interest:

- GoA 2 means semi-automatic train operation where starting and stopping is automated, but a driver operates the doors, drives the train if needed and handles emergencies.
- GoA 4 means unattended train operation, where starting and stopping, operation of doors and handling of emergencies are fully automated without on-train staff.

In the scope of timetable and capacity modelling, ATO impacts various aspects such as easily inserting new trains to react to unforeseen peak of demand. Some of them, as identified so far, are addressed in the following paragraphs. While the consequences for modelling are mostly elaborated, there is a lack of published study results or even of in-field experiences. At least the outcomes of a study on potential improvements on the suburban "S-Bahn" network in the Stuttgart area are available (Website VM Baden-Württemberg).

Reduction of Approaching Times

In Supervised Speed Envelope Management (SSEM), the on-board-ATO establishes the maximum speed the train can run without interfering with the ETCS speed limits. In TSM this means, the ATO-OB shall drive the train so as not to reach the EBI curve. For this purpose, the ATO-OB (re-)computes various speed-distance functions within the current MA using the information sent by the ETCS-OB. In particular the ATO-OB inhibits the service-brake command being triggered by overpassing an SBI supervision limit as well as any "Sinfo" sounds in relation to speed and distance monitoring. In consequence, the set of ETCS information/intervention curves (cf. Figure 2) melts down to the outermost EBI.

Instead, an appropriate representation of ATO-borne requires consideration of ATO-OB specific time components. As this part of the system behaviour is vendor-specific one has to make assumptions how to represent the properties in a generic manner, which allows fine-tuning as soon as in-field experiences have been collected. According to expert judgement, the following model seems promising:

- By means of maximum brake decelerations, a SBD-equivalent for ATO is defined.
- An ATO service brake build up time allows deriving an SBI-equivalent.
- To avoid ETCS emergency brake intervention, a "buffer time" ensures sufficient computation times for the ATO-OB control loop and missing synchronisation between ETCS-OB braking curves and their ATO-OB replica. Whenever the ATO-SBD injures the ETCS-EBI, it is replaced accordingly.

In Figure 4 the two additional ATO speed-distance functions and their interaction with

ETCS speed-distance functions are visualised. By comparison to Figure 2 it becomes evident which time components lose their relevance for the indication distance.

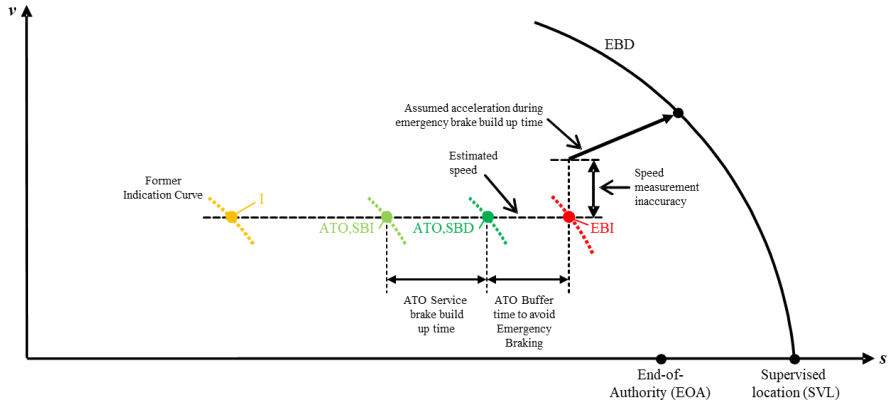


Figure 4: Indication distance in case of ATO-over-ETCS

If ETCS Level 1/2 is implemented as overlay to the signalling layout of a high-capacity Class B system, often an increase of indication distance and thus a loss of capacity emerges. This is due to the multiple convenience and safety margins, which also take effect if the system is configured without service-brake intervention in TSM. By means of ATO-over-ETCS, a considerable share of the convenience margins are suppressed and disadvantages from the ETCS implementation on capacity consumption are mitigated.

Reduction of Regular Supplements

Best practice amongst railway operators is to augment technical/minimum running times by supplements to compensate randomness of technical/minimum running times to an appropriate extent. Such randomness may arise from dwelling time delays, human driving behaviour, weather conditions, track-wheel adhesion or availability of full rolling-stock characteristics. (Running-time extensions because of track works are taken into account separately.) The margins are usually defined as a relative increase of technical/minimum running times or as an additional running time per running distance. In the first metric, values of 3 % to 7 % are common. The resulting running time is referred to as regular running time. It forms the basis for the timetable compilation. As well, the corresponding speed profile serves the computation of approaching times in blocking-time theory.

By means of (at least) semi-automatic driving, at least the stochastic (human) impact of the driver is eliminated from acceleration, coasting and deceleration processes. As long as there is the same journey profile, same (version of the) on-board equipment and the same train/track conditions, the same trajectory shall result thanks to ATO-OB. Furthermore, an automatic train operation can regulate the jerk very accurately, which can reduce the running time thanks to too high acceleration and deceleration. This means less deviation of the actual running times. An example of human-driven trajectories versus ideal trajectories on the Belgian network is provided by (Bienfait et al (2012)). (Unfortunately, speed-distance diagrams of later test-runs on the Brussels – Leuven line with ATO in comparison to drivers are not published. As well, similar experiences from ATO operation on the RATP network are not publicly available.)

Following the logic of regular supplements to eliminate the impact of the largest share of randomness to the timetable, if the new technical systems does not need new time supplements, the magnitude of regular supplements could be reduced. This would allow to:

- Reduce scheduled running times and thus reduce travel times. (It may result in a little increase of capacity consumption because of higher approaching time and because of higher heterogeneity of train paths, anyway.)
- Keep scheduled running times and facilitate the gained supplements for delay reduction whenever needed. This way, stability of the overall timetable concept is increased.

While the mechanisms are known, knowledge on their magnitude still needs to be gathered as only very few main-line systems have been taken into operation with ATO. (For comparability, urban metro systems can only be compared with limitations, as they are often operation CBTC-borne from their early days.) In some countries a specific regular time supplement is added to compensate variations of adhesion and driving style. At a first glance, a reduction of the regular supplements by a third seems appropriate according to expert judgement.

Backward Compatibility

In tools of railway operation research, randomness is taken into account whenever simulations of operation are performed and is being represented. Usually, operation is disturbed by delays at entry and by primary delays at stops. Furthermore, running times may be increased randomly. To ensure a precise representation of knock-on delays, reaction times are incorporated. The general algorithmic representation of railway operations is quite deterministic, nonetheless. As we see above, ATO contributes to the reduction of a part of the stochasticity from railway operation. From system design one can conclude, that reactions times are replaced by transmission times with transmission times being relatively short as ATO is intended to be a non-safe system. Thus, reacceleration after an unscheduled standstill may start earlier than in today's operation.

In the close future, there will be various studies to assess the benefits of ATO on specific infrastructures. To ensure reliable outcomes and a precise differentiation between system behaviour with/without ATO, stochastic properties have to be represented properly. With regard to the mostly deterministic representation of the current non-ATO operation, the challenge within the tools may be rather to create a more random representation of the status quo operation while only fine-tuning to the aforementioned ATO aspects.

3.3 Constraints of Moving Block Application

Applying the moving block principle is subject to various constraints as it has already been shown in (Büker et al (2010)). Depending on the specific infrastructure design, certain sections have to be operated in (virtual) blocks in any case, as trains should not come to standstill for various reasons:

1. Moveable elements (switches, bridges) can only be occupied as a whole.
2. Overhead catenary design (OCS) design does not allow standstill.
3. Initial traction effort is too low to ensure reacceleration.
4. Maximum coupling forces may avoid reacceleration.

While the two latter constraints may rather happen on the open line along steep inclinations, the two first constraints usually take effect in station areas. In consequence, they frequently become decisive for minimum headway times and foil the benefits of the moving block principle as shown by Figure 5:

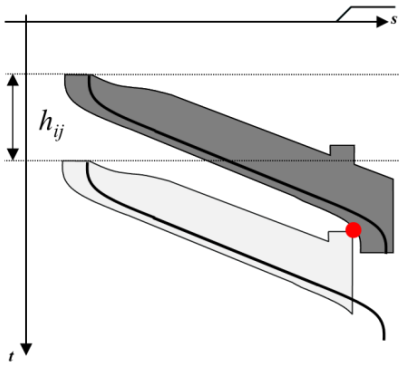


Figure 5: Moving block being interrupted in switchpoint

Virtual blocks may also be caused in station/bifurcation areas to avoid partial occupation of heavily loaded track sections: Let there be a train sequence using the same track beyond a bifurcation with the first train being a suburban train stopping in the vicinity of the bifurcation and the second train being a heavy freight service. With the first train moving on the moving block of the second train consequently occupies the bifurcation as a feature of geometric interlocking. Once the stop of the first train takes significantly longer than expected, any third movement along the bifurcation is excluded.

In conventional block sectioning this effect might be avoided by the static block logic not granting a route to the second train at all. In geometric signalling this – and the implication 2-4 stated above – have to be mitigated by the TMS layer instead.

4 Implementation Aspects

The implementation of ETCS toolboxes or modules for different levels, specifications and purposes itself is a challenge. Starting from scratch, the implementation of the aforementioned features might seem more or less straightforward. Anyway, integrating the new aspects into the traditional approaches for running-time and occupation calculation (including derived functionalities like conflict detection) has to consider existing data structures as well as integration and extension of data-exchange interfaces, backend databases or graphical interfaces and output graphics and diagrams.

4.1 Running Time Computation

The common understanding and basic approach for current microscopic tools realizing detailed running time computations is to start with a (static) speed profile and stopping policy for a single train run. Based on this input, the highest (technical/physical) train speed is determined and the shortest running time is calculated. Additional time margins, regular and specific supplements are applied afterwards, resulting in a (decreased) running speed corresponding to the new running times.

Within a forward-oriented loop, the train speed is increased as long as it remains within speed profiles. Due to traction forces resulting accelerations might be negative ("missing traction power"), but principally these phases intends to determine the highest physical running speed possible.

In contrast, the character of braking phases usually follows a backward-oriented approach: Within the forward oriented determination of running speed the change of profile speeds towards a lower value requires braking phases towards the (new) target speed. Typically, braking phases are determined by (fixed or value-dependent) braking acceleration values. With these values, one braking curve anchored at target speed and location can be computed. Following that curve, the intersection point with the former train run can be identified as braking point. Braking accelerations are usually predefined (theoretical) values representing the "usual" braking behaviour of trains respectively train drivers. Braking phases towards scheduled stops are treated equally with a target speed of zero. Moreover, a "green wave" is assumed usually implying that ATP/ATC influence remains inactive with respect to running time computation.

The introduction of ETCS, especially ETCS Level 3 plus dynamic/moving blocks affects the well-proven running time principles for some reasons:

- Braking acceleration values become much more dynamic due to braking models, national parameters and multiple dependencies from train and traction characteristics.
- A large set of probably overlapping braking curves is computed, the most restrictive curve has to be derived from that set and the resulting curve might be a section-wise partitioning of multiple curves.
- For lowering speed changes the most restrictive curves might become relevant, the static braking acceleration cannot be used any more.
- Most restrictive curve or any derived simplification like static braking accelerations underriding the most restrictive curve is much more complex than any other legacy supervision curve or braking model.

Moreover, especially for ETCS Level 3 that eliminates "traditional" spatial blocking, the semantic of braking processes has shifted towards a "forward oriented" braking computation, imitating the behaviour and decisions on a train driver with respect to DMI visualizations.

4.2 Curve Variations and Dependencies

Depending on the usage of ATO the braking curves might vary, time components have to be considered respectively ignored. The availability of ATO functionality is dependent from train characteristics as well as trackside equipment that moreover might be installed only partially.

Therefore the implementation has to extend the trains by appropriate properties but also the infrastructure model has to be enhanced by information about availability and type of ATO infrastructure, e.g. directed begin and end graph nodes for a microscopic network graph including technical ATO properties. These begin and end elements have to be evaluated for each train passing these elements and the validity and consequences for the train runs respectively running time and occupation computations have to be realized, resulting in dynamic property vectors that have to be considered by the ETCS curve computation. Finally, to ensure that existing ATO functionality is available, the complete braking curve

from braking point to EOA has to be within such an ATO enabled area. This might imply a repeated computation if the assumption, that ATO is available fails while computing due to ATO area leaving.

4.3 Occupation-time Computation

The classic blocking-time theory determines braking points based on braking curves with static braking acceleration values. Depending on the train control system used, the point of MA submission is derived from braking points, system or transmitting times are added and begin time of succeeding block sections are derived from these indication points. Similar to running-time computation, the determination of ETCS braking curves has to happen for each possible EOA and for each change to a lower speed.

The quantity of braking curve computations raises in case of ETCS Level 3 with dynamic block sectioning, because EOA becomes approximately continuous which increases the more complex ETCS curve computation extensively. Therefore, the implementation considerations concerning shifting paradigms from backward- to forward-oriented braking computation become much more relevant. Finally, the integration of this reasonable paradigm shift and its integration into existing tools and applications are another implementation problem. Current railways operation functionalities add another implementation complexity dimension, e. g. tools simulating train operation have to handle dynamic aspects like occupied block sections etc. that disturb the ideal world of green wave planning. For train timing, red signal and knock-on-delays the management of complex ETCS braking curves additionally complicates the implementation.

With ETCS Level 3, one new aspect has to be incorporated: Cyclic system times and partial discrete occupation element. While cycle times resulting from system component frequencies discretize the theoretically continuous occupation band in time (cf. paragraph 5.1), e.g. technical times for switch changes moreover contradicts occupation band continuity which has to be considered by the implementation as a new challenge additionally.

Conflict detection has to be modified, too. While former conflict detection evaluates overlapping time ranges for single occupation elements, microscopic conflict detection for ETCS level 3 has to evaluate band overlaps instead, where upper and lower boundaries of occupation bands are derived as location-time regions. A regional overlapping has to be performed for conflict detection instead. The determination of e.g. minimum headway times also has to be adapted if implementing the new approaches. While former times are derived from time buffers between two occupation blocks, the determination of buffer times between occupation bands follows a continuous distance detection approach between two regions in space. From an implementation point of view it might be interesting, which algorithms and probably which granularity of discretization are used to detect overlapping regions.

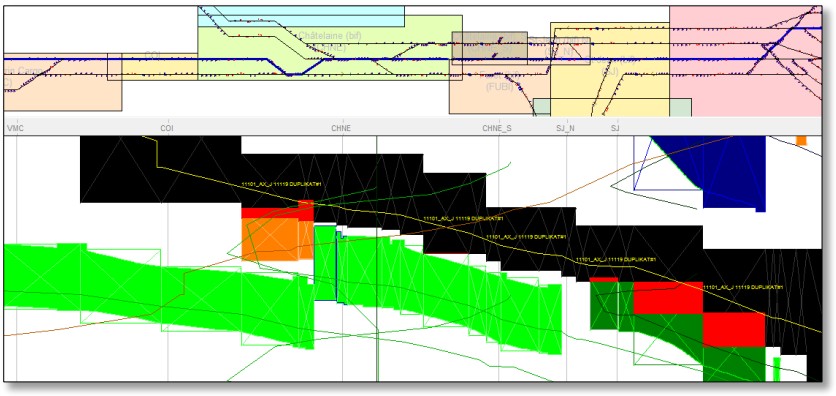


Figure 6: Sample implementation of ETCS Level 3 occupation bands (green & blue) and traditional occupation blocks (black) including conflict detection (red), sectional bounding boxes and discrete switch occupations.

4.4 Occupation Points, Occupation Areas and Model Synchronization

While the two previous paragraphs described pure implementation aspects of running-time and occupation-time computation, more complexity arises when synchronizing different occupation paradigms, e.g. mixing up classical, section based occupation graphs and occupation bands from ETCS Level 3. Integrating the continuous occupation bands into sectional occupation blocks requires some activities:

- (Spatial) segmentation of occupation bands due to corresponding occupation sections.
- Relating band segments as “virtual occupations” to occupation elements.
- Detecting the bounding box for each band segment and setting begin and end times of the associated virtual occupation accordingly.

When segmenting occupation bands, some interesting implementation aspects are:

- Occupation areas around switches are assigned to two occupation blocks, the switch itself and the preceding block.
- The discretization of occupation bands results in more or less detailed saw-tooth bands within occupation blocks.
- Switches have additional technical setup times, e.g. route setting times, that do not move in the same way the ETCS blocks do. Therefore switches can usually be identified easily due to the “blister” shown on top of the occupation band.

Mapping occupation bands in that way, it is possible to study and analyse “mixed scenarios”, e.g. ETCS Level 3 controlled trains and trains operated under conventional train control (fall back or mixture of differently equipped trains). This approach moreover directly fits into existing conflict detection and solving paradigms of succeeding tool functionalities like capacity assessment (analytical approaches, simulations, UIC 406 etc.),

therefore the approached presented within this paper can completely be integrated into existing tool implementations and their functionalities.

5 Decisive Effects on Headway Times

As described above, ETCS Level 3 (with absolute braking distance) in combination with GIXL and TMS architecture requires an enhancement of the blocking-time theory taking into account all previously new defined time components. Resulting minimum headway times drop, as expected, but various effects have to be taken into account. For the discussion of decisive effects of an ETCS Level 3 system in combination with geometric interlockings we will focus on the minimum headway times as well as the maximum element occupation time of a set of trains depending on the used CCS.

5.1 Cycle Times

For the horizon of smartrail 4.0 it is assumed that for example driven by autonomous street cars, the localization technology will be affordable and precise enough to be used within the railway system. It is currently aimed for a localization accuracy of 1 m and a fail-safe data transmission (Website smartrail). Since in current ETCS level 2 implementations the TPR runs in cycles of around 5 seconds and needs to be evaluated by the RBC (working in cycles as well), this leads to the steps on the bottom of the moving block, which are shown in Figure 3. For the evaluation of minimum headway times it has to be assumed that in the worst case two trains follow each other with TPR cycles not being synchronized, which will increase the headway time by T_CYCLOC.

Since the TPR runs periodically and requires radio transmission, it has to be evaluated, if the TPR is sufficient to be used in heavily used track topologies for track-clearance detection. A way to avoid problems with the needed timespan would be to use track-clearance equipment (for example axle counters) in dense areas, since they might work faster and use the TPR for track-clearance on the line. To get any major benefits in terms of headway times a TPR-cycle would need to be reduced to about one second (or less), which would result in even smaller vertical steps in the running-time band (manufacturer survey).

5.2 Switchpoints

In legacy IXL a switchpoint is always covered by a signal. This causes, that in the blocking-time theory a switchpoint is already occupied as soon as the approaching distance of the corresponding signal is reached. In case of L3/GIXL, the occupation of switchpoints has to be taken into consideration separately. As stated above, switchable elements can only be occupied as a whole. Speaking in terms of blocking-time theory, the whole length of a switchpoint has to be preoccupied at once, taking into account that already before reaching the approaching distance a time span for the turnaround of a switchpoint is necessary.

In Figure 5 it is visible, that especially switchpoints may become the decisive element for determine headway times, since the occupation of a switchpoint starts earlier (due to switchpoint turnaround time) than the occupation of the track section directly in front of and behind the switchpoint. The capacity effects of a moving block are noticeable limited by switchpoints. The negative effect on capacity could be reduced by different approaches:

1. Use of smaller switches to reduce the occupation length
2. Investigation, if it is possible (and has benefits) to only preoccupy the moveable

parts of a switchpoint instead of the whole switchpoint

3. In advance swinging (by TMS) to avoid turnaround time affecting blocking time
4. Possibility of turning all switches simultaneously instead of one after one
5. Reduction of switchpoints

5.3 Speed Changes

One issue that currently reduces the positive effects of the moving block is the speed supervision in speed changes with ETCS as standardised with Baseline 3. For running-time calculations, the permitted curve (cf. grey curve Figure 2) is used, even though ATO is assumed. A yellow status of the DMI is accepted, thus. In Figure 7 we take a look at a train sequence of two (passenger) trains. It is visible, that the permitted curve reaches the target speed significantly earlier as the speed change is mandatory. This distance is mostly dependent of the trains' braking characteristics. If both trains are following each other with the minimum headway time needed at 140 km/h (in this example 70 seconds), the train sequence of these two trains is not conflict free in the speed change any longer, resulting in an occupation time conflict of 21 seconds. If the headway time is increased up to 91 seconds (second train shifted by 21 seconds), the train sequence would be conflict free again.

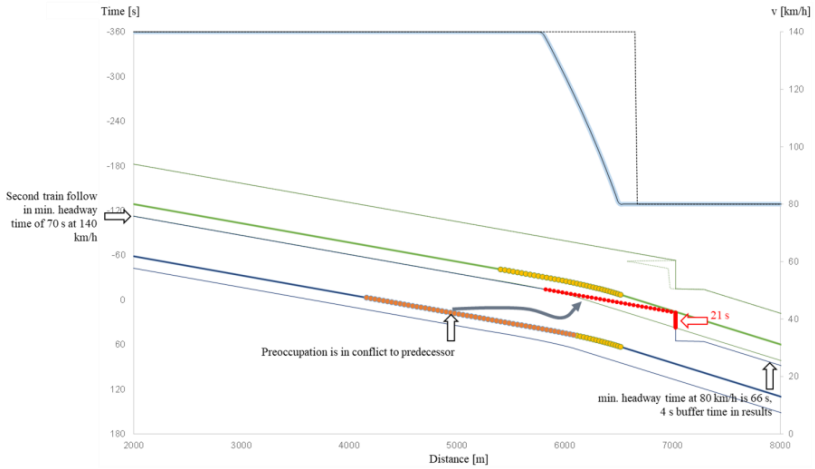


Figure 7: Permitted Curve towards v-Step

The negative influences on the minimum headway time in speed changes can be minimized, if a second speed step is implemented and the whole sequence is run at yellow DMI. This scenario is presented in Figure 8. This does increase the running time of the trains slightly, but reduces the minimum headway time at the same time.

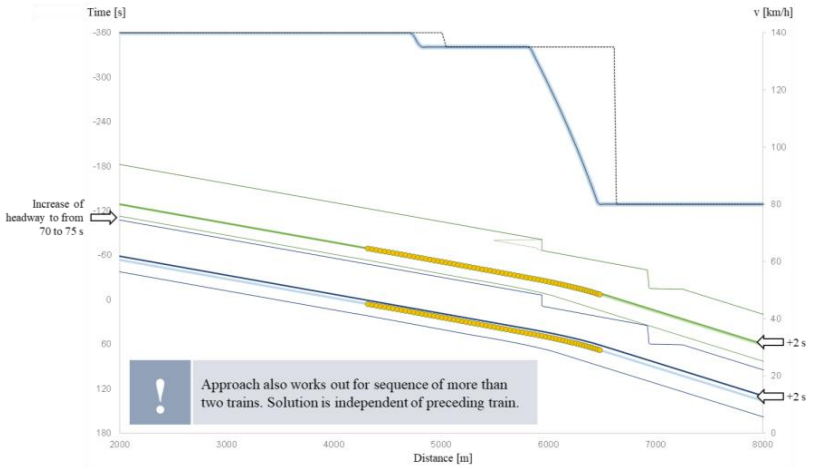


Figure 8: Optimized train sequence in v-Step

The benefit in using two speed changes instead of one is that we preoccupy the section between the two speed changes with the permitted distance and not with the indication distance, which results in a reduction in minimum headway time. In the given example, the minimum headway time between the two trains could be reduced by 16 seconds (from 91 to 75 seconds) with the introduction of the intermediate v-step. The impact of the intermediate v-step on the running time is comparable small (2 seconds) and should be accepted for a reduced minimum headway time. With this approach it is possible to increase capacity (and reduce the minimum headway time) if a speed change is the decisive section for the minimum headway time of two trains following each other directly. Adding more than one intermediate speed change does not reduce the minimum headway time any more since the benefits come mostly from the use of the permitted distance in between the two v-steps.

The issue of increased running time with ETCS Full Supervision (FS) due to more restrictive speed supervision in speed changes is valid for all current ETCS FS systems. It has to be discussed, whether an ATO has to follow the permitted curve in speed changes, or if it is possible to run beyond the permitted curve to reduce the running time again.

5.4 Exemplary Capacity Gains by smartrail (EI)

To evaluate the capacity effects that can be gained in the smartrail setup, we have conducted various analysis, with examples being presented here. In a first analysis we calculate the occupation times for three different existing railway lines (one optimised for freight trains, one for long distance passenger trains and one for local passenger trains) for every main signal on these lines. One major result of the calculation of occupation times is, that with EI in combination with ETCS Level 3 the distribution of occupation times becomes more homogeneous and all in all the occupation times can be reduced. Details can be seen in table 8, being based on the following assumptions:

- ETCS braking model without Service Braking (SRS 3.6.0)
- Preoccupation starts with the indication distance

- Fixed odometry confidence interval
- Running time sufficiently long to pass a diverging switchpoint with the given speed

Table 8: Exemplary distribution of technical occupation times

	Conventional signalling [s]	**EI with L3 and ATO [s]**
5 %	78,9	68,8
25 %	99,7	74,9
50 %	123,7	88,8
75 %	150,1	141,3
95 %	193,1	163,2
Average	128,0	103,5

The enhancement of the blocking time theory is prototypical implemented in LUKS®, which is a software tool for the assessment of capacity on railway networks. With this tool, we are able to validate timetable concepts. For timetable compilation we use the following assumptions:

- Preoccupation: Indication distance
- Running time calculation: permitted curve
- Buffer for operational quality is included in the route clearance time, thus not visible

In Figure 9 a screenshot of a future timetable is given. On the top part, all trains use conventional signalling. There it is visible, that the desired train sequence is not possible without blocking-time conflicts (purple). In the bottom part, we see the same time-slot in new architecture. There it is obvious, that the same timetable does not have blocking-time conflicts anymore. This given example is taken from a current timetable validation project and the same issues can be seen in different locations and constellations.

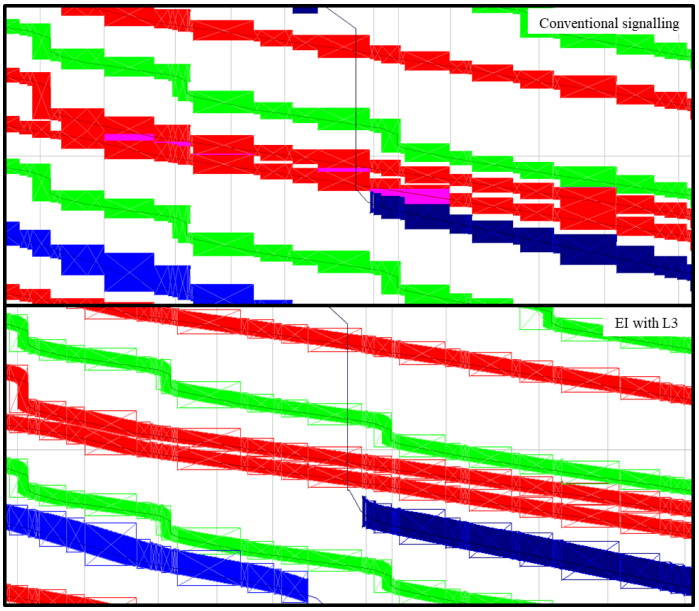


Figure 9: Exemplary time distance diagram with conventional IXL (top) and EI (bottom)

At large, it can be said that EI in combination with ETCS Level 3 can lead to a reduction of minimum headway times, which can increase the capacity of a railway infrastructure. Never the less there are more boundary conditions which determine the overall capacity of a railway infrastructure which cannot all be solved with this new architecture.

6 Conclusion and Future Challenges

Recent studies lead to an enhancement of the blocking-time model to represent future architectures (ATO, GIXL, RBC and TMS). With their introduction, a contribution towards capacity increases on today's railway network can be made – but mostly only, if all subsystems are introduced as a whole.

In case of ETCS, the safety level is adjusted by a combination of distance EOA-SVL, by National Values and by rules how to compute the train's brake capability. Even though the underlying Braking Curve Model is flexible, it cannot be adapted to every situation. This results in a safety surplus in certain situations. A continuous re-examination of the overall situation shall be part of geometric interlocking and the safety distance is virtual in case of ETCS Level 3. So far it has not yet been elaborated, to which extent a continuous adaptation of the safety distance at the desired safety level may serve a reduction of approaching time.

As popularity of ATO-over-ETCS grows, we see further needs for research and development. To our best knowledge, the following things should be addressed in the future:

- Behaviour of ATO-OB for timetable compilation tools and for TMS

- Influence of ATO on system times
- Influence of ATO on regular supplements (less stochastic influences)
- Legal aspects of the use of ATO

After the uncertainties regarding the ATO-specification and -implementation have been resolved and it becomes foreseeable to which extend the assumptions made have been met, it will be possible to integrate especially ATO in software tools for timetable compilation. Before that has been achieved, all assumptions on capacity will be preliminary and we strongly recommend to not overestimate the system by making unlikely assumptions.

References

Bienfait, B.; Zoetardt, P.; Barnard, B.: *Automatic Train Operation: The mandatory improvement for ETCS applications*. At: ASPECT IRSE, London 2012.

Büker, Th.; Kogel, B.; Nießen, N., 2010. *Influence of the European Train Control System (ETCS) on the capacity of nodes*, Paris, May 2010, ISBN 978-2-7461-1801-0.

Büker, Th.; Kuckelberg, A., 2013. *ETCS approximation by legacy ATP/ATC systems under railway operation research methods.* In: Signal+Draht 105(2013) issue 10.

Grabowski, D.; Schmidt, S.: *The "ETCS Interlocking".* Signal+Draht 110(2018) issue 10.

Happel, O., 1959. *Sperrzeiten als Grundlage für die Fahrplankonstruktion*. In: Eisenbahntechnische Rundschau (ETR) 8 (1959) 2, pp. 79-90.

Hansen, I.; Pachl, J., 2008. *Railway Timetable & Traffic: Analysis - Modelling – Simulation*, Eurailpress (2008).

https://www.smartrail40.ch (checked 19th Jan. 2019)

https://www.deutschebahn.com/de/presse/suche_Medienpakete/medienpaket_digitale_schiene_deutschland-1177310 (checked 19th Jan. 2019)

https://www.banenor.no/en/startpage1/Projects/ERTMS-National-Implementation (checked 19th Jan. 2019)

https://vm.baden-wuerttemberg.de/fileadmin/redaktion/m-mvi/intern/Dateien/Praesentationen/181024_Praesentation_Verkehrsausschuss_ETCS_DSTW_Pilotprojekt_Stuttgart.pdf (checked 19th Jan. 2019)

Pachl, J., 2005. *Modelling Specific Signalling Features in Computer-based Scheduling Systems*. In: Rail Delft 2005 Proceedings, International Association of Railway Operations Research.

Weidmann, U., Laumanns, M., Montigel, M., Rao, X, 2014. *Dynamische Kapazitätsoptimierung durch Automatisierung des Bahnbetriebs*. In: Eisenbahn-Revue 12/2014.

Fact-checking of Timetabling Principles: a Case Study on the Relationship Between Planned Headways and Delays

Fabrizio Cerreto [a,1], Megan Holt Jonasson [a]
[a] Traffic Division, Rail Net Denmark
Carsten Niebuhrs gade 49, 1577 Copenhagen, Denmark
[1] E-mail: fceo@bane.dk, Phone: +45 82349172

Abstract
The tradeoff between reliability and level of service is a central focus for railway operators and infrastructure managers. A well-performing timetable must include an optimal level of buffer time between conflicting train movements, such that a high service delivery and a high service quality are maintained. This focus on buffer time has informed the research within the fields of timetable optimization, capacity utilization and delay propagation modeling. Despite recent and ongoing advancements in these fields, there are still disconnects between the theoretical models and their application in the design, planning and evaluation of railway timetabling. Parameters that are used in timetabling, as well as, as input to the analytical assessment models, are typically derived from practical experience and based on the macroscopic limitations of a system, rather than the microscopic conflicts inherent in its signaling system.

The objective of this paper is to support the design of fact-based timetables by introducing a method of applying statistical analysis of the relationship between planned headways and recorded delays to estimate the minimum feasible headway between conflicting train movements in a railway system. This method is applied on the busiest railway line in Denmark and the results from recorded operations are validated through microsimulation.

Keywords
Railway Delays, Headways, Timetables, Data Analysis, Train detection systems

1 Introduction

The reliability and punctuality of a railway system are of utmost importance to its operators and infrastructure managers, as these factors directly influence the service delivery and service quality of the system. Both performance measures can be improved by decreasing the risk of conflict between trains in the network. One well-established method for reducing the risk of conflict in a timetable is the addition of buffers to the individual timetable components, such as running time and dwell time. Buffer time can also be added between conflicting train movements to ensure that the timetable can be operated, even in the case of moderate disruption; this is referred to as headway buffer.

Headway buffer is defined as the difference between the planned headway time and the minimum headway time, which is a function of the infrastructure, as well as, the features of the trains involved in the interaction (Goverde & Hansen, Performance indicators for railway timetables, 2013). The larger the headway buffer between trains, the lower the chance that the delay of one train will propagate to the other trains in the network (Hansen & Pachl, 2014). While buffer time increases the robustness of a system, it also increases the

capacity consumption and thus leads to a reduction in the level of service for passengers. This tradeoff between reliability and level of service is a central focus of research within railways, particularly in the fields of timetable optimization (Huisman & Boucherie, 2001; Schittenhelm, 2011; Sels, et al., 2015; Jovanović, Kecman, Bojović, & Mandić, 2017), capacity utilization (Gibson, Cooper, & Ball, 2002; Landex, 2008; Armstrong & Preston, 2017; Jensen, Landex, Nielsen, Kroon, & Schmidt, 2017) and delay propagation modeling (Hofman, Madsen, Groth, Clausen, & Larsen, 2006; Şahin, 2017; Zieger, Weik, & Nießen, 2018).

Many of the models presented or applied in these fields of research emphasize the importance of minimum headway in assessing the performance of a railway timetable and identifying the optimal buffer times that should be used in the planning of these timetables. Although it is included as an input parameter in all the referenced models, the minimum headway was either left as a theoretical concept or was applied as a generalized value without reference to its validation.

In their simulation model for testing timetable robustness and recovery strategies on the DSB S-train, Hofman et al. (2006) applied a general value of 1,5 minutes for the minimum headway between all trains at all locations in the network. However, they admitted that this generalization decreased the precision of the model and that it could be improved by applying actual, verified minimum headways values. Zieger et al. (2018), who used Monte-Carlo simulation to model delay propagation, explained that the minimum headway is dependent on the train type and infrastructure, and asserted that it is the responsibility of the infrastructure manager to identify this parameter to ensure that all timetables are planned with respect to it.

While a realistic estimation of the minimum feasible headway is proven to be essential for the design of robust timetables with adequate buffers to absorb the most common disturbances, it is still common practice in railway planning for practitioners to design planned headways based on experience and rule-of-thumb estimations at an aggregated line level and without consideration of the actual conflicts at the block-section level (Andersson, Peterson, & Törnquist Krasemann, 2011; Palmqvist, Olsson, & Hiselius, 2018). A poor estimation of the minimum headway time leads to infeasible timetables and sequences of trains with a negative headway buffer and thus, an increase in the delay across consecutive trains.

In this paper, the relationship between the planned headways separating conflicting movements and the change in delay of the second train involved in the conflict is investigated. Historical data recorded by the signaling system and the automatic train detection system is deployed to estimate the minimum feasible headway between conflicting movements. These values could then be used as input to models or calculation methods that assist in the designing and planning of optimal railway timetables.

The following section includes a review of the relevant literature. Section 2 introduces the methodology that is applied in this research and presents the developed method for deriving the minimum headway from the distribution of planned headway and change in delay. Section 3 applies these methods to a case study on a Danish railway line; the results are presented and their significance is discussed. Finally, a conclusion is given in Section 4.

1.1. Literature survey

Headway times, and particularly minimum headway times, serve as input parameters to the models of delay generation and propagation found in the literature. However, there is a smaller set of research studies that have used empirical data to focus specifically on the

relationship between realized delay and planned headway.

The relationship between delay and headway was studied by Landex (2008) by identifying a delay propagation factor as a function of capacity consumption and an initial delay value, given in terms of the minimum headway. The author asserted that the planned headway, along with the minimum headway and the initial delay, could be used to estimate the realized secondary delay but did not explore this assertion further. Haith et al. (2014) validated this assertion and concluded that planned headway values increase the precision of finding and assessing the reactionary delays in a system in comparison to using a compression method to assess capacity usage and the corresponding realized delay.

Hansen (2004) modelled the stochastic nature of realized block occupation by analysing the distributions of the realized time registrations of trains in relation to their planned values. The author then asserted that these findings could be used to determine the optimal planned headway since it assured that there was an acceptable probability that conflicts would be avoided. This analysis focused on the planned headway at the line level, rather than at the detailed signal level.

Daamen et al. (2009) developed a conflict identification tool that uses detailed historical operations data, including signal aspect data, as input to the model. Goverde & Meng (2011) extended the usability of this tool by introducing a statistical analysis tool that automatically identifies secondary delays based on the identification of route conflict chains. The focus of this research was to provide a method for identifying the signals in the system with the greatest number of conflicts or largest changes in delay in order to identify systemic bottlenecks.

Richter (2012) had a similar research goal and used an aggregated dataset of detailed signal aspect records to study the source of train delays on both the train level and the signal level. The authors investigated the change in delay between consecutive trains, but only connected this to the planned headway through visual inspection. A similar method was applied by van Oort et al. (2015), who assessed the service quality on a bus line through visual comparison of the realized headways and realized delays at each stopping location on the line. A value for the minimum headway could have been estimated through this visualization technique, but it is not sufficient for clarifying its direct relationship to delay, nor does it include the relationship between the planned headway and the realized delay.

Corman & Kecman (2018) assessed the relationship between the planned headway between two consecutive trains and the change in delay of the second train, in the case that at least one of the trains was a freight train. They used visual inspection to assert that, in general, large changes in delay correspond to shorter planned headway times. The authors also took this investigation one step further and used regression analysis to conclude that the change in delay for this subset of trains could not be explained statistically by the planned headway.

Minimum headway and its direct relationship to delay was investigated by Yabuki et al. (2015) in their assessment of the effectiveness of a delay reduction measure applied on a metro line. This delay reduction measure involved upgrading the signalling system to enable a decrease in the minimum headway on the line, and therefore, an increase in the buffer time when the planned headway is unchanged. The authors analysed empirical data by association rules and concluded that reducing the minimum headway was successful in reducing the level of delays in the network. However, they did not extend their research to include the derivation of the minimum feasible headway time inherent in the system.

There is agreement throughout the literature on the importance of understanding the relationships between minimum feasible headway, planned headway and realized delay. There is also a clear need for the derivation of accurate values of minimum headway to be

used as input for models of timetable optimization, capacity utilization and delay propagation. This research focuses on the relationship between planned headway and realized secondary delays; it expands the usefulness of this relationship by identifying a method for applying statistical analysis to derive the minimum feasible headway inherent in a railway system. In addition to the derivation of the minimum headway from standardly accessible historical operations data, the second major contribution of this work is the focus on specific conflicting movements, rather than on conflicting train paths at the line level.

2 Identification of the minimum feasible headway

Headway times in railway planning describe the time separation between conflicting train movements at a specified location. The planned headways can be considered as the summation of two main components. The first is the minimum feasible headway, which describes the technical time necessary for the itinerary reset after a train passes and for the transfer of movement authority to the second train. The second part is commonly referred to as headway buffer, and it is used to reduce the interferences between train movements in case of small disturbances (Hansen, 2004). This relationship is described in (1), with h_i being the planned headway between trains i and $i-1$, $h_{i_{min}}$ being the minimum feasible headway, and b_i being the headway buffer.

$$h_i = h_{i_{min}} + b_i. \tag{1}$$

When the planned headway between conflicting movements of two trains is equal to the minimum feasible headway, any delay of the first train will be transferred and result in a delay of the second train at least equal to the delay of the first. This delay can only be recovered if there is a buffer in the planned headway between the trains. In this case, the delay of the second train is greater than or equal to the delay of the first train minus the planned headway buffer. The headway buffer represents, thus, the upper limit in the delay recovery between consecutive trains at a specified location. This relationship is explained by the equations below:

$$d_i \geq d_{i-1} - b_i \tag{2}$$

$$\Delta d_i := d_i - d_{i-1} \geq h_{i_{min}} - h_i, \tag{3}$$

where d_i is the delay measured for train i at a timing point, and Δd_i is the difference in delay measured between consecutive trains. Note that the relations are valid both for positive and negative deviations from the schedule, respectively delays and earliness, as the minimum headway between conflicting movements is independent from the timetable. From (1), the minimum feasible headway corresponds to a value of planned headway that contains no buffer and therefore allows for no recovery between consecutive trains.

Railway schedules are often characterized by few discrete values of planned headway, due to the rounding to entire minutes in the public timetables (Hansen, 2004). The continuous domain of (1) becomes thus discrete, and the distributions of realized changes in delays can be analyzed as conditional to the individual values of planned headway. The minimum feasible changes in deviations from the schedule still lie on the straight line defined in (1), as depicted in Figure 1.

In this paper, the relationship between the planned headways and the change in deviation between consecutive trains is investigated through historical data recorded by the signaling system and the automatic train detection system. The timestamps of all the trains operated at one location are compared to the schedule to identify the deviations. The time differences between the scheduled times of consecutive trains represent the planned headway. The

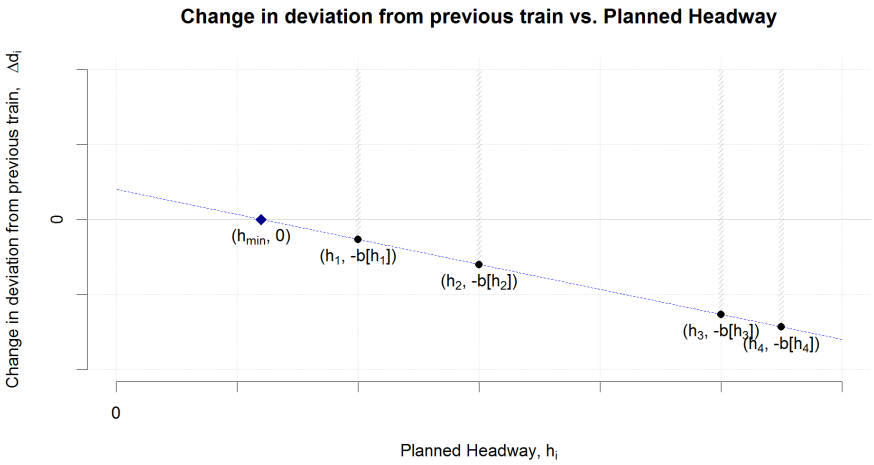


Figure 1: Relationship between planned headway and change in deviation between trains with discrete values on planned headways.

change in deviation between consecutive trains is then compared to the respective planned headway. For a given value of planned headway, the minimum change in deviation recorded between trains identifies a lower boundary to the buffer as it expresses the maximum recorded recovery between consecutive trains (cf. (2), (3)). The regression of the minimum changes in deviation against the planned headways returns the linear relationship between the headway buffer and the planned headway. The minimum headway between conflicting movements can be calculated, then, as the value of planned headway that gives zero buffer.

The analysis of historical records can be disaggregated by different factors with a potential influence on the minimum feasible headway. Examples are the train length and dynamic performance, the train category, and the speed profile of the conflicting itineraries. In the following section, the method described above finds application on a Danish case.

3 A Danish case: the West Line

The *Vestbane* (West Line) is a premarily double tracked railway in the Copenhagen region. This is the the busiest railway line in the Danish railway network of Banedanmark, and it is operated by a manifold traffic: regional, intercity, and international passenger trains, as well as domestic and international freight trains. The passenger service is typically operated from the central station in Copenhagen (KH) to Høje Tåstrup (HTÅ) and beyond, whereas the typical route for freight trains originates from Malmø (Sweden) through the Øresund bridge and reaches the Vestbane at the junction in Hvidovre. Figure 2 depicts the line scheme with the train detection points. Only the westbound tracks are reported as the analysis only includes trains in this direction.

At Copenhagen central station, four platform tracks are connected to the Vestbane, but these tracks all share the same timing point, located just beyond the junction. On the contrary, the two westbound tracks in Høje Taastrup are provided with individual timing points, as the line continues as four-tracked up to Roskilde.

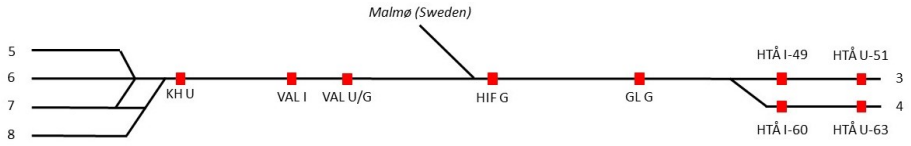


Figure 2: The Vestbane line scheme. Westbound track. The timing point locations are reported in red.

Table 1: Station codes and names on the Vestbane.

Station code	Station name	Distance from KH [km]	Type
KH	Copenhagen central station	0,0	Passenger Station
VAL	Valby	3,9	Halt
HIF	Hvidovre Fjern	7,3	Junction
GL	Glostrup	11,2	Technical station
HTÅ	Høje Taastrup	19,5	Passenger Station

In the resulting charts, the stations are identified by a code specified by the infrastructure manager. The station codes and names are reported in Table 1.

The set of timestamps included in the analysis state the scheduled and realized times of the trains at every timing point on the Vestbane during the period from August to December 2018, as this is the most recent long period without major modifications to the timetable. The daily timeframe of the records spans from 5AM to 8PM to exclude the influences of track possession for routine works and the consequent traffic modifications. A total of 118.965 records were collected and analyzed between Copenhagen central station and Høje Taastrup. The records include information about the operations and the timing points, such as the station name, track section ID, train ID, train category, scheduled time, and recorded deviation. The data is generated by Banedanmark's automatic train detection system, which uses the sensors from the interlockings and the signaling system components. Typically, the track circuit boundaries do not correspond exactly to the platforms and an offset is generated between the time recorded by the automatic system and the actual time a train arrives at or departs from the platform. A correction factor was calculated by Banedanmark using statistical analyses of GPS positions of train trajectories in collaboration with the main rail operator, DSB (Richter, Landex, & Andersen, 2013). The recorded timestamps are, therefore, an approximation of the real platform times.

The timestamps are divided into three types, which describe the associated types of movement. *"I"* records indicate the arrival times at the stations (*Indkørsel*, Entrance), whereas *"U"* records indicate the departure times (*Udkørsel*, Exit). *"G"* records indicate the pass-through time in case of non-stopping trains (*Gennemkørsel*, pass-through) and are measured at the same locations as the *"U"* records.

The planned headways and changes in deviation across consecutive trains were calculated from the timestamps by means of the free software R 3.5.1 by the R Foundation for Statistical Computing. For every timing point, the conflicting movements of interest were identified in terms of track ID and type of records (I, U, or G).

The relationship between the planned headway and the realized change in deviation was explored on a subset of the records, which only included passenger trains operated in the scheduled order. Freight and empty trains were excluded as there are fewer timestamps for these trains and they are characterized by larger variations in the recorded deviations

(Corman & Kecman, 2018). The dataset was further filtered according to the sequences of trains, as the comparison between planned headway and realized change in deviation, in fact, is only valid if the realized sequence of trains corresponds to the plan.

From (2), the minimum recorded change in deviation constitutes a lower boundary for the actual headway buffer and does not necessarily correspond to its magnitude. For this reason, only a subset of the recorded minimum changes in deviation as a function of the planned headway can be considered in the regression to the headway buffer. As a starting point, the selection of the valid points is based on the number of observations recorded for each value of planned headway. The underlying assumption is that, for a large enough sample of observations of train sequences planned with a given headway, there finds at least one case of full recovery. In such cases, the full buffer contributed in the reduction of delay propagation and the delay of the second train of the pair was reduced by exactly an amount corresponding to the headway buffer. In this study, the selection of the valid points was based on the number of observations as a percentage of the total number of observations in the complete dataset. The percentage was defined for individual headway studies.

3.1. Results

Two representative graphs are reported in this article, as a result of the analysis of the Danish case. Figure 3 and Figure 4 show the relationship between recorded changes in deviation and planned headways.

The minimum feasible headways were calculated for the main conflicting movements on the line and compared to the minimum feasible headway times measured through microsimulation. The results are reported in Table 2.

The simulation tests were operated in the commercial software RailSys 10.3.322, by Rail Management Consultants GmbH.

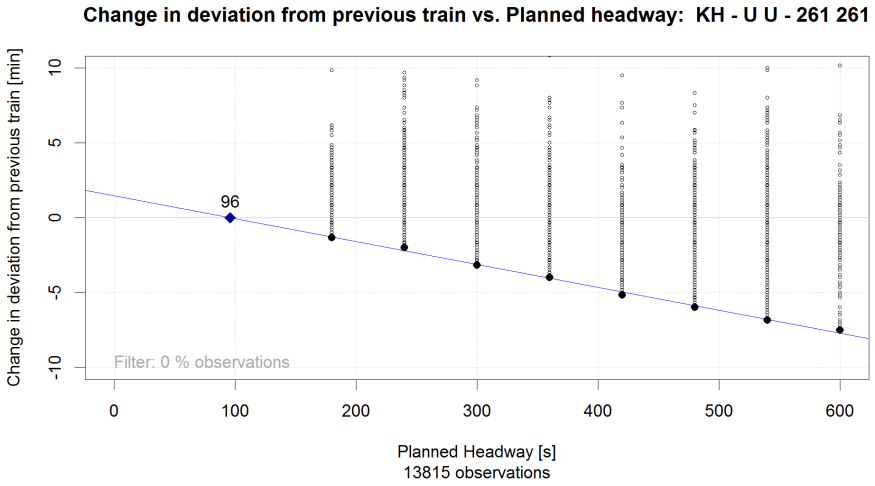


Figure 3: Change in deviation in relation to the planned headway for departures from Copenhagen central station. The bold dots are the minimum changes in deviation recorded for given planned headways. The blue line is the regression line of the headway buffer as a function of the planned headway. The diamond is the calculated minimum feasible headway.

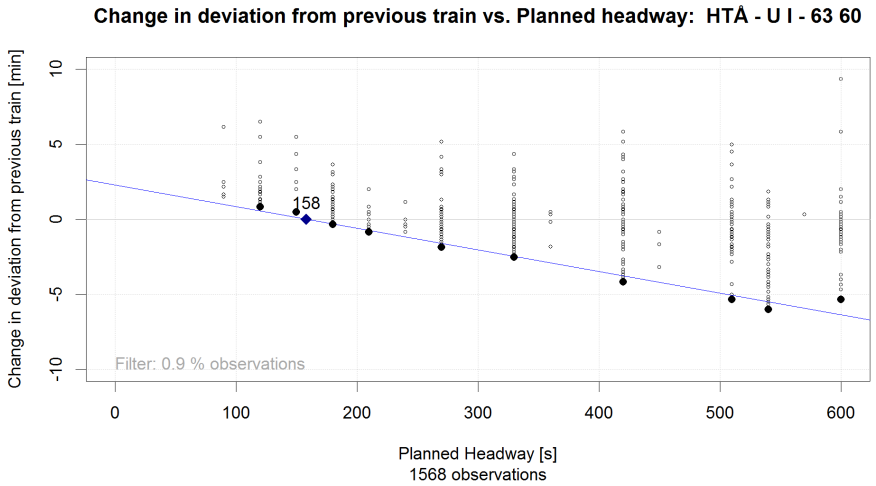


Figure 4: Change in deviation in relation to the planned headway for sequences of exits and entrances at Høje Taastrup station, track 4. The bold dots are the minimum changes in deviation recorded for given planned headways. The blue line is the regression line of the headway buffer as a function of the planned headway. The diamond is the calculated minimum feasible headway.

Table 2: Results from historical data analysis compared to microsimulation.

Station	Registr. pattern	Section ID 1	Section ID 2	Track no.	h_{min} [s]		Diff. [s]
					Hist. data	Microsim.	
KH	UU	261	261	5/6/7/8	96	94	-2
VAL	GI	2042	2033	2	118	113	-5
VAL	GU	2042	2042	2	142	139	-3
VAL	UG	2042	2042	2	101	116	15
VAL	UU	2042	2042	2	150	164	14
HIF	GG	452	452	2	64	82	18
HTÅ	UI	51	49	3	176	148	-28
HTÅ	UI	63	60	4	158	148	-10
HTÅ	UU	51	51	3	154	211	57
HTÅ	UU	63	63	4	234	211	-23
HTÅ	II	49	49	3	243	211	-32
HTÅ	II	60	60	4	236	211	-25
HTÅ	II	49	60	3-4	81	102	21
HTÅ	II	60	49	4-3	79	102	23

3.1. Discussion

Table 2 shows limited differences between the analysis of historical data and the microsimulation of minimum feasible headways. In general, the deviation between the two methods lies within a [-30, +30] s interval, apart from records at HTÅ, track 3. This specific case is affected by few outliers, possibly inaccurate time measures, shown in Figure 5. In particular, the estimated minimum feasible departure time at HTÅ track 3 seems infeasible, highlighting the necessity for further investigation.

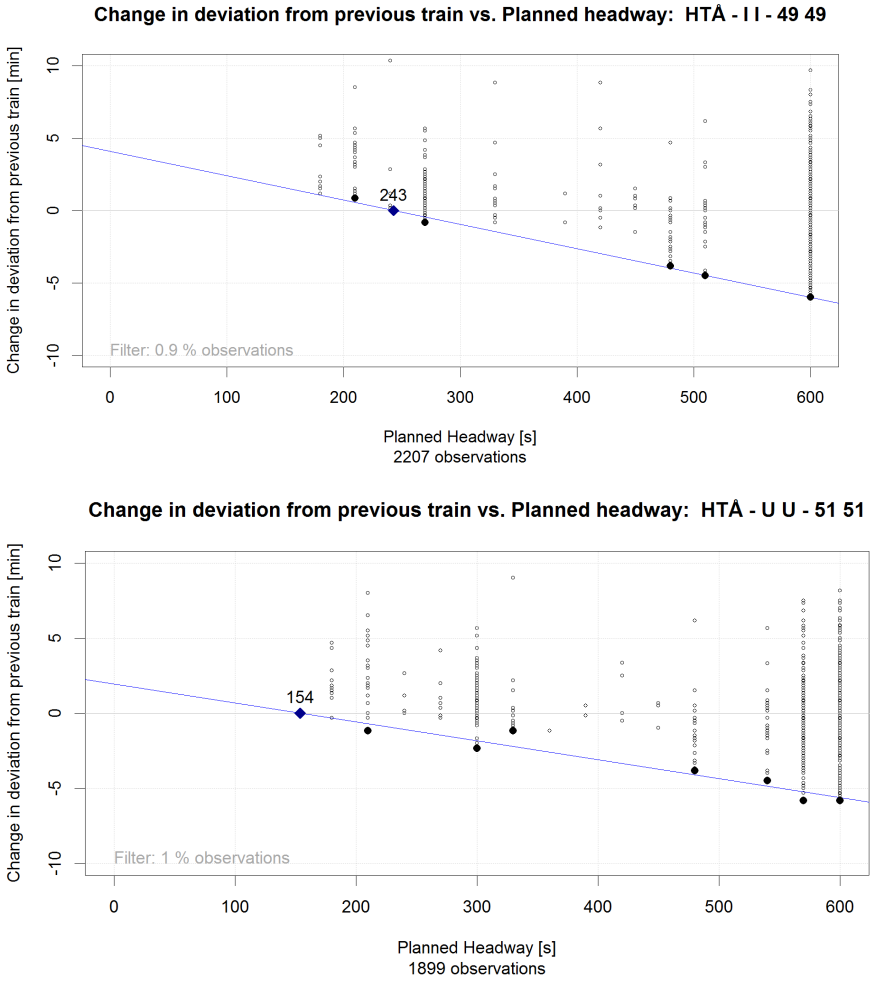


Figure 5: Minimum feasible headways at HTÅ, track 3. Arrival headways on the left, departure headways on the right.

In the other cases, the deviation between the two estimation methods finds partial explanation in the different granularity of the measuring systems. On the one hand, while it is possible to measure passing times with a second-precision in RailSys, the current time granularity for the trackside measurements on the Danish rail network is 10s. On the other hand, the microsimulation results depend on the quality of the modeling assumptions, including a deterministic minimum dwell time, and approximated driving behaviors.

The presence of resulting negative buffers at HTÅ, visible in Figure 5, is noteworthy. At this station, a 4-tracked line section starts to fork into two lines at Roskilde, about 10 km beyond HTÅ. The minimum feasible headway between movements operated on the same track is clearly larger than movements occupying different tracks. The planned headway

between trains originally scheduled on different tracks is smaller than the minimum feasible headway between movements operated on the same track. This is the case for points registered on the left side of the minimum feasible headway in Figure 5, left side. This results in a positive change in deviation, namely a secondary delay.

Note that some of the influencing characteristics could not be measured. For example, the railway undertakings do not have to state the length and type of rolling stock used in operation, even though it might differ from the original plan. However, microsimulation tests suggested very limited differences in the liberation time of the blocking sections among different settings of rolling stock. The most relevant factor, the stopping pattern, is taken into account by means of the record type (I, U, or G).

4 Conclusions

This paper presents a historical data-based method to estimate the actual minimum feasible headway between conflicting movements in railway systems. The relationship between planned headways and recorded delays is investigated from the train timestamps automatically generated by the signaling system. The method is applied on the busiest railway line in Denmark and the results from recorded operations are validated through microsimulation.

The identified minimum feasible headways constitute the input data for multiple applications. Timetable optimization problems, simulation models at both mesoscopic and macroscopic level, and capacity and robustness assessment methods often require the minimum feasible headway times as input. The method supports, thus, the improvement of railway schedules through a fact-based planning of the process times and buffers, as opposed to the current tradition of experience-based planning. Microsimulation models can also be calibrated and validated using the proposed method, through a systematic comparison of the minimum feasible headways measured from realized operation and from simulation. Further applications include the evaluation of the timetable reliability, as it is possible to extract the actual available headway buffer in the already planned schedules by subtracting the minimum feasible headways.

While previous methods described the relationship between headways and delay propagation from a theoretical perspective (Landex, 2008), this research presents a method based on the realized operation. Nevertheless, this method does not require detailed signal timestamps (Daamen Winnie and Goverde, 2009; Goverde & Meng, 2011; Richter T. , 2012), which simplifies the data acquisition process. The resulting minimum feasible headways clearly identify the potential conflicts in the timetables, whereas previous research based the identification of conflicts mainly on visual inspection of the delay and realized headway profiles (van Oort, Sparing, Brands, & Goverde, 2015). The found relationship between planned headway agrees with previous research (Yabuki, Ageishi, & Tomii, 2015; Corman & Kecman, 2018), even though this relationship had not been used to identify the minimum feasible headways.

The case study presented in Section 3 showed some weakness of the method against irregular data. In fact, a more sophisticated approach is under development to account for the recorded conditional distribution of changes in deviation for given values of planned headways. This will provide a method for assessing the probability that the minimum record value corresponds to the actual minimum possible change in deviation, thus providing a better selection of the regression points and returning more accurate values of the minimum feasible headways.

References

Andersson, E., Peterson, A., & Törnquist Krasemann, J. (2011). Robustness in Swedish Railway Traffic Timetables. In S. Ricci, I. Hansen, G. Longo, D. Pacciarelli, J. Rodriguez, & E. Wendler (Ed.), *Proceedings of the 4th International Seminar on Railway Operations Modelling and Analysis*, (pp. 1-18). Rome.

Armstrong, J., & Preston, J. (2017, 12 1). Capacity utilisation and performance at railway stations. *Journal of Rail Transport Planning and Management, 7*(3), 187-205.

Corman, F., & Kecman, P. (2018). Stochastic prediction of train delays in real-time using Bayesian networks. *Transportation Research Part C: Emerging Technologies*.

Daamen Winnie and Goverde, R. (2009, 3). Non-Discriminatory Automatic Registration of Knock-On Train Delays. *Networks and Spatial Economics, 9*(1), 47-61.

Gibson, S., Cooper, G., & Ball, B. (2002). Developments in transport policy: The evolution of capacity charges on the UK rail network. *Journal of Transport Economics and Policy, 36*(2), 341-354.

Goverde, R., & Hansen, I. (2013). Performance indicators for railway timetables. *2013 IEEE International Conference on Intelligent Rail Transportation Proceedings* (pp. 301-306). IEEE.

Goverde, R., & Meng, L. (2011). Advanced monitoring and management information of railway operations. *Journal of Rail Transport Planning and Management, 1*(2), 69-79.

Haith, J., Johnson, D., & Nash, C. (2014, 1 2). The case for space: the measurement of capacity utilisation, its relationship with reactionary delay and the calculation of the capacity charge for the British rail network. *Transportation Planning & Technology, 37*(1), 20-37.

Hansen, I. (2004). Increase of capacity through optimised timetabling. *Advances in Transport, 15*, 529 - 538.

Hansen, I., & Pachl, J. (2014). *Railway Timetabling & Operations : Analysis, Modelling, Optimisation, Simulation, Performance Evaluation* (2nd and ex ed.). Hamburg: Eurailpress.

Hofman, M., Madsen, L., Groth, J., Clausen, J., & Larsen, J. (2006). Robustness and Recovery in Train Scheduling - a simulation study from DSB S-tog a / s. (R. Jacob, & M. Müller-Hannemann, Eds.) *6th Workshop on Algorithmic Methods and Models for Optimization of Railways (ATMOS)*, 97-118.

Huisman, T., & Boucherie, R. (2001, 3). Running times on railway sections with heterogeneous train traffic. *Transportation Research Part B: Methodological, 35*(3), 271-292.

Jensen, L., Landex, A., Nielsen, O., Kroon, L., & Schmidt, M. (2017, 1). Strategic assessment of capacity consumption in railway networks: Framework and model. *Transportation Research Part C: Emerging Technologies, 74*, 126-149.

Jovanović, P., Kecman, P., Bojović, N., & Mandić, D. (2017, 1). Optimal allocation of buffer times to increase train schedule robustness. *European Journal of Operational Research, 256*(1), 44-54.

Landex, A. (2008). *Methods to estimate railway capacity and passenger delays.* Technical University of Denmark (DTU), Transport.

Palmqvist, C.-W., Olsson, N., & Hiselius, L. (2018). The Planners' Perspective on Train Timetable Errors in Sweden. *Journal of Advanced Transportation, 2018*, 17.

Richter, T. (2012). Data aggregation for detailed analysis of train delays. In C. Brebbia, N. Tomii, J. Mera, B. Ning, & P. Tzieropoulos (Ed.), *WIT Transactions on the Built Environment. 127*, pp. 239-250. WITPress.

Richter, T., Landex, A., & Andersen, J. (2013). Precise and accurate train run data: Approximation of actual arrival and departure times. *WCRR (World Congress Railway Research).* Sydney (Australia): International Association of Railways.

Şahin, İ. (2017). Markov chain model for delay distribution in train schedules: Assessing the effectiveness of time allowances. *Journal of Rail Transport Planning and Management*.

Schittenhelm, B. (2011). Planning With Timetable Supplements in Railway Timetables. *Annual Transport Conference at Aalborg University.* Aalborg, DK: trafikdage.

Sels, P., Meisch, K., Parbo, J., Møller, T., Dewilde, T., Cattrysse, D., & Vansteenwegen, P. (2015). Towards a Better Train Timetable for Denmark Reducing Total Expected Passenger Time. *CASPT2015.* Rotterdam.

van Oort, N., Sparing, D., Brands, T., & Goverde, R. (2015). Data driven improvements in public transport: the Dutch example. *Public Transport, 7*(3), 369-389.

Yabuki, H., Ageishi, T., & Tomii, N. (2015). Mining the Cause of Delays in Urban Railways based on Association Rules. *CASPT2015*, (pp. 1-16). Rotterdam.

Zieger, S., Weik, N., & Nießen, N. (2018). The influence of buffer time distributions in delay propagation modelling of railway networks. *Journal of Rail Transport Planning and Management*.

Train Rescheduling for an Urban Rail Transit Line under Disruptions

Yihong Chang [a], Ru Niu [a], Yihui Wang [a,1], Xiaojie Luan [b] Marcella Samà [c] Andrea D'Ariano [c]

[a] State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University
Beijing 100044, P.R. China

[1] E-mail: yihui.wang@bjtu.edu.cn, Phone: +86 10 5168 8607

[b] Section Transport Engineering and Logistics, Delft University of Technology
2628 CD Delft, the Netherlands

[c] Deparment of Engineering, Roma Tre University
via della vasca navale 79,00146 Rome, Italy

Abstract

Disruptions in urban rail transit systems usually result in serious incidents due to the high density and the less flexibility. In this paper, we propose a novel mathematical model for handling a complete blockage of the double tracks for 5-10 minutes, e.g., lack of power at a station, where no train can pass this area during the disruption. Under this disruption scenario, train services may be delayed or cancelled, some rolling stock may be short-turned at the intermediate stations with either single or double crossovers. To ensure the service quality provided to passengers, the back-up rolling stock inside depots may also be put into operation depending on the consequences of the disruptions. Thus, the number of rolling stock in the depot is considered. We discuss the disruption management problem for urban rail transit systems at a macroscopic level. However, operational constraints for the turnaround operation and for the rolling stock circulation are modelled. A mixed-integer non-linear programming (MINLP) model, which can be transformed into mixed-integer linear programming (MILP) problem, is proposed to minimize the train delays and the number of cancelled train services as well as to ensure a regular service for passengers, while adhering to the departure and arrival constraints, turnaround constraints, service connection constraints, inventory constraints, and other relevant railway constraints. Existing MILP solvers, e.g. CPLEX, are adopted to obtain near-optimal solutions. Numerical experiments are conducted based on real-world data from Beijing subway line 7 to evaluate the effectiveness and efficiency of the proposed model.

Keywords

Urban rail transit, Train rescheduling, Complete blockage, Short-turn, Rolling stock circulation

1 Introduction

Urban rail transit is of crucial importance for transporting commuters and travelers in big cities due to its advantages, such as large capacity, high efficiency, and the ability to provide safe, reliable and fast service. However, with the rapid development of urban rail transit,

plenty of new technologies and new equipment have been used, which bring in many uncertain factors that affect the normal operation of urban rail transit systems. Unexpected events, such as infrastructure failures, rolling stock failures and signal malfunctions, happen frequently and have significant impacts on the operation of train services as well as the safety of passengers. When a disruption occurs, it is important that dispatchers quickly present a good solution to reschedule trains so as to recover to the planned schedule as quickly as possible and minimize the inconvenience of passengers. On the one hand, the headway of urban rail transit lines has become smaller and smaller due to the increasing passenger demand, e.g., the headway is 2 minutes during peak hours for most of the metro lines in Beijing. On the other hand, the layout, especially the station layout, of urban rail transit lines is much simpler when compared with mainline. In most of the urban rail transit lines, trains do not overtake or meet each other in general during normal operations due to the limited infrastructure (in terms of tracks and platforms) available. So the disruptions in urban rail transit systems usually cause serious consequences due to the dense traffic and the limited operation flexibility.

The real-time railway traffic management problem has attracted more and more attention in recent years. Advances in scheduling theory have made it possible to handle railway traffic management problem effectively, in which not only the adjustment of running time and dwell time is considered (Ginkel and Schöbel (2007)), but also reordering, rerouting, cancellation of trains and other measures are adopted to change the connection between trains to ensure the quality of service provided to passengers (Corman et al. (2012)). According to Clausen et al. (2010), a disruption is an event or a series of events that render the planned schedules for trains, crews, etc. infeasible. When a disruption occurs, some effective measures which can quickly help the system return to normal operation and reduce the negative impact on passengers should be taken to adjust train schedules in a safe, effective and well-organized way. Jespersen-Groth et al. (2013) split the disruption management process for passenger railway transportation as three main sub-problems: timetable adjustment, rolling stock rescheduling and crew rescheduling. For more information, we direct to the review papers (Cacchiani et al. (2014); Narayanaswami and Rangaraj (2011)).

However, most existing literatures on train rescheduling problems are based on mainline railway systems. Since extra tracks, platforms and multiple routes are available, rescheduling in mainline railway systems usually involves reordering and rerouting strategies. Ghaemi, Cats and Goverde (2017) considered a complete blockage of double tracks for several hours, a MILP model is proposed at the microscopic level to select the optimal short-turning stations and reroute for all the services to continue operating in opposite direction. Louwerse and Huisman (2014) focused on adjusting the timetable of a passenger railway system in case of major disruptions, in which both partial and complete blockage of tracks are formulated. They also investigated the trade-off between delaying and cancelling trains. Zhan et al. (2015) investigated the real-time rescheduling of railway traffic on a high speed railway line in case of a complete blockage of double tracks, in which disrupted trains do not turn around but wait at stations until the disruption ends. Main decisions, including in which stations do trains wait, in which order do they leave after the disruption, and the cancellation of trains, are optimized by a MILP model. Zhan et al. (2016) rescheduled train services on a double-track high speed railway under disruptions, in which one of the double tracks is temporarily unavailable. They assumed that the exact duration of the disruption is not known as a priori but been updated gradually, thus trains are rescheduled according to the latest information of the disruption. Alternative graph models, which combine job

shop and alternative graph techniques, are developed in a series of papers (D'Ariano et al. (2008); D'Ariano and Pranzo (2009); D'Ariano, Pranzo and Hansen (2007)) and applied in a real traffic management system ROMA (railway traffic optimization by means of alternative graphs) to resolve conflicts in recent years. In the alternative graph model, the operation of trains is regarded as jobs associated to a prescribed sequence of operations which denote the processing on block sections.

The researches with regard to the rescheduling problems for urban rail transit systems are limited. In comparison to mainline railway systems, the objectives and formulation approaches for urban rail transit systems are slightly different due to their specific characteristics. As an early literature on train rescheduling in urban rail transit systems, Eberlein et al. (1998) tried to improve the headway regulation after a disturbance by using deadheading strategy. A MIP model is constructed to determine which trains should be deadheaded and how many stations should be skipped by certain trains to shorten the average passenger waiting time. Kang et al. (2015) proposed a model to reschedule the last trains in urban rail networks after a disturbance. The objective is to minimize the running time and the dwelling time, and meanwhile to maximize the average transfer redundant time and the network accessibility, as well as to minimize the difference between the planned timetable and the rescheduled one. A genetic algorithm was developed to solve the problem. Gao, Yang and Gao (2017) proposed a mathematical optimization model to calculate real-time automatic rescheduling strategy for an urban rail line by integrating the information of fault handling. However, they just considered small faults and recovered the timetable by modifying dwelling time and running time at a macroscopic level. Xu, Li and Yang (2015) considered an incident on one track of a double-track subway line and formulated an optimization model to calculate the rescheduled timetable with the objective to minimize the total delay time of trains. Crossover tracks are considered to balance the service quality under emergent situations. Taking passengers demand in consider, Gao et al. (2016) proposed an optimization model to reschedule a metro line with an over-crowded and time-dependent passenger flow after a short disruption, in which the pure running time between consecutive stations is fixed and stop-skip strategy is presented in the model to speed up the circulation of trains. An iterative algorithm is used to solve the model.

In this paper, we focus on a complete blockage of the double tracks for 5-10 minutes, e.g., an accident happened and the operator shut down the power supply system at a station, where no train can pass this area during the disruption. Therefore, some rolling stock may be short-turned at the intermediate stations with either single or double crossovers. The rolling stock circulation is also formulated in our disruption management model, where the rolling stock performed a disrupted service can turn around at a turnaround station and take over another service in the opposite direction. To ensure the service quality provided to passengers, the back-up rolling stock inside the depot may also be put into operation depending on the consequences of the disruptions, thus the number of rolling stock in the depot is considered. A mix integer non-linear programming (MINLP) model is proposed to handle the disruption management problem, which can be transformed into mix integer linear programming (MILP) model and then solved by exciting solvers.

The remainder of this paper is organized as follows: Section 2 describes the disruption management problem considered in this paper. The MINLP model for the disruption management problem in urban rail transit systems in term of a complete blockage of the double tracks for 5-10 minutes is proposed in Section 3. In Section 4, the formulated optimization model is transformed into an MILP problem. Experimental results based on the real-world

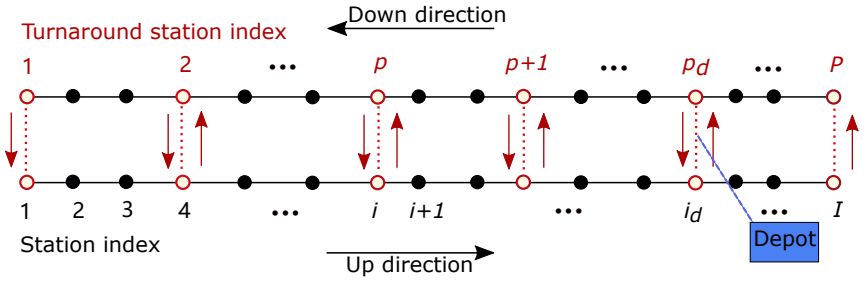


Figure 1: The layout of an urban rail transit line

data from Beijing subway line 7 are given in Section 5. The paper ends with conclusions in Section 6.

2 Problem Description

2.1 Operation of An Urban Rail Transit Line

An urban rail transit line mainly consists of stations, turnaround stations, open tracks, crossovers and depots. Figure 1 shows the layout of an urban rail transit line, which has I stations, P turnaround stations and a depot linked to turnaround station pd. One station is separated into two platforms. Open tracks are separated into two directions and each track is designed for rolling stock to operate in only one direction during normal operation but can be used in opposite direction under emergent situations. The crossovers connecting two parallel open tracks at turnaround stations can be used by rolling stock to turn around and take over another train service in the opposite direction.

This paper considers the disruption management problem for urban rail transit systems at a macroscopic level, however, the sufficient details for the turnaround operation and the rolling stock circulations are involved. In this paper, “train service” is defined as a rolling stock operating in one direction from its origin to destination. In detail, we use “service” to represent a rolling stock’s operation from station 1 to station I in the up direction or from station I to station 1 in the down direction. Once a rolling stock turns around using crossovers at turnaround stations, the corresponding “service” ends, while the rolling stock keeps circulating in the urban rail transit line. Rolling stock is stored in the depots when out of usage and the number of rolling stock in depots is limited.

2.2 Dispatching Measures

This paper considers the rescheduling problem in case of an incident of the railway infrastructure. Due to the disruption, the double tracks in a railway segment are out of order for 5-10 minutes and no train services can pass this area during the time period. The dispatching measures used to ensure the capacity of urban rail transit systems and quickly recover from the disruption include:

- Adjustments of running times and dwell times for train services;

- Rolling stock performed a disrupted service in one direction can turn around at the turnaround stations and take over another service in opposite direction;
- The back-up rolling stock inside depots can be put into operation when necessary, e.g., performing a train service that cannot be executed by the predefined rolling stock;

2.3 Assumptions

In order to formulate the disruption management model for the complete blockage scenario, we make following assumptions according to the special characteristics of urban rail transit systems:

- Rolling stock do not meet or overtake each other during operation due to the limited infrastructure (in terms of tracks and platforms) available;
- Connection between train services will change when rolling stock turning around at intermediate stations, cancelling train services and using the back-up rolling stock inside depots;
- Stopping in an interval is not allowed to avoid panicking passengers;
- Since the potential accumulation of rolling stock on the line due to the disruption, adding of new train services is not avaiable;
- Train services can depart before the departure time specified in the timetable, since the urban rail transit is more focus on the headway between train services and the passengers do not know the exact departure times;

3 Mathematical Formulation

3.1 Parameters and Variables

Parameters and decision variables adopted in the mathematical model are listed in Table 1 and Table 2 for the convenience of formulating the disruption management problem.

3.2 Objective Function

The objective function of the disruption management problem involves three parts:

- Minimize the train delay times at all visited stations;
- Minimize the deviation of the current train operations and the predefined timetable in terms of the number of cancellation services and intermediate turnaround services;
- Minimize the headway deviations between train services to ensure a regular operation and minimize passengers' waiting time;

Table 1: General subscripts, sets, input parameters

Symbol	Description
$\mathbf{I}$	set of stations, I is the last station in the line
$\mathbf{P}$	set of turnaround stations, P is the last turnaround station in the line
$\mathbf{F}$	set of train services in the up direction
$\mathbf{G}$	set of train services in the down direction
i	station index, $i \in \mathbf{I}$, i_d is the station corresponding to turnaround station p_d
p	turnaround station index, $p \in \mathbf{P}$, p_d is the turnaround station connected with depot
f	train service index in the up direction, $f \in \mathbf{F}$
g	train service index in the down direction, $g \in \mathbf{G}$
$\bar{x}^{\mathrm{up}}_{f,p,p+1}$	given binary value, $\bar{x}^{\mathrm{up}}_{f,p,p+1} = 1$ if service f in the up direction operates between turnaround station p and $p+1$ for $p \in \{1, 2, \ldots, P-1\}$ in the timetable
$\bar{y}^{\mathrm{up}}_{f,i,i+1}$	given binary value, $y^{\mathrm{up}}_{f,i,i+1} = 1$ if service f in the up direction operates between station i and $i+1$ for $i \in \{1, 2, \ldots, I-1\}$ in the timetable
$\bar{x}^{\mathrm{dn}}_{g,p,p-1}$	given binary value, $x^{\mathrm{dn}}_{g,p,p-1} = 1$ if service g in the down direction operates between turnaround station p and $p-1$ for $p \in \{2, 3, \ldots, P\}$ in the timetable
$\bar{y}^{\mathrm{dn}}_{g,i,i-1}$	given binary value, $y^{\mathrm{dn}}_{g,i,i-1} = 1$ if service g in the down direction operates between station i and $i-1$ for $i \in \{2, 3, \ldots, I\}$ in the timetable
$\bar{\beta}^{\mathrm{up}}_{f,g,p}$	binary variable, $\bar{\beta}^{\mathrm{up}}_{f,g,p} = 1$ if service f in the up direction is connected with service g in the down direction at turnaround station p in the timetable
$\bar{\beta}^{\mathrm{dn}}_{g,f,p}$	binary variable, $\bar{\beta}^{\mathrm{dn}}_{g,f,p} = 1$ if service g in the down direction is connected with service f in the up direction at turnaround station p in the timetable
$\bar{a}^{\mathrm{up}}_{f,i}/\bar{d}^{\mathrm{up}}_{f,i}$	planned arrival/departure time of service f at station i in the up direction in the timetable
$\bar{a}^{\mathrm{dn}}_{g,i}/\bar{d}^{\mathrm{dn}}_{g,i}$	planned arrival/departure time of service g at station i in the down direction in the timetable
h_{min}	minimum headway between two successive train services in the same direction in the timetable
$w^{\mathrm{up,max}}_i/w^{\mathrm{up,min}}_i$	maximum/minimum dwell time of train services at station i in the up direction
$w^{\mathrm{dn,max}}_i/w^{\mathrm{dn,min}}_i$	maximum/minimum dwell time of train services at station i in the down direction
$r^{\mathrm{up,max}}_{i,i+1}/r^{\mathrm{up,min}}_{i,i+1}$	maximum/minimum running time between station i and station $i+1$ in the up direction
$r^{\mathrm{dn,max}}_{i,i-1}/r^{\mathrm{dn,min}}_{i,i-1}$	maximum/minimum running time between station i and station $i-1$ in the down direction
$t^{\mathrm{turn,max}}_p/t^{\mathrm{turn,min}}_p$	maximum/minimum turnaround time at turnaround station p
w_{cr}	extra waiting time at turnaround stations needed to let all the passengers alight from the train
N_{p_d}	number of rolling stock in the depot before the disruption, $N_{p_d} \geq 1$
t_d	the start time point for disruption

Table 2: Decision variables

Symbol	Description
$x^{\mathrm{up}}_{f,p,p+1}$	binary variable, $x^{\mathrm{up}}_{f,p,p+1} = 1$ if service f in the up direction operates between turnaround station p and $p+1$ for $p \in \{1, 2, \ldots, P-1\}$
$y^{\mathrm{up}}_{f,i,i+1}$	binary variable, $y^{\mathrm{up}}_{f,i,i+1} = 1$ if service f in the up direction operates between station i and $i+1$ for $i \in \{1, 2, \ldots, I-1\}$
$x^{\mathrm{dn}}_{g,p,p-1}$	binary variable, $x^{\mathrm{dn}}_{g,p,p-1} = 1$ if service g in the down direction operates between turnaround station p and $p-1$ for $p \in \{2, 3, \ldots, P\}$
$y^{\mathrm{dn}}_{g,i,i-1}$	binary variable, $y^{\mathrm{dn}}_{g,i,i-1} = 1$ if service g in the down direction operates between station i and $i-1$ for $i \in \{2, 3, \ldots, I\}$
$\beta^{\mathrm{up}}_{f,g,p}$	binary variable, $\beta^{\mathrm{up}}_{f,g,p} = 1$ if service f in the up direction is connected with service g in the down direction at turnaround station p
$\beta^{\mathrm{dn}}_{g,f,p}$	binary variable, $\beta^{\mathrm{dn}}_{g,f,p} = 1$ if service g in the down direction is connected with service f in the up direction at turnaround station p
$a^{\mathrm{up}}_{f,i}/d^{\mathrm{up}}_{f,i}$	arrival/departure time of service f at station i in the up direction
$a^{\mathrm{dn}}_{g,i}/d^{\mathrm{dn}}_{g,i}$	arrival/departure time of service g at station i in the down direction
$w^{\mathrm{up}}_{f,i}$	dwell time of service f at station i in the up direction
$w^{\mathrm{dn}}_{g,i}$	dwell time of service g at station i in the down direction
$r^{\mathrm{up}}_{f,i,i+1}$	running time of service f between station i and station $i+1$ in the up direction
$r^{\mathrm{dn}}_{g,i,i-1}$	running time of service g between station i and station $i-1$ in the down direction
$t^{\mathrm{turn}}_{f,p}/t^{\mathrm{turn}}_{g,p}$	turnaround time of service f/g at turnaround station p
$\alpha^{\mathrm{up}}_{f,p_d}$	binary variable,$\alpha^{\mathrm{up}}_{f,p_d} = 1$ if the rolling stock performing service f in the up direction go back to the depot at turnaround station p_d
$\alpha^{\mathrm{dn}}_{g,p_d}$	binary variable,$\alpha^{\mathrm{down}}_{g,p_d} = 1$ if the rolling stock performing service g in the down direction go back to the depot at turnaround station p_d
$\theta^{\mathrm{up}}_{f,p_d}$	binary variable,$\theta^{\mathrm{up}}_{f,p_d} = 1$ if the rolling stock performing service f in the up direction come out from the depot at turnaround station p_d
$\theta^{\mathrm{dn}}_{g,p_d}$	binary variable,$\theta^{\mathrm{down}}_{g,p_d} = 1$ if the rolling stock performing service g in the down direction come out from the depot at turnaround station p_d
$N^{\mathrm{in}}_{f,p_d}/N^{\mathrm{in}}_{g,p_d}$	total number of rolling stock going back to depot before the departure of train service f/g at turnaround station p_d
$N^{\mathrm{out}}_{f,p_d}/N^{\mathrm{out}}_{g,p_d}$	total number of rolling stock coming out from depot before the departure of train service f/g at turnaround station p_d

Thus, the objective function can be formulated as

$$
\begin{aligned}
Z = \min \Bigg(& w_1 * \Bigg(\sum_{f \in \mathbf{F}} \sum_{i \in \mathbf{I}, i \neq 1} y^{\mathrm{up}}_{f,i-1,i} \Big(\max \left(0, (d^{\mathrm{up}}_{f,i} - \bar{d}^{\mathrm{up}}_{f,i})\right) \Big) \\
& \quad + \sum_{g \in \mathbf{G}} \sum_{i \in \mathbf{I}, i \neq I} y^{\mathrm{dn}}_{g,i+1,i} \Big(\max \left(0, (d^{\mathrm{dn}}_{g,i} - \bar{d}^{\mathrm{dn}}_{g,i})\right) \Big) \Bigg) \\
& + w_2 * \Big(\sum_{f \in \mathbf{F}} \sum_{p \in \mathbf{P}, p \neq P} \left(\bar{x}^{\mathrm{up}}_{f,p,p+1} - x^{\mathrm{up}}_{f,p,p+1}\right) + \sum_{g \in \mathbf{G}} \sum_{p \in \mathbf{P}, p \neq 1} \left(\bar{x}^{\mathrm{dn}}_{g,p,p-1} - x^{\mathrm{dn}}_{g,p,p-1}\right) \Big) \\
& + w_3 * \Big(\sum_{f \in \mathbf{F}, f \neq 1, f \neq F} \sum_{i \in \mathbf{I}, i \neq 1} \left(y^{\mathrm{up}}_{f-1,i-1,i} y^{\mathrm{up}}_{f,i-1,i} y^{\mathrm{up}}_{f+1,i-1,i} (d^{\mathrm{up}}_{f+1,i} + d^{\mathrm{up}}_{f-1,i} - 2d^{\mathrm{up}}_{f,i})\right) \\
& \quad + \sum_{g \in \mathbf{G}, g \neq 1, g \neq G} \sum_{i \in \mathbf{I}, i \neq I} \left(y^{\mathrm{dn}}_{g-1,i+1,i} y^{\mathrm{dn}}_{g,i+1,i} y^{\mathrm{dn}}_{g+1,i+1,i} (d^{\mathrm{dn}}_{g+1,i} + d^{\mathrm{dn}}_{g-1,i} - 2d^{\mathrm{dn}}_{g,i})\right) \Big) \Bigg)
\end{aligned} \tag{1}
$$

3.3 Operational Constraints

Departure and Arrival Times

As shown in Figure 2, in the disruption scenario considered in this paper, train service f in up direction can operate continuously to the next station or turn around to connect with train service g in down direction at station i (corresponding to turnaround station p). Thus, the calculation of departure times can be analysed into two cases according to the layout of station i:

- Normal Stations
  In this case, service f can only depart from station i and operate to station $i+1$, the departure time of service f at station i can be calculated by

$$d^{\mathrm{up}}_{f,i} = y^{\mathrm{up}}_{f,i-1,i}(a^{\mathrm{up}}_{f,i} + w^{\mathrm{up}}_{f,i}), \forall f \in \mathbf{F}, i \in \{2, 3, \ldots, I\}, \tag{2}$$

  where $w^{\mathrm{up}}_{f,i}$ denote the dwell time of service f at station i, which satisfies the following constraint

$$w^{\mathrm{up,min}}_{i} \leq w^{\mathrm{up}}_{f,i} \leq w^{\mathrm{up,max}}_{i}, \forall f \in \mathbf{F}, i \in \mathbf{I}. \tag{3}$$

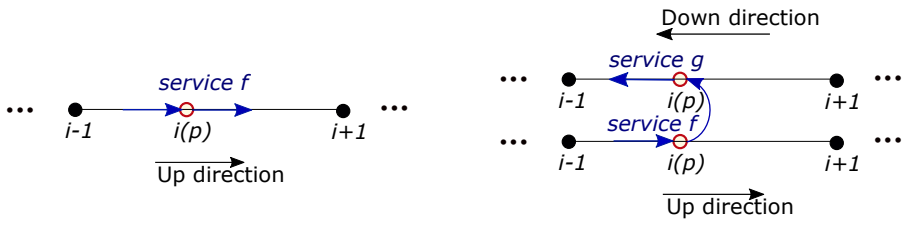


Figure 2: Departure options of train service f at station i

- Turnaround Stations
  If service f in the up direction turns around at station i (corresponding to turnaround station p) and connects with service g in the down direction, i.e., $\beta^{\mathrm{up}}_{f,g,p} = 1$, then we have

$$d^{\mathrm{up}}_{f,i} = y^{\mathrm{up}}_{f,i-1,i}(a^{\mathrm{up}}_{f,i} + w^{\mathrm{up}}_{f,i} + \beta^{\mathrm{up}}_{f,g,p} w_{\mathrm{cr}}), \forall f \in \mathbf{F}, g \in \mathbf{G}, p \in \mathbf{P}, i \in \{2, 3, \ldots, I\}, \tag{4}$$

  where w_{cr} is the extra time needed to let all the passengers alight from the train.

The calculation of arrival times can also be analysed into two cases:

- Normal Stations
  The arrival time of service f at station i from station $i-1$ can be calculated by

$$a^{\mathrm{up}}_{f,i} = y^{\mathrm{up}}_{f,i-1,i}(d^{\mathrm{up}}_{f,i-1} + r^{\mathrm{up}}_{f,i-1,i}), \forall f \in \mathbf{F}, i \in \{2, 3, \ldots, I\}, \tag{5}$$

  where $r^{\mathrm{up}}_{f,i-1,i}$ denotes the running time of service f between station $i-1$ and i, which satisfies the following constraint

$$r^{\mathrm{up,min}}_{i-1,i} \leq r^{\mathrm{up}}_{f,i-1,i} \leq r^{\mathrm{up,max}}_{i-1,i}, \forall f \in \mathbf{F}, i \in \{2, 3, \ldots, I\}. \tag{6}$$

- Turnaround Stations
  If train service f is taken over by the rolling stock performed train service g in the down direction, which turns around at turnaround station i (corresponding to turnaround station p), i.e., $\beta^{\mathrm{dn}}_{g,f,p}$=1,the arrival time of service f at station i in up direction can be calculated by

$$a^{\mathrm{up}}_{f,i} = (1-y^{\mathrm{up}}_{f,i-1,i}) y^{\mathrm{dn}}_{g,i+1,i} \beta^{\mathrm{dn}}_{g,f,p} (d^{\mathrm{dn}}_{g,i} + t^{\mathrm{turn}}_{g,p}), \forall f \in \mathbf{F}, g \in \mathbf{G}, p \in \mathbf{P}, i \in \{2, 3, \ldots, I\}, \tag{7}$$

  where $t^{\mathrm{turn}}_{g,p}$ denotes the turnaround time of service g at turnaround station p, which satisfies the following constraint

$$t^{\mathrm{turn,min}}_{p} \leq t^{\mathrm{turn}}_{g,p} \leq t^{\mathrm{turn,max}}_{p}, \forall f \in \mathbf{F}, p \in \mathbf{P}. \tag{8}$$

  When combining equation (5) and equation (7), the arrival time of service f at station i in the up direction can be calculated by

$$a^{\mathrm{up}}_{f,i} = \beta^{\mathrm{dn}}_{g,f,p}(1 - y^{\mathrm{up}}_{f,i-1,i}) y^{\mathrm{dn}}_{g,i+1,i}(d^{\mathrm{dn}}_{g,i} + t^{\mathrm{turn}}_{g,p}) + (1 - \beta^{\mathrm{dn}}_{g,f,p}) y^{\mathrm{up}}_{f,i-1,i}(d^{\mathrm{up}}_{f,i-1} + r^{\mathrm{up}}_{f,i-1,i}), \\ \forall f \in \mathbf{F}, g \in \mathbf{G}, p \in \mathbf{P}, i \in \{2, 3, \ldots, I-1\}. \tag{9}$$

Similarly, the departure time and arrival time for train service g at station i can be calculated in two cases as well.

Headway Constraints

In the disruption scenario, the headway between train services should be larger than the minimum headway determined by the train control systems. Therefore, we have the headway between service $f-1$ and f

$$y^{\mathrm{up}}_{f-1,i-1,i} y^{\mathrm{up}}_{f,i-1,i}(d^{\mathrm{up}}_{f,i} - d^{\mathrm{up}}_{f-1,i}) \geq y^{\mathrm{up}}_{f-1,i-1,i} y^{\mathrm{up}}_{f,i-1,i} h_{\min}, \\ \forall f \in \{2, 3, \ldots, F\}, i \in \{2, 3, \ldots, I\}. \tag{10}$$

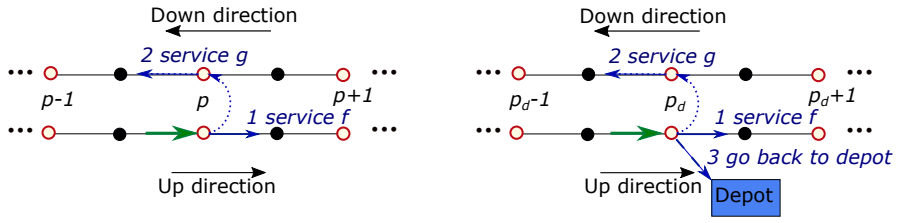


Figure 3: Departure directions of train service at turnaround stations

If $y^{\text{up}}_{f,i-1,i} = 0$ or $y^{\text{up}}_{f-1,i-1,i} = 0$ (one of the two consecutive train services was cancelled or turn around at intermediate stations), the constraint above is satisfied automatically. However, if train service f at station i is canceled, i.e., $y^{\text{up}}_{f,i-1,i} = 0$, then we need to calculate the headway using service $f+1$ and $f-1$ as follow:

$$y^{\text{up}}_{f-1,i-1,i} y^{\text{up}}_{f+1,i-1,i} (1 - y^{\text{up}}_{f,i-1,i})(d^{\text{up}}_{f+1,i} - d^{\text{up}}_{f-1,i}) \geq y^{\text{up}}_{f-1,i-1,i} y^{\text{up}}_{f+1,i-1,i} (1 - y^{\text{up}}_{f,i-1,i}) h_{\min}, \quad \forall f \in \{2,3,\ldots,F-1\}, i \in \{2,3,\ldots,I\}. \tag{11}$$

Service Connection Constraints

The rolling stock performed train service f in the up direction can turn around at turnaround stations and take over another service in the opposite direction in the disruption scenario. However, train service f can be connected with at most one train service in the down direction, i.e.,

$$\sum_{g} \sum_{p} \beta^{\text{up}}_{f,g,p} \leq 1, \forall f \in \mathbf{F}, g \in \mathbf{G}, p \in \mathbf{P}, \tag{12}$$

where $\beta^{\text{up}}_{f,g,p}$ denotes the connection between service f in the up direction and service g in the down direction.

Similarly, we have

$$\sum_{f} \sum_{p} \beta^{\text{dn}}_{g,f,p} \leq 1, \forall f \in \mathbf{F}, g \in \mathbf{G}, p \in \mathbf{P}. \tag{13}$$

to ensure train service g is connected with at most one train service in the up direction.

As shown in Figure 3, train service f in the up direction has more than one departure option at turnaround stations, especially turnaround stations with depot. Therefore, services connection constraints should be discussed separately according to different turnaround stations.

- Turnaround Stations without Depot
  In this case, service f in up direction at turnaround station p has two options: operate continuously to next station in the up direction or turn around at turnaround station p and connect to service g in the down direction. The relationship between $\beta^{\text{up}}_{f,g,p}$ and $x^{\text{up}}_{f,p,p+1}$ can be formulated as follow

$$\beta^{\text{up}}_{f,g,p} + x^{\text{up}}_{f,p,p+1} = x^{\text{up}}_{f,p-1,p}, \forall f \in \mathbf{F}, g \in \mathbf{G}, p \in \{2,3,\ldots,P-1\}. \tag{14}$$

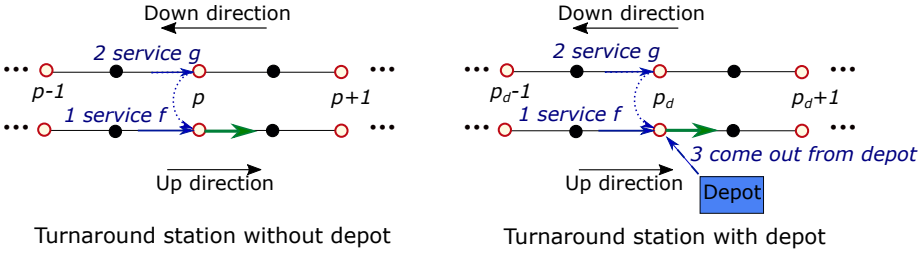


Figure 4: Sources of train service at turnaround stations

- Turnaround Stations with Depot
  Except the two options described above, service f can also go back to depot directly at turnaround station p_d which connects with depot, the equation can be proposed as

$$\beta^{\text{up}}_{f,g,p_d} + x^{\text{up}}_{f,p_d,p_d+1} + \alpha^{\text{up}}_{f,p_d} = x^{\text{up}}_{f,p_d-1,p_d}, \forall f \in \mathbf{F}, g \in \mathbf{G}, \tag{15}$$

  where $\alpha^{\text{up}}_{f,p_d}$ denotes whether service f goes back to depot at turnaround station p_d.

At the same time, train service f departs from turnaround station p in the up direction also has different sources according to the layout of turnaround stations as shown in Figure 4:

- Turnaround Stations without Depot
  In this case, service f departs from turnaround station p has two sources: come from station $p-1$ in the up direction or connect with service g in the down direction, so we have

$$\beta^{\text{dn}}_{g,f,p} + x^{\text{up}}_{f,p-1,p} = x^{\text{up}}_{f,p,p+1}, \forall f \in \mathbf{F}, g \in \mathbf{G}, p \in \{2,3,\ldots,P-1\}. \tag{16}$$

- Turnaround Stations with Depot
  Except the two sources described above, service f departs from turnaround station p_d in the up direction may also come from depot directly, the equation can be proposed as

$$\beta^{\text{dn}}_{g,f,p_d} + x^{\text{up}}_{f,p_d-1,p_d} + \theta^{\text{up}}_{f,p_d} = x^{\text{up}}_{f,p_d,p_d+1}, \forall f \in \mathbf{F}, g \in \mathbf{G}, \tag{17}$$

  where $\theta^{\text{up}}_{f,p_d}$ denotes whether service f is come out from the depot at turnaround station p_d.

Since the adding of new train services is not included in the this model, we have

$$x^{\text{up}}_{f,p,p+1} \leq \bar{x}^{\text{up}}_{f,p,p+1}, \forall f \in \mathbf{F}, p \in \{1,2,\ldots,P-1\}. \tag{18}$$

Similarly constraints about service connection of train service g in the down direction can be presented.

Inventory Constraints

For turnaround stations connected with the depot, train services can be performed by rolling stock coming out from the depot directly and the rolling stock performed a service can also go back to the depot. However, the number of back-up rolling stock inside depots for urban rail transit lines is fixed. We need to consider the availability of rolling stock when adjusting the connection between train services at turnaround stations with depot.

When a rolling stock inside the depot is required to perform train service f, i.e., $\theta^{\mathrm{up}}_{f,p_d} = 1$, the number of rolling stock going back to and coming out from the depot before train service f should satisfy inventory constraints

$$\theta^{\mathrm{up}}_{f,p_d}(N^{\mathrm{out}}_{f,p_d} - N^{\mathrm{in}}_{f,p_d}) \leq N_{p_d} - 1, \forall f \in \mathbf{F}, \tag{19}$$

where N_{p_d} is the number of rolling stock in the depot before the disruption, N^{out}_{f,p_d} and N^{in}_{f,p_d} denote the total number of rolling stock coming out from and going back to the depot before the departure of train service f at turnaround station p_d after the disruption happened, which can be calculated by

$$N^{\mathrm{out}}_{f,p_d} = \sum_{f'} \epsilon^{\mathrm{up}}_{f',p_d} \delta^{\mathrm{up}}_{f',f,p_d} \theta^{\mathrm{up}}_{f',p_d} + \sum_{g'} \epsilon^{\mathrm{dn}}_{g',p_d} \delta^{\mathrm{dn}}_{g',f,p_d} \theta^{\mathrm{dn}}_{g',p_d}, \forall f \in \mathbf{F}, f' \in \mathbf{F}, g' \in \mathbf{G}, \tag{20}$$

$$N^{\mathrm{in}}_{f,p_d} = \sum_{f'} \lambda^{\mathrm{up}}_{f',p_d} \eta^{\mathrm{up}}_{f',f,p_d} \alpha^{\mathrm{up}}_{f',p_d} + \sum_{g'} \lambda^{\mathrm{dn}}_{g',p_d} \eta^{\mathrm{dn}}_{g',f,p_d} \alpha^{\mathrm{dn}}_{g',p_d}, \forall f \in \mathbf{F}, f' \in \mathbf{F}, g' \in \mathbf{G}. \tag{21}$$

A set of binary variables is presented to describe the sequence between train services, in which $\delta^{\mathrm{up}}_{f',f,p_d} = 1$, means service f' in the up direction departs from turnaround station p_d (corresponding to station i_d) before the departure of service f, i.e.,

$$d^{\mathrm{up}}_{f,i_d} - d^{\mathrm{up}}_{f',i_d} \geq 0, \forall f \in \mathbf{F}, f' \in \mathbf{F}, \tag{22}$$

$\delta^{\mathrm{dn}}_{g',f,p_d} = 1$, means service g' in the down direction departs from turnaround station p_d before the departure of service f, i.e.,

$$d^{\mathrm{up}}_{f,i_d} - d^{\mathrm{dn}}_{g',i_d} \geq 0, \forall f \in \mathbf{F}, g' \in \mathbf{G}, \tag{23}$$

$\eta^{\mathrm{up}}_{f',f,p_d} = 1$, means service f' in the up direction arrives at turnaround station p_d before the departure of service f, i.e.,

$$d^{\mathrm{up}}_{f,i_d} - a^{\mathrm{up}}_{f',i_d} \geq 0, \forall f \in \mathbf{F}, f' \in \mathbf{F}, \tag{24}$$

$\eta^{\mathrm{dn}}_{g',f,p_d} = 1$, means service g' in the down direction arrives at turnaround station p_d before the departure of service f, i.e.,

$$d^{\mathrm{up}}_{f,i_d} - a^{\mathrm{dn}}_{g',i_d} \geq 0, \forall f \in \mathbf{F}, g' \in \mathbf{G}, \tag{25}$$

Moreover, a set of binary variables is considered to identify if the train service arrives at or depart from turnaround station p_d after the disruption happened, in which $\epsilon^{\mathrm{up}}_{f',p_d} = 1$ means service f' in the up direction departs from turnaround station p_d after the disruption happened, i.e.,

$$d^{\mathrm{up}}_{f',i_d} - t_d \geq 0, \forall f' \in \mathbf{F}, \tag{26}$$

$\epsilon^{\mathrm{dn}}_{g',p_d} = 1$, means service g' in the down direction departs from turnaround station p_d after the disruption happened, i.e.,

$$d^{\mathrm{dn}}_{g',i_d} - t_d \geq 0, \forall g' \in \mathbf{G}, \tag{27}$$

$\lambda^{\mathrm{up}}_{f',p_d}$, means service f' in the up direction arrives at turnaround station p_d after the disruption happened, i.e.,

$$a^{\mathrm{up}}_{f',i_d} - t_d \geq 0, \forall f' \in \mathbf{F}, \tag{28}$$

$\lambda^{\mathrm{dn}}_{g',p_d}$, means service g' in the down direction arrives at turnaround station p_d after the disruption happened, i.e.,

$$a^{\mathrm{dn}}_{g',i_d} - t_d \geq 0, \forall g' \in \mathbf{G}, \tag{29}$$

Similarly, when a rolling stock inside the depot is required to perform train service g, i.e., $\theta^{\mathrm{dn}}_{g,p_d} = 1$, the inventory constraints can also be proposed.

4 MILP Solution

The mixed-integer nonlinear programming (MINLP) model which is formulated in Section 3 can be transformed into a mixed-integer linear programming (MILP) problem according to the transformation properties introduced in (Bemporad et al. (1999)).

- Property I: Consider a real-valued variable $f(x)$ and a logical variable $\theta \in [0,1]$. if we let $M = f(x)_{max}$, $m = f(x)_{min}$, the product term $\theta f(x)$ can be replaced by an auxiliary real variable $z = \theta f(x)$, where $z = \theta f(x)$ is equivalent to

$$\begin{cases} z \leq M\theta, \\ z \geq m\theta, \\ z \leq f(x) - m(1-\theta), \\ z \geq f(x) - M(1-\theta). \end{cases} \tag{30}$$

- Property II: Consider two logical variables $\theta_1 \in [0,1]$ and $\theta_2 \in [0,1]$. the product term $\theta_1\theta_2$ can be replaced by a logical variables $\theta_3 \in [0,1]$, where $\theta_3 = \theta_1\theta_2$ is equivalent to

$$\begin{cases} -\theta_1 + \theta_3 \leq 0, \\ -\theta_2 + \theta_3 \leq 0, \\ \theta_1 + \theta_2 - \theta_3 \leq 1. \end{cases} \tag{31}$$

- Property III: Consider a real-valued variable $f(x) \leq 0$, and let $M = f(x)_{max}$, $m = f(x)_{min}$. If we introduce a logical variable $\theta \in [0,1]$, it can be verified that $[f(x) \leq 0] \longleftrightarrow [\theta = 1]$ is true if

$$\begin{cases} f(x) \leq M(1-\theta), \\ f(x) \geq \epsilon + (m-\epsilon)\theta. \end{cases} \tag{32}$$

Through property I the nonlinear constraints (4) and (9) can be transformed by using auxiliary real variables. Constraints (20) and (21) can be transformed by adding another logical variables according to property II. Constraints (9), (10) and (11) can be transformed by combining property I and II. The statements (22) to (29) can be transformed into logical dynamic constraints through property III.

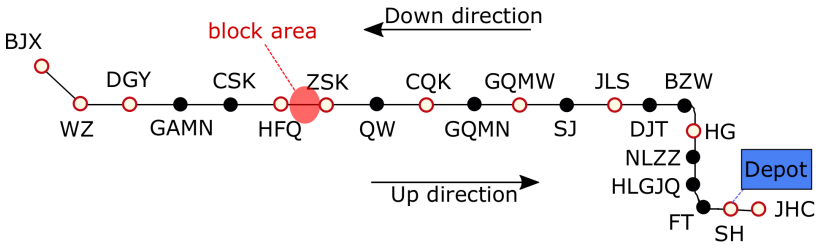


Figure 5: Layout of Beijing subway line 7

Table 3: Detailed status of train services

Number of train services	Direction	Status
f1	up	dwelling at DJT
f2	up	running from GQMN to GQMW
f3	up	running from QW to CQK
f4	up	running from CSK to HFQ
f5	up	running from DGY to GQMN
f6	up	dwelling at BJX
g1	down	running from QW to ZSK
g2	down	running from GQMN to CQK
g3	down	running from JLS to SJ
g4	down	running from HG to BZW
g5	down	running from FT to HLGJQ

5 Case Study

In this section, the experimental results of the proposed model is demonstrated based on the data from Beijing subway line 7 and IBM CPLEX 12.8 is used as the solver for the MILP problem.

The layout of Beijing subway line 7 is shown in Figure 5, which is 23.7 km long with 21 stations and one depot connected with SH station. Stations denoted by red circles are turnaround stations which provide single or double crossovers for rolling stock to turn around and take over another service in opposite direction, while stations denoted by black dots are normal stations where train services can only run directly to next station in the same direction. Train services running from BJX to JHC are in up direction while services running from JHC to BJX are in down direction. In this case study, we consider the time period from 11:00 am to 12:00 am, 10 services in each direction, which departure from its origin during this period are considered. The track blockage between HFQ and ZSK starts at 11:29 am and ends at 11:39 am, during which no trains can pass the block area. At 11:29 am, the time point which the disruption occurs, 6 services in the up direction as well as 5 services in the down direction considered in this case study are operating on the line, the detailed status are given in table 3. The maximum and minimum running times in each section are defined by adding extra 10s or reducing 10s based on the predefined timetable. The minimum dwell times at each station are defined as 20s to let passengers get on or alight from the trains while the maximum dwell times are defined by adding extra 120s in

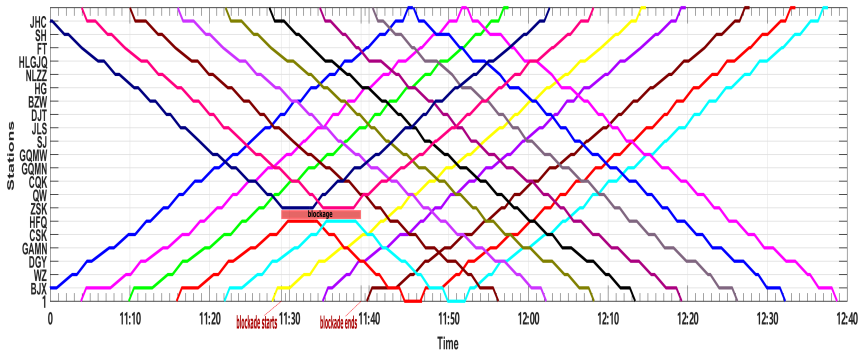


Figure 6: The rescheduled timetable

case of holding the train in station if necessary. Furthermore, the turnaround time should be between 120s and 600s. The headway of two consecutive train services should be more than 240s. The number of rolling stock in the depot is taken as 2 at the beginning of disruption. The extra waiting time at turnaround stations is 60s. The weights in the objective function are set to $w_1 = 2$, $w_2 = 100$ and $w_3 = 1$ based on several experiments.

The rescheduled timetable for train services in this disruption scenario is shown in Fig-

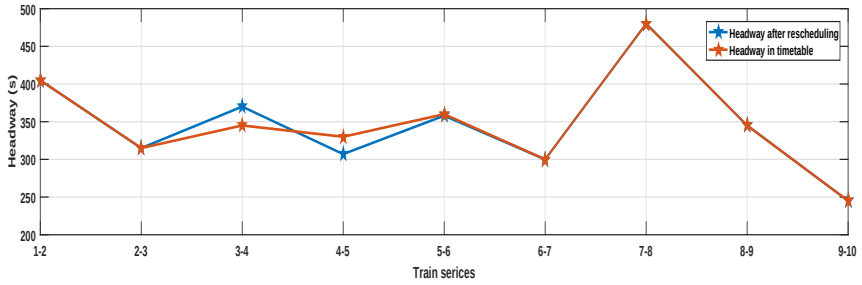


Figure 7: The headway in up direction

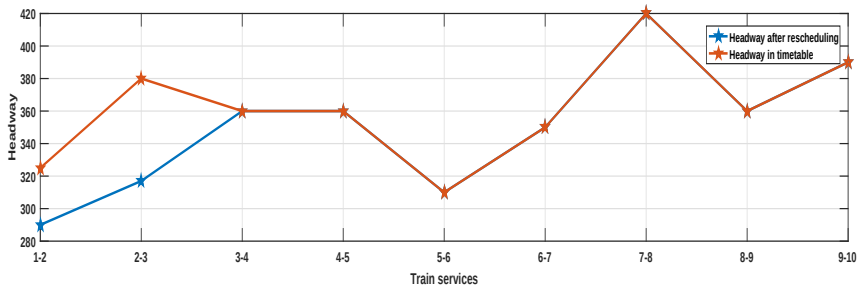


Figure 8: The headway in down direction

ure 6, in which different colors denoted train services performed be different rolling stock and the track blockage is denoted by a red rectangular inserted between HFQ and ZSK, which appears at 11:29 am and disappears at 11:39 am. It can be observed that, two train services ($f3$ and $f4$) in the up direction turn around at turnaround station HFQ and connect to train services ($g1$ and $g2$) in the down direction, accordingly, train services ($g1$ and $g2$) in the down direction turn around at turnaround station ZSK and connect to train services ($f3$ and $f4$) in the up direction without huge impact on other train services. The circulation plan of rolling stock does not change, in which train services $g8$ and $g9$ in the down direction are performed by the rolling stock which performed $f1$ and $f2$ in the up direction and turned around at JHC. The headways between train services at the station close to the block area in up and down direction are illustrated in Figure 7 and Figure 8 respectively, in which the red line denoted the predefined headway in timetable and the blue line denoted the headway after rescheduling. As can be observed in Figure 7, the headway between service $f2$ and $f3$ and the headway between service $f3$ and $f4$ are slightly changed since services $f3$ and $f4$ are disrupted and turn around before the block area, while other headways remain the same in timetable. The result is similar in down direction.

The experimental results demonstrate the effectiveness and efficiency of the proposed disruption management model. A rescheduled timetable and rolling stock circulation plan can be obtained in a few seconds, which can be used to handle disruptions so as to ensure the capacity of urban rail transit and the service quality provided to passengers.

6 Conclusion

In this paper, a disruption management model is proposed to rescheduling train services in term of a complete blockage of the double tracks for 5-10 minutes in urban rail transit systems. The objective of the model is to minimize the train delays and the number of canceled train services as well as to ensure a regular service for passengers, while constrains, such as departure and arrival constraints, turnaround constraints, service connection constraints, inventory constraints are considered. The case study based on the real-world data from Beijing subway line 7 demonstrated that an acceptable rescheduled timetable and rolling stock circulation plan can be obtained within a few seconds, which can be adopted in real-time disruption management.

References

Bemporad, A. (ed.), 1999. “Control of systems integrating logic, dynamics, and constraints”, *Automatica*, vol. 35, pp. 407–427.

Cacchiani, V. (ed.), 2014. “An overview of recovery models and algorithms for real-time railway rescheduling”, *Transportation Research Part B: Methodological*, vol. 63, pp. 15–37.

Clausen, J. (ed.), 2010. “Disruption management in the airline industry-Concepts, models and methods”, *Computers and Operations Research*, vol. 37, pp. 809–821.

Corman, F. (ed.), 2012. “Bi-objective conflict detection and resolution in railway traffic management”, *Transportation Research Part C: Emerging Technologies*, vol. 20, pp. 79–94.

D’Ariano, A. (ed.), 2008. “Reordering and local rerouting strategies to manage train traffic in real time”, *Transportation science*, vol. 42, pp. 405–419.

D'Ariano, A., Pranzo, M., 2009. "An advanced real-time train dispatching system for minimizing the propagation of delays in a dispatching area under severe disturbances", *Networks and Spatial Economics*, vol. 9, pp. 63–84.

D'Ariano, A., Pranzo, M., Hansen, I.A., 2007. "Conflict resolution and train speed coordination for solving real-time timetable perturbations", *IEEE Transactions on intelligent transportation systems*, vol. 8, pp. 208–222.

Eberlein, X.J. (ed.), 1998. "The real-time deadheading problem in transit operations control", *Transportation Research Part B: Methodological*, vol. 32, pp. 77–100.

Gao, Y. (ed.), 2016. "Rescheduling a metro line in an over-crowded situation after disruptions", *Transportation Research Part B: Methodological*, vol. 93, pp. 425–449.

Gao, Y., Yang, L., Gao, Z., 2017. "Real-time automatic rescheduling strategy for an urban rail line by integrating the information of fault handling", *Transportation Research Part C*, vol. 81, pp. 246–267.

Ghaemi, N., Cats, O., Goverde, R.M.P.,2017. "A microscopic model for optimal train short-turnings during complete blockages", *Transportation Research Part B: Methodological*, vol. 105, pp. 423–437.

Ginkel, A., Schöbel, A., 2007. "To wait or not to wait? The bicriteria delay management problem in public transportation", *Transportation Science*, vol. 41, pp. 527–538.

Jespersen-Groth, J. (ed.), 2013. "Disruption Management in Passenger Railway Transportation", In: *Robust and Online Large-Scale Optimization*, Berlin, Heidelberg.

Kang, L. (ed.), 2015. "A practical model for last train rescheduling with train delay in urban railway transit networks", *Omega*, vol. 50, pp. 29–42.

Louwerse, I., Huisman, D., 2014. "Adjusting a railway timetable in case of partial or complete blockades", *European Journal of Operational Research*, vol. 235, pp. 583–595.

Narayanaswami, S., Rangaraj, N., 2011. "Scheduling and rescheduling of railway operations: A review and expository analysis", *Technology Operation Management*, vol. 2, pp. 102–122.

Xu, X., Li, K., Yang, L., 2015. "Rescheduling subway trains by a discrete event model considering service balance performance", *Applied Mathematical Modelling*, vol. 40, pp. 1446–1466.

Zhan, S. (ed.), 2016. "A rolling horizon approach to the high speed train rescheduling problem in case of a partial segment blockage", *Transportation Research Part E: Logistics and Transportation Review*, vol. 95, pp. 32–61.

Zhan, S. (ed.), 2015. "Real-time high-speed train rescheduling in case of a complete blockage", *Transportation Research Part B: Methodological*, vol. 78, pp. 182–201.

Punctuality and Capacity in Railway Investment: A Socio-Economic Assessment for Finland

Luca Corolli [a, 1], Giorgio Medeossi [a, 2], Saara Haapala [b, 3], Jukka-Pekka Pitkänen [b, 4], Tuomo Lapp [b, 5], Aki Mankki [b, 6], Alex Landex [c]
[a] TRENOlab s.r.l.s., via Maniacco 7/A, 34170 Gorizia, Italy, Phone: +39 0481 30031
E-mail: [1] l.corolli@trenolab.com, [2] g.medeossi@trenolab.com
[b] Ramboll Finland, Espoo, Finland, E-mail: [3] saara.haapala@ramboll.fi,
[4] jukka-pekka.pitkanen@ramboll.fi, [5] tuomo.lapp@ramboll.fi, [6] aki.mankki@ramboll.fi
[c] Ramboll Denmark, Copenhagen S, Denmark, E-mail: alex.landex@ramboll.dk

Abstract
This paper presents two methods designed to provide quantitative data for analysing the socio-economic impacts of rail network improvements developed for the Finnish Transport Agency. The first is a capacity estimation method; it adapts the UIC 406 method to the characteristics of the Finnish rail network. The second method estimates delay propagation based on the key characteristics of lines; in this case distinct formulas were developed using regression for single- and double-track lines. The proposed methods were evaluated based on actual and simulated data from Finland and the UK. They provide network saturation and delay data for evaluation of capital improvements by network managers. The study results were approved and adopted by the Finnish Transport Agency.

Keywords
Capacity estimation, Delay propagation, UIC 406, Mathematical regression, Finland.

1. Introduction

The Finnish Transport Agency (FTA) requires preparation of socio-economic assessments for all major infrastructure investments. This requirement covers many types of railway projects from track rehabilitation to major network improvements. Unfortunately, there is currently no established quantitative method for assessing the capacity and traffic punctuality impacts of railway investments, and therefore they are only assessed qualitatively.

This paper presents results of research conducted for the FTA to develop quantitative methods for assessing the capacity and traffic punctuality impacts of railway investments for use in FTA's socio-economic assessments (Finnish Transport Agency, 2018). The first method assesses railway line capacity, enabling the rail network manager to determine line saturation, and thereby estimate the effect of investments on capacity. The second method evaluates delay propagation given a set of line parameters, enabling the network manager to estimate the effect of investments on train punctuality. Both methods were developed with the aim of being easy to apply by non-experts in socio-economic analyses.

This paper is organised as follows: Section 2 describes the capacity analysis method, focusing on its interpretation for Finland and results obtained by applying it to a real single-track line. Section 3 describes the delay propagation methods developed using regression for use on single- and double-track lines. Finally, Section 4 presents conclusions.

2. Capacity analysis method

The main concern in socio-economic assessments is railway network utilisation, making capacity consumption the key performance indicator. Railway capacity can be defined as the maximum throughput of a given set of trains on a specific line section or station area. Many methods have been developed to estimate railway line capacity including UIC 405 (UIC, 1996), CAPACITY (Pitkänen, 2005), and CAP1/CAP2 (Moreira et al., 2004). A basic way to calculate capacity consumption is to determinate the share of time reserved for train operations during a given time period. The result is a percentage, as shown in Equation (1):

$$\text{Cc\%} = 100 \text{ x } \frac{\text{Time reserved for train operations}}{\text{Analysed time period}} \qquad (1)$$

The most widely used method for estimating capacity consumption in Europe is UIC 406 (UIC, 2004). A key shortcoming of this method is that it does not clearly define many important parameters, leading to a wide room for interpretation (Lindner, 2011). As a result, multiple interpretations have been proposed including the UK's Capacity Utilisation Index (CUI) and Denmark's Train Mix (Landex, 2008).

An alternative method for capacity consumption estimation uses capacity indices. For example, heterogeneity indices have been developed based on the observation that heterogeneity has a clear negative correlation to disturbance tolerance (Vromans, 2005). Similarly, rail yard conflict indices have been developed based on railway layout, conflict probability, or minimum train headways (Pitkänen, 2005).

In addition to timetable-based calculation methods, capacity can also be estimated using microscopic simulation. Simulation is typically used when detailed information on the impact of various alternative infrastructure scenarios or fault situations is needed. An advantage of simulation models is that they can take human behaviour into account using stochastic parameters. A drawback is that they typically require users to define a complete microscopic model, which can be time consuming.

In the Finnish context, a study (Pitkänen, 2005) was aimed at calibrating the SBB's CAPACITY method for application in Finland. An important finding during model calibration was that results are always dependent on specific infrastructure, rolling stock and timetable assumptions, making it very difficult to study independent measures. In socio-economic assessments, these parameters frequently differ between alternatives, making comparison impossible.

An important requirement of the socio-economic assessments being considered in this research is that they should be tackled using macroscopic analysis. Therefore, microscopic methods (i.e., simulation) are not suitable. As a result, it was decided to develop an interpretation of the UIC 406 method based on characteristics of the Finnish rail network. The goal of developing a UIC 406 interpretation for Finland was to create a simple and accurate method for estimating capacity applicable to both single- and double-track lines.

2.1 UIC 406 interpretation for Finland

Developing an interpretation of UIC 406 for Finland started with Equation (1). Defining the equation denominator (the time period) is straightforward; defining the numerator (the time reserved for train operations) is more complicated.

Determining the time reserved for train operations depends on many parameters including features of the Finnish interlocking system and rolling stock. These parameters are listed and discussed in Table 1.

Parameter	Description	Notes
$\boldsymbol{K}$	Capacity consumption	Measured in percentage
$\boldsymbol{T}$	Analysed time period	Suggested value is 60 minutes
$\boldsymbol{h_A}$	Sum of minimum headway times	Sum of the time intervals between two consecutive trains running in the same direction
$\boldsymbol{t_D}$	Sum of driving time differences	Sum of time intervals between two consecutive trains running in the same direction with different driving times
$\boldsymbol{t_O}$	Sum of occupation times	Sum of time intervals between two consecutive trains running in opposite directions on a single-track line
$\boldsymbol{t_{EPD}}$	Earliest possible departure time, compared with the beginning of the time period	Time interval referring to the impact of partial trains in the beginning of the time period
$\boldsymbol{t_M}$	Sum of supplementary time for maintenance	Time that the line section is not available for normal operations due to maintenance
$\boldsymbol{t_S}$	Sum of station and crossing times	Amount of time needed for switch turning operations during the time period

Table 1. Parameters used to determinate time reserved for train operations. All time measurements are expressed in minutes.

Using the parameters listed in Table 1, Equation (1) can be expressed as:

$$K = \frac{h_A + t_D + t_O + t_{EPD} + t_M + t_S}{T} \qquad (2)$$

The first step in calculating this equation is to define the set of trains to be analysed. Next, the data must be prepared for each of the parameters. This is described below.

Definition of the set of trains to be analysed. Capacity consumption is typically calculated for hourly time periods. Trains are assigned to time periods based on the time of departure from the first station they leave in the studied area. For double-track sections, areas can span over multiple locations (i.e., stations, halts, or junctions). Single-track sections, on the other hand, are only defined between two consecutive locations.

Calculation of minimum headway times ($\boldsymbol{h_A}$). Minimum headway times depend on the driving speed and signalling. Block sections can vary by direction and therefore headway values must be calculated separately for each direction. Theoretically, the minimum headway time depends on the driving speeds of two consecutive trains. Defining:

- n block sections factor: $n = 1$ for single block sections, $n = 2$ for multiple block sections
- d average block section length, in km
- s weighted average speed, in km/h

Let us denote with h_i the minimum headway time for a train *i*, that is $h_A = \sum_i h_i$. For each train *i*, headway h_i is calculated as shown in Equation (3) (notice that 60 = mins/hour):

$$h_i = \frac{(n * d * 60)}{s} \qquad (3)$$

Calculation of running time differences (t_D). The running time difference describes the extra time needed when a slow train is followed by a faster train. This calculation, for double-track sections, depends on the operations of consecutive trains. Let us denote with t_i the additional headway to be assigned to a train *i*. If a train *i* is followed by a slower or equally fast train, there is no additional headway: $t_i = 0$. Otherwise, t_i is calculated as the difference between the running times of the two trains over the area being analysed. The total t_D value is then calculated as the sum of all t_i values, i.e. $t_D = \sum_i t_i$.

Calculation of occupation times (t_O). In this context, the term occupation time describes the reserved period after an operation on a single-track line. It is equal to the train running time on the line section being analysed.

Calculation of earliest possible departure times (t_{EPD}). This parameter is used to describe the impact of trains that only partially operate during the analysed time period, i.e. that span over multiple time periods in the studied area. In the following, we call such trains "partial trains". Four cases can be identified:

1. There is no partial train in the analysed time period: $t_{EPD} = 0$
2. There is only one partial train *t* in the scenario, departing before the beginning of the scenario and arriving at destination during the timetable period:

   $$t_{EPD} = at_i + hl - rt_{i'} - bg$$

   where:
   - at_i arrival time of train *i*
   - hl headway of the last line section
   - $rt_{i'}$ running time of first train *i'*
   - bg beginning of the time period
3. Multiple partial trains (arriving during the considered period) are present in the scenario: only the last partial train is considered.
4. There is at least one train running through the scenario, i.e. departing before and arriving after the scenario period: t_{EPD} is set to the length of the time period, resulting in full capacity consumption (100%).

Calculation of station and junction crossing times (t_S). Station and junction crossing times consist of the extra occupation time needed to account for turning a switch between two train operations. They are location-specific and should be provided by a signalling specialist. t_S is calculated as the sum of these values.

Calculation of supplementary time for maintenance (t_M). Timetables may or may not include planned capacity reservations for maintenance work or shuntings. If these operations are planned and known, they can be included in the analysis by simply adding their total duration, in minutes, to the t_M parameter.

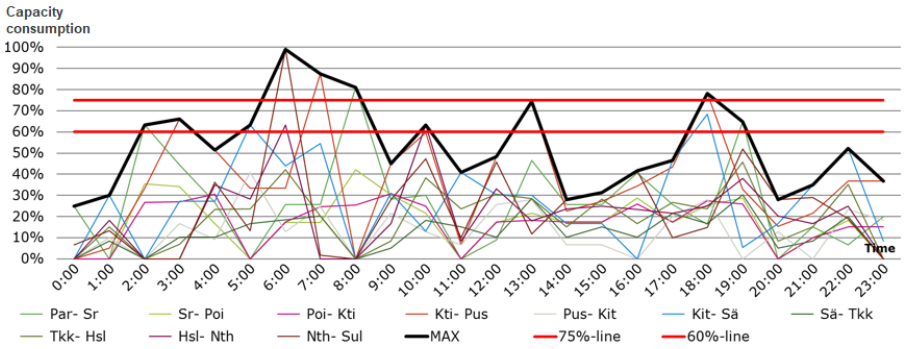


Figure 1. Capacity consumption in different single-track line sections.

2.2 Capacity consumption analysis results

Once capacity consumption values are defined for all line segments and time periods being analysed, the consumption value for the whole line can be determined. Line capacity is given by the largest value during the considered time period. For single-track sections, both directions are considered together, so the maximum value between the two directions is taken. For double-track sections, the two directions are considered separately.

When capacity consumption values are calculated over a full day, peak times typically stand out. The sharpness of these peaks gives important information on the likelihood of delays on track sections. For track sections with both passenger and freight trains, UIC has set two threshold values for congestion: 75% in peak hours, and 60% off-peak.

Figure 1 illustrates capacity consumption for different sections of a single-track line with mixed operations in Finland. Each line depicts the variation in capacity consumption over the whole day for a particular line segment. The thick black line highlights maximum values, while the two red straight lines indicate the UIC threshold values. As shown in Figure 1, during peak hours there is congestion in multiple areas, with capacity consumption remaining above 75% for three hours. This indicates a high risk of unpunctuality and little room for effective delay recovery. Similarly, the 60% off-peak threshold is exceeded three times.

3. Delay propagation method

The second method developed to better quantify socio-economic assessments of railway investments was a delay propagation method. This method calculates the relationship between capacity-related parameters and delays. Several well-accepted methods using capacity to calculate delays are already available for double-track lines (Landex, 2008). Conversely, for single-track lines, no direct relationship can be consistently identified following the theoretical evidences first identified by Potthoff (Potthoff, 1962). As a result, single- and double-track lines must be analysed separately using ad-hoc methods. These methods require large sets of data with a wide range of capacity usage. Such data can be obtained by either analysing operational data from several lines (including some heavily used lines) or using simulation (which allows testing several increasing traffic density scenarios and analysing the corresponding simulated delays). The next two sections describe the distinct methods for analysing delay on single- and double-track lines.

3.1 Single-track line delay propagation method

Based on theoretical considerations and an analysis of actual Finnish data, the delay propagation P for a group of trains on a single-track line can be defined as a function of:

- the number of crossings N_x in the timetable
- the margin t_m (it is a function of the running time: 10% for passenger trains, and 12.5% for freight trains)
- the initial delay $t_{d,i}$, with early- and late-running trains accounted for separately: $t_{d,i}^{+}$ and $t_{d,i}^{-}$. Notice that early arriving trains are considered as trains with negative initial delay, i.e. they contribute to the $t_{d,i}^{-}$ parameter.

Since initial delay is the delay given as input and final delay is delay given as output, in the following text we call them "input delay" and "output delay", respectively. Both input and output delays include all delays regardless of the cause of the delay. For input delays this is not an issue since infrastructure investments can only affect delays that propagate in the track section affected by the investment. For output delays, days with heaviest delay propagation within each line need to be filtered out of the data set since they include major train or infrastructure failures, which are not related to railway investments.

As part of this research, one year's worth of input data were aggregated by day and line. These data were supplemented by simulation data since historical data do not cover all possible parameter combinations. The simulations were run using OpenTrack software (Nash and Huerlimann, 2004) on timetables with 12 different numbers of crossings per train (each corresponding to a specific headway value) and 5 different input delay values. This showed how delays changed altering one parameter at a time. One hundred simulations were run for each combination of crossings and input delay, for a total of $12 \cdot 5 \cdot 100 = 6{,}000$ simulations.

The simulation results are illustrated in Figure 2 which shows the relationship between input delay (x-axis) and output delay (y-axis). The lines show the output delay variation for a given headway (in seconds), while the vertical bars show the average output delay across all headway values. As shown in Figure 2 the relationship between output and input delay appears to be slightly super-linear.

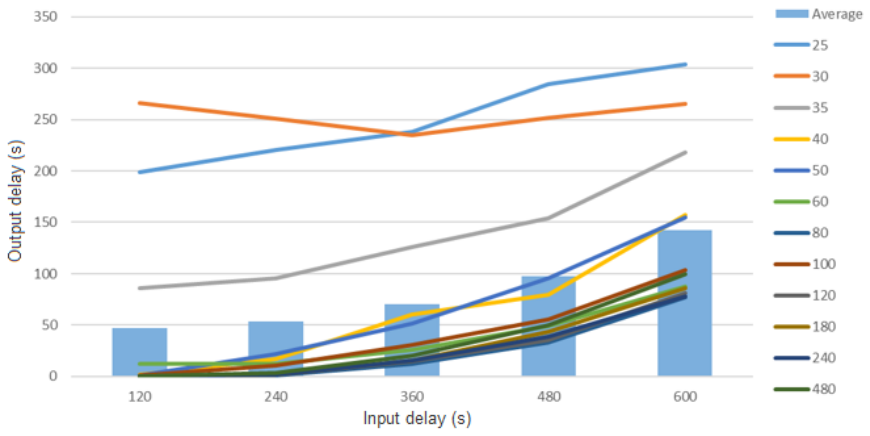


Figure 2. Simulation results analysis: output delay vs input delay

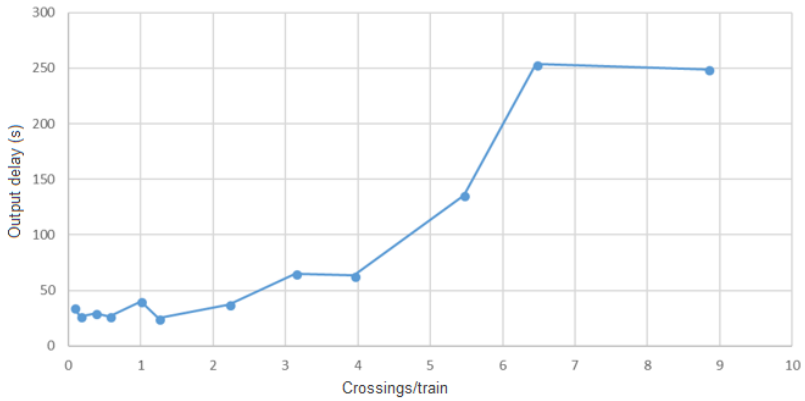


Figure 3. Simulation results analysis: output delay vs crossings.

Figure 3 illustrates the relationship between number of crossings and output delays. For small numbers of crossings there is no appreciable tendency for delays to increase. At approximately 3 crossings/train output delay starts increasing, and after 4 crossings/train it rapidly grows. The growth in output delay does not continue beyond 7 crossings/train since deadlocks in the simulation prevent trains from arriving at all.

The delay propagation model was developed to obtain a mathematical formula for estimating output delays based on input delays and crossings. In this case mathematical regression, an approach consistent with other railway delay propagation research (Marković et al., 2015) was used to develop the formula.

The first step in a regression analysis is to examine the data to determine the type of relationship. Figure 4 illustrates a quadratic trend line plotted for the relationship between input delay and output delay, while Figure 5 illustrates a quadratic trend line plotted for the relationship between crossings/train and output delay. In both cases quadratic approximations appear to be reasonable. Since quadratic equations are also easy for non-experts to apply, they were chosen for use in developing the assessment method.

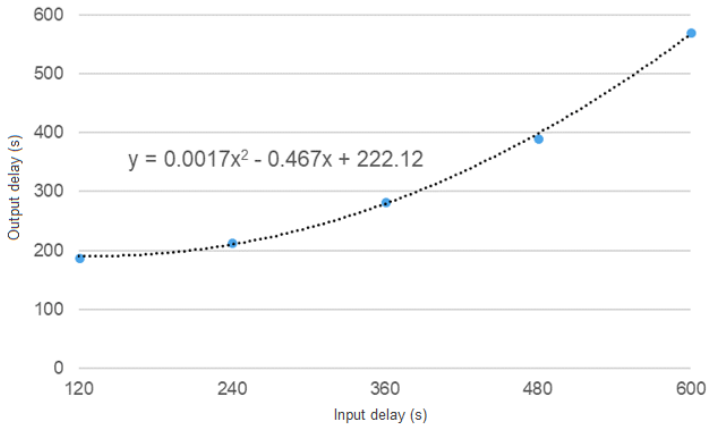


Figure 4. Output delay vs input delay: quadratic trend line.

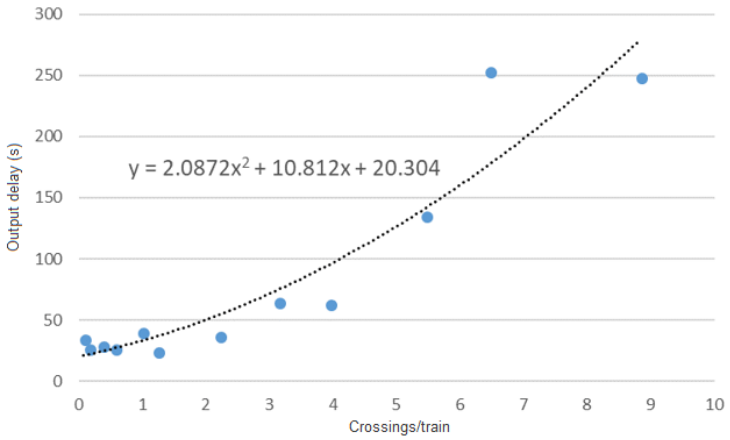


Figure 5. Output delay vs crossings per train: quadratic trend line.

The regression model for the single-track case is a combination of two quadratic formulas: one considering positive input delays and one considering the number of crossings. Denoting with $N_{x,t}$ the average number of crossings per train, and with $\beta, \gamma, \delta, \epsilon$ the regression parameters, the total expected output delay for a group of trains can be calculated with Equation (4):

$$t^{+}_{d,e} = f\left(t^{+}_{d,i}, N_{X,t}\right) = \beta \cdot {t^{+}_{d,i}}^{2} + \gamma \cdot t^{+}_{d,i} + \delta \cdot N^{2}_{x,t} + \epsilon \cdot N_{x,t} \qquad (4)$$

The parameters were obtained by running a regression on the simulation results using the XLSTAT data analysis Excel add-on. The parameters found were: $\beta = 0.0005$, $\gamma = -0.152$, $\delta = 2.127$, $\epsilon = 10.392$. Next a goodness of fit indicator was calculated to evaluate results. Denoting with $t^{+}_{d,o}$ the observed (measured) positive output delays, goodness is defined in Equation (5):

$$\text{goodness} = 1 - \frac{\sum(|t^{+}_{d,e} - t^{+}_{d,o}|)}{\sum t^{+}_{d,o}} \qquad (5)$$

The goodness measure was calculated using the identified parameters and Finnish historical data from 12 railway lines. The goodness was equal to just 10.8%, calling for an alternative approach. Thus, a mixed approach was studied. In the mixed approach, simulated data were used to estimate crossings parameters δ and ϵ (since real data do not have a sufficient range of crossings values), and real data were used to determine the input delay parameters β and γ. The regression analysis of real input delay data resulted in a negligible value for β (so it was removed from the formula), and $\gamma = 0.918$. The goodness measure calculated with these parameters was equal to 61.0% which is reasonable. The final proposed formula for estimating total output delay for a group of trains on a single-track line is presented in Equation (6):

$$t^{+}_{d,e} = 0.918 \cdot {t^{+}_{d,i}}^{2} + 2.127 \cdot N^{2}_{x,t} + 10.392 \cdot N_{x,t} \qquad (6)$$

3.2 Double-track line delay propagation method

The key parameters used to evaluate delay propagation on a double-track line are:

- the buffer times, i.e. the additional spacing between trains provided to reduce the risk of delay propagation. It is especially important to examine cases when the buffer time is limited (so called "critical headways"). Buffer times are included using a set denoted with B, with $b \in B$ buffer thresholds. Buffer thresholds are indices to denote buffers of size s_b. Each buffer, measured in minutes, ranges from a minimum $s_{min}(b)$ to a maximum $s_{max}(b)$, thus $s_{min}(b) < s_b \leq s_{max}(b)$
- the initial delay $t_{d,i}$ (referred to as "input delay" in the following text)
- the running and stop time margins t_{mr} and t_{ms}

As for the single-track case, the formula for estimating output delays for a group of trains from a set of input parameters can be obtained using mathematical regression. First, the input data were prepared aggregating values for all parameters for each line, direction, day, and time-band. The total expected output delay for a group of trains can be calculated using Equation (7):

$$t^{+}_{d,e} = \beta \cdot \sum_{b \in B} (w(b) \cdot bf_b) + \gamma \cdot t_{mr} + \delta \cdot t^{+}_{d,i} + \epsilon \cdot t^{-}_{d,i} \qquad (7)$$

Parameter bf_b is the number of buffers in a threshold b, and w(b) is the weight associated to buffer *b*. Thus, the effect of buffer times is evaluated considering the criticality of having a small buffer time, with w(b) defined to reflect this criticality: $w(b) = 2^{-s_{max}(b)}$.

Input data include 10 double-track lines with both directions separately accounted for. One-year worth of traffic data were considered, defining one train group per day/line. Buffer thresholds were subdivided into five 1-minute wide groups, from 0 up to 5 minutes. Train groups without buffers between 0 and 1 minute (the most critical ones) were not considered.

Regression performed on the input data provided the following parameter values: $\beta = 22.443$, $\gamma = -0.033$, $\delta = 1.029$, and $\epsilon = -0.001$. All parameters have a reasonable practical interpretation, and the corresponding goodness is 73.91%. Thus, they may be used in Equation (7) to create Equation (8) for estimating delay propagation on double-track lines:

$$t_{d+,e} = 22{,}443 \cdot \sum_{b \in B} (w(b) \cdot bf_b) - 0.033 \cdot t_{mr} + 1{,}029 \cdot t^{+}_{d,i} - 0{,}001 \cdot t^{-}_{d,i} \qquad (8)$$

Regression results were tested to evaluate the impact of timetable changes, by applying the proposed delay propagation formula to 4 scenarios from the UK's Crossrail project. The input delay was set at zero to simplify the analysis. Results showing the estimated effect of all timetable-dependent parameters on output delay are illustrated in Table 2.

Scenario	**SC0**	**SC1**	**SC2**	**SC3**
Number of trains	40	48	22	11
0–1 min buffers	397	559	37	0
Buffer weight	210.668	291.219	36.688	0
Margin	547.5	678	0	0
Expected delay [s/train/day]	100.4	114.1	36.4	0

Table 2. Validation of the double-track line delay propagation method. Scenario N is denoted with SCN (e.g. SC0 = Scenario 0)

The base scenario (SC0) represents the current timetable. SC1 adds 8 trains to the base timetable, resulting in a large number of small buffers. SC2 and SC3 have lighter traffic levels: SC2 has about half the trains from the base scenario, and SC3 further divides the number of trains in half. This test case study shows that the proposed mathematical model is sensitive to train frequency and provides reasonable results.

4. Conclusions

This paper discusses research carried out for the Finnish Transport Agency to develop quantitative methods for evaluating the socio-economic impacts of railway investments. Two methods were developed, the first determines capacity consumption and the second determines delay propagation. These methods are designed to provide railway network managers with simple formulas for evaluating the impacts of railway line investments without performing complex simulations.

The capacity consumption method was developed by applying the characteristics of the Finnish railway (e.g., interlocking, rolling stock) to the UIC 406 capacity formula. The paper describes the development of the parameters and highlights the differences between single- and double-track line cases. The method was then applied to a Finnish line to illustrate use of capacity over the course of a day.

The delay propagation forecasting method was developed using mathematical regression with both simulated and historical traffic data. The regression results were evaluated using a goodness measure. Separate methods were developed for the single- and double-track line cases to account for the different factors triggering delay propagation.

References

- Finnish Transport Agency, 2018. “Capacity and Punctuality in Railway Investment Socio-Economic Assessment”. Technical report.
- International Union of Railways (UIC), 1996. “Links between railway infrastructure capacity and the quality of operations” (UIC code 405 OR). Paris, France.
- International Union of Railways (UIC), 2004. “UIC leaflet 406”, France
- Landex, A., 2008. “Methods to estimate railway capacity and passenger delays”. PhD thesis, Copenhagen, Technical University of Denmark.
- Lindner, T., 2011. “Applicability of the analytical UIC Code 406 compression method for evaluating line and station capacity”. Journal of Rail Transport Planning & Management 1, pp.49-57.
- Marković, N., Milinković, S., Tikhonov, K.S., Schonfeld, P., 2015. “Analyzing passenger train arrival delays with support vector regression”. Transportation Research Part C: Emerging Technologies, vol. 56, pp.251-262.
- Moreira, N., Garcia, L., Catarrinho, P., 2004. “Network Capacity”. Computers in Railways IX Conference, Dresden, Germany.
- Nash, A., Huerlimann, D., 2004. “Railroad simulation using OpenTrack”; Computers in Railways IX, WIT Press, Southampton, pp. 45-54.
- Pitkänen, J-P., 2005, “Radan välityskyvyn mittaamisen ja tunnuslukujen kehittäminen”. Helsinki: Ratahallintokeskus, Liikennejärjestelmäosasto (in Finnish language).
- Potthoff, G., 1962. “Verkehrsströmungslehre I - Die zugfolge auf Strecken und in Bahnhöfen”, Transpress, Berlin, Germany (in German language).
- Vromans, M., 2005. “Reliability of Railway Systems”. Rotterdam: The Netherlands TRAIL School.

Transforming Automatic Scheduling in a Working Application for a Railway Infrastructure Manager

Florian H.W. Dahms [a], Anna-Lena Frank [b],
Sebastian Kühn [b], Daniel Pöhle [b,1]
[a] Vulpes AI GmbH
Textorstrasse 97, 60596 Frankfurt am Main, Germany
[b] neXt Lab, Timetable and Capacity Management, DB Netz AG
Rotfeder-Ring 3, 60327 Franfurt am Main, Germany
[1] E-mail: daniel.poehle@deutschebahn.com, Phone: +49 (0) 69 265 48267

Abstract
In this article, we present a practical approach for the optimized creation of railway timetables. The algorithms are intended to be used by Deutsche Bahn, Germanys largest railway infrastructure provider. We show how our methods can be used, both for creating a timetable in advance and for answering ad-hoc requests coming in via a digital app. Numerical experiments are provided to show that our solution exceeds manual creation of timetables in terms of capacity usage, travel times and the time taken for creating the timetable.

Keywords
railway timetable computation, traffic networks, network optimization

1 Introduction

The usual process of creating a timetable, especially for freight trains, is a manual *make-to-order* process. But as Feil and Pöhle (2014) already pointed out: It is necessary to overcome the manual process to be able to find global optimal solutions in an industrialized creation of timetables as the amount of data to be processed is steadily increasing. Additionally, in order to improve the efficiency of the process of creating a timetable as well as making better use of the available infrastructure it is convenient to transform the procedure to an *assemble-to-order* process.

This transformation as well as different approaches for an automatic creation of rail freight timetables have been widely discussed in the literature. Planning of slots was first introduced for cyclic timetables using a periodic event scheduling problem (PESP) by Nachtigall (1998) and extended by Opitz (2014). Großmann et al. (2012) showed that the PESP can also be encoded as a SAT problem which leads to significantly lower computation times. For non cyclic timetables, Großmann et al. (2013) proved that a mixed integer formulation works for practical problem sizes. The train path assignment problem as the second step after planning of slots was introduced by Nachtigall and Opitz (2014) and extended for optimization with different traffic days by Nachtigall (2015). Most recently within the DFG project ATRANS considered all aspects of automatic creation of timetables (Streitzig et al. (2016); Li et al. (2017)).

In this paper, we show how the academic models mentioned above can be transformed

into an application in real world optimization with an innovation that is threefold: First, we are able to generate more capacity without actually building additional infrastructure but by making better use of it. Second, we are able to reduce the travel time for most of the scheduled trains by simultaneously considering all requests and using global optimization instead of manual, local optimization. Third, we reduce response times by implementing the first fully automated process in short-term capacity planing from ordering to the actual departure of the train. All resulting in better customer value.

The organization of the paper is as follows. First, we describe the modelling approach in Section 2. After that, we discuss the use cases and the data used in our numerical experiments alongside the results in Section 3. We close with an outlook (Section 4) to possible extensions of our approach in the future.

2 Train Path Planning and Train Path Assignment

In this section we briefly describe our modelling approach to automatically create timetables. It is divided into two main parts: First, we precalculate slots, which are located on the most frequently used parts of the infrastructure (Section 2.1). Therefore, we extend the idea of Opitz (2014). The second part is the train path assignment (Section 2.2), based on the idea of Nachtigall (1998, 2015) and Nachtigall and Opitz (2014), where some single train paths are assigned to a complete timetable.

2.1 Train Path Calculation

This section will give a brief overview of the process used for creating a train path. The entire process consists of three main steps. First, we search one or more different routes through the infrastructure. In the second step we create a network of discrete building blocks (called snippets) along these routes for which we calculate travel and blocking times. Finally, we put these snippets together to form a non conflicting train path. The same process is used to calculate train paths for individual requests as well as the precalculated slots for the annual timetable. For the slots we consider frequently travelled relations, each starting and ending in a *Betriebstelle*. A Betriebstelle is an organizational unit into which the German railway infrastructure is divided. For each relation we consider up to three different train characteristics, chosen to be representative for most of the traffic expected on the relation.

Routing
To reduce the problem size, our first step consists of finding routes that could be relevant for the train. We use an A* algorithm for finding a shortest route on our infrastructure, which is a digital representation of the German railway network. The algorithm utilizes geo coordinates for calculating the beeline distance between Betriebstellen as the lower bound. For each Betriebstelle we keep a list of all possible ways to traverse it. Each possibility is termed a *Fahrweg*. The Fahrwege form a graph where each Fahrweg is a vertex, with directed edges indicating which Fahrweg is a direct successor of another. The edge costs are based on the Fahrwege lengths. As certain Fahrwege are preferred to others, we multiply the costs with factor greater or equal to one, with larger factors for less desirable Fahrwege. In this way the use of intersections and the use of tracks designated for the opposite direction can be discouraged, but we do not consider track or congestion charges due to regulatory

reasons. It is this graph we use for finding routes for our train paths.

While exploring the graph the algorithm filters out all paths which are incompatible with the characteristics of the train. For example the train might exceed maximum mass or width of the Fahrweg or might require an electrified track. As it is not always the best option to use the shortest path, we create multiple alternative routes. For searching the subsequent routes, we increase the costs of edges that where used in already found routes. A route is only accepted as a real alternative if it differs from all already calculated ones in at least one Betriebstelle. We set an upper limit to the length of alternative routes which is a multiple (1.3 in the experiments) of the shortest route found.

Snippet Creation

For each route we calculate the travel and blocking times that would be required for a train using the route without intermediate stops. To enable stopping, we search for all tracks that could be used for a stop of the train along the route. A track is only considered for stopping, if it branches off the main route and joins it again within a single Betriebstelle. For each such track we create one snippet leading from the main route to the stop and one from the stop to the main route. These snippets have a length of 7km, a length that ensures that acceleration and deceleration to and from the main travel speed is always possible within the snippet. For each snippet we again calculate travel and blocking times. In order to connect these snippets with the main route, we cut it into snippets at the points where our stopping snippets branch off or join it. In addition to these stopping snippets, we create snippets for alternative non stopping traversals of a Betriebstelle for all tracks that traverse the Betriebstelle similar to the original main route. Sometimes it can be beneficial for a train to travel with less than its maximum speed to match the speed of a preceeding train without stopping unnecessarily. To enable this we create alternatives of each snippet with reduced maximum speeds.

In this way we get a directed, acyclic graph of snippets representing the possible ways the train can travel along each route including intermediate stops. Each snippet has travel and blocking times calculated. A path from a source snippet (those without predecessors) to a sink snippet (without successors) will always represent a valid train path (without a specific starting time of the train).

Train Path Calculation

The last step to calculate the train path is to determine a starting time and a path through the snippet graph such that no blocking time is in conflict with another train path and the number of stops is minimized. For each snippet we can calculate the possible departure times of the snippet that are not conflicting with other trains by projecting the blocking times of other trains back to the snippets start. Thus, we get a list of intervals for each snippet where each interval represents the allowed start times. Next we reduce the intervals further to ensure that only such intervals remain that can be reached by one of the snippets predecessors. For each snippet we calculate the possible arrival times given the known departure intervals. Note, that for a snippet not ending in a stop, the resulting intervals will have the same length but are shifted by the travel time of the snippet. For stopping snippets the intervals can increase in size as long as the stop is not used by a different train. For each snippet we take the union of all its predecessors arrival intervals and intersect them with the possible departures of the snippet. This yields the new departure times of the snippet. As the snippet graph is directed and acyclic, we can find a topological ordering of the snippets such that

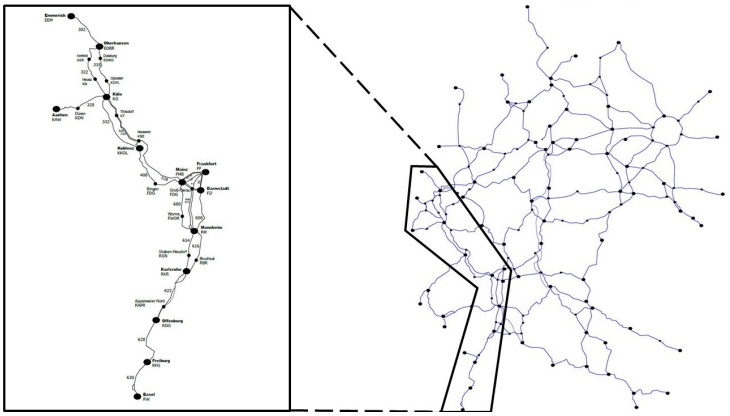


Figure 1: Right: a macroscopic map of the infrastructure administrated by DB Netze, on which slots are created. The time dependency is neglected for a better overview. Left: Zoom into the German part of the Rhine-Apline Corridor.

this reduction of departure times needs to be done only once per snippet.

The possible arrival times of the sink snippets will now be such that there must exist a non conflicting valid train path arriving at the given time. We calculate train paths by choosing the earliest arrival time. For a given arrival time we can select the predecessor snippets that have a compatible departure time. From them we select those that lead to the least number of stops (we prefer non stopping predecessors) first and then to the latest departure. The resulting train path is guaranteed to be conflict free, while keeping the number of stops and travel time low. The algorithm runs sufficiently fast, requiring $\mathcal{O}(m + n)$ calculations where m is the number of snippets and n is the number of edges in the snippet graph. Note that the maximum number of intervals per snippet is bounded by a constant, the number of seconds per day.

An example for calculated train paths is shown in Figure 2 within a path-time curve.

2.2 Train Path Assignment

As a result of the slot calculation presented in the previous section (Section 2.1), we are provided with a set of slots starting and ending at a specific Fahrweg at a specific time. The slots form a graph, we call $\mathcal{G}$ for later reference, where the slots are the edges and the tuple of Fahrweg and point in time are the vertices. In a process consisting of four steps, this graph is used to assign train paths to the requests made by customers. First, we map the requests of the customer to the graph $\mathcal{G}$, resulting in so-called *break-in* and *break-out points*. Then, we search for the shortest path within $\mathcal{G}$, where a path represents a consecutive chain of slots. In the third part, we calculate train paths from the start of a request to the break-in point and from the break-out point to the target of the request. Finally, we optimize the result for all requests using a column generation approach.

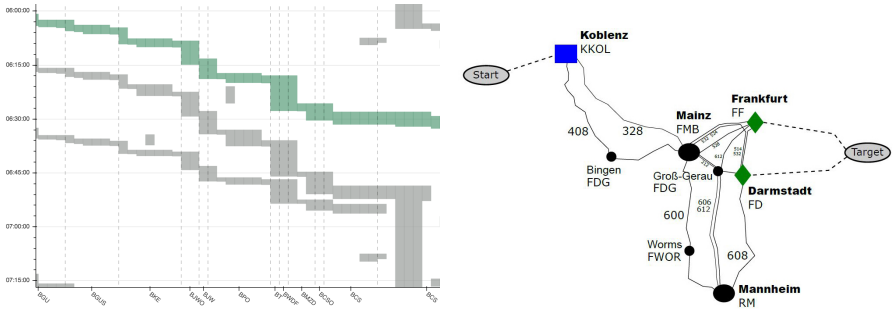


Figure 2: Left: parts of multiple train paths shown as blocking times within the time interval between 6.00am and 7.15am and Betriebsstelle BGU and BCS. Right: a small section of Figure 1. The customer request starts at *start* and ends in *target*. The break-in and break-out points are marked as blue box (□) and green diamond (◇), respectively. The time dependency is neglected for a better overview.

Calculation of Break-in and Break-out Points

As most customer requests do not start or end at vertices of the graph $\mathcal{G}$ (compare Figure 2), we need to map the start and end of a request to vertices of $\mathcal{G}$, which we call break-in and break-out points, respectively. Therefore, we calculate a number of different routes (10 in the experiment), using the A* algorithm previously described. Then, beginning at the start of the route, we consecutively check for each Fahrweg of each route whether it is associated to a vertex of the graph $\mathcal{G}$. If so, we found a break-in point and stop; otherwise no sequence of predefined slots can be used and the request has to be fulfilled with an individually calculated train path. In order to find the break-out points, we repeat the above process starting from the end of each route.

Routing on Slots

The mapping of a request previously described may not be unique (as indicated in Figure 2), so that multiple sources and sinks have to be considered. Furthermore, due to the technical properties of the requests such as e.g. length, acceleration or width, the use of slots is restricted depending on the request. Those as well as the slots also have time restrictions. Thus we end up with a restricted, time-dependent multi-source-multi-sink shortest path problem to be solved for each request.

We use standard techniques to reduce the complexity of the problem like time expansion to tackle the time dependency. We model the multi-source-multi-sink problem with dummy edges from a super source and to a super sink. We cope with the restriction due to the technical properties by using dynamic filters. In that way, we are able to reduce the problem to a standard shortest path problem, which we solve using a Dijkstra algorithm.

Individual Train Path Calculation

In the previous steps for each request, we either create a path of slots from a break-in point to a break-out point or we know that there is none. So in order to provide a train path from start to end for each request, we individually create train paths for the missing parts of each request, i.e. either both a train path from start to break-in point and a train path from break-

out point to end or a train path from start to end. For those train paths we use the approach presented in Section 2.1. Consequently, after that step we provide exactly one train path per request.

Optimization
The process described above is sufficient if only one request has to be fulfilled at a time like in the use case of the app. But if there is more than one request – like in the use case of creating an annual timetable – the simple assignment of the shortest path for each request leads to conflicts between the train paths. We solve these conflicts using a column generation approach, where in each iteration we generate new train paths which resolve more conflicts than in the iteration before. Our experiments show, that the process terminates within up to 10 hours for sufficiently large problems (compare Section 3.2).

3 Use Cases and Numerical Experiments

In this section, we describe two use cases (Section 3.1 and 3.2) to which we are able to apply the methods mentioned before. For each use case we provide numerical experiments showing the threefold benefit of faster response times, increased capacity usage and reduced travel times.

3.1 Click&Ride App

For a short-term train path request, e.g. a train run for the next day, we can improve the response time to the railway operator by using our approach in a fully automated process. We will introduce the new way of booking a train path with a mobile application called *Click&Ride-App*. We commit to provide the railway operator with a train path offering in at most three minutes. In comparison, today's process for manual planning takes several hours in most cases and may require up to three days. To ensure a maximum duration of three minutes we need to automate every single step in the planning process. A simplified process sequence is shown in Figure 3.

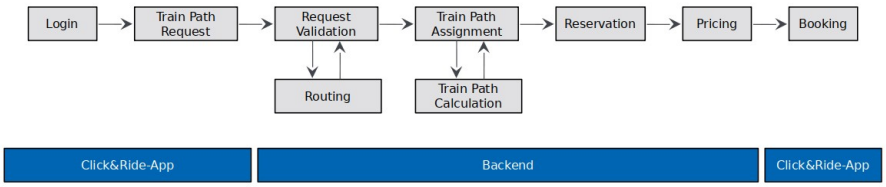


Figure 3: Simplified process of Click&Ride. The process is fully automated in the backend.

Click&Ride is a new B2B channel. After logging in, the train path request will be submitted to the back-end processes. Figure 4 gives an impression of the user interface of Click&Ride. At first there is a validation service that ensures formal and technical fit to the tracks that will be used. For example an electric vehicle cannot use a track with no overhead line. If there are any problems with the train path request, the railway operator will get instant feedback on the app's screen and the possibility to change the request. If

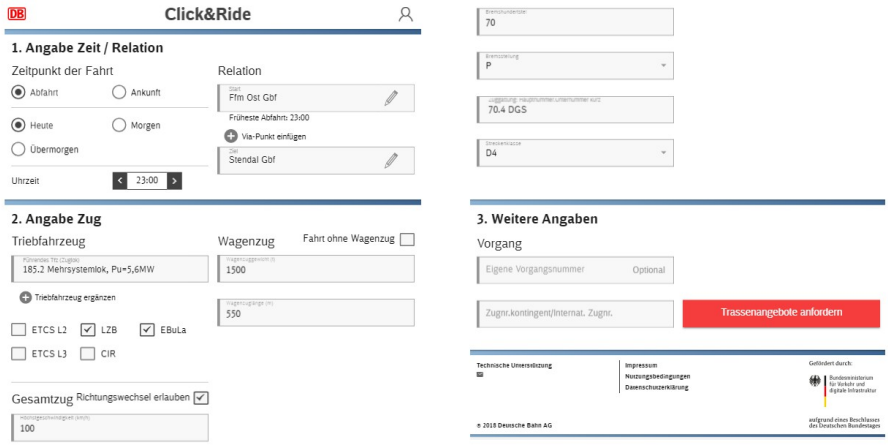


Figure 4: GUI Click&Ride to put a request. First the time requirements and waypoints need to be entered (upper left). Then the characteristics of the requested train need to be provided (lower left and upper right). Thirdly, in the lower right, there is space for additional information.

the request is validated, it is handled by the optimization in train path assignment (see Section 2.2). If the technical properties of the requested train do not match the existing slots on the lines, a train path will be generated automatically by train path calculation as described in Section 2.1. Railway operators have a limited time to check the offered train path before booking. To make sure the capacity on the track cannot be given to other railway operators in the meanwhile, the train path is reserved for a maximum of ten minutes. To complete the response the total price for the train path is calculated and displayed in the Click&Ride-App. An example response is shown in Figure 5.

Numerical Experiments

For our experiments we use 1301 real customer requests from November, 14th, 2013 in Germany. A customer request consists of the waypoints, time requirements and the characteristics of the train. The waypoints are at least the start of the request and its target, but may also include some stops which should be served in between. The time requirements for our data consists only of an interval at the start of the request. But it is also possible to provide further time restrictions on the other waypoints. The characteristics of the request include all data necessary to calculate its dynamic properties such as acceleration and its static parameters like length, width and mass of the requested train. We consider the actual German infrastructure available in 2013. Furthermore, we also regard the blockages of passenger trains, which were scheduled on November, 14th, 2013 in order to have a realistic setup. We measure the quality of a train path in a metric called BFQ (*Beförderungszeitquotient*) which is the travel time actually required by the train divided by the travel time that would have been necessary without additional stops. Note that for the Click & Ride BFQ we use the shortest travel time on the shortest route in the denominator while the BFQ for the manual planners was calculated using the shortest travel time on the route chosen by the planner.

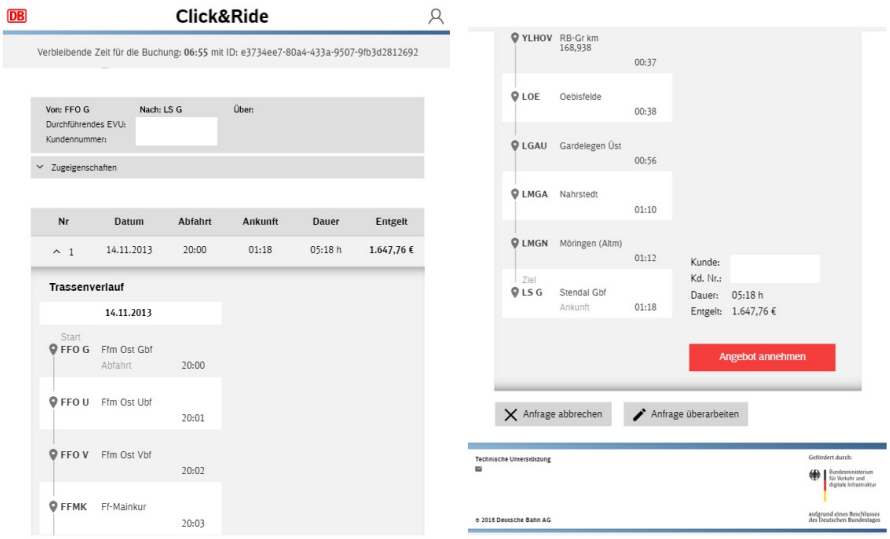


Figure 5: Part of the response, i.e. a train path, to the request shown in Figure 4.

This slight discrepancy in the numbers is to the advantage of the manual planners, as the route in the denominator could be longer.

Table 1: Percentiles for response times, automatic BFQ and manual BFQ for 1301 customer requests.

Metric	**50%**	**90%**	**95%**	**99%**	**Max**
C&R response time	3.68 sec.	50.28 sec.	82.98 sec.	147.89 sec.	222.34 sec.
C&R BFQ	1.03	1.46	1.73	2.53	6.07
Manual BFQ	1.22	1.94	2.52	4.35	10.54

The experimental results, as provided in Table 1, show that 95% of the request are served within 82.98 seconds. 95% of requests have a BFQ of at most 1.73 while the same percentile had a BFQ of 2.52 when planned manually. The maximal response time is 222.34 seconds. Our target of serving a response in less than three minutes can be fulfilled in all but a few edge cases (4 out of the 1301). This is a vast improvement compared to the up to three days required in the current manual process. The BFQ shows that the automatic process on average leads to faster train paths compared to the manual process.

3.2 Annual Timetable

The introduced methods will allow us to change the process of the creation of an annual timetable. In this use case all customer requests, for passenger and freight trains, are put at the same time and have to be provided with a timetable fulfilling all requests within 50 days. This currently means a huge effort to DB Netz and can be eased by planning the freigth trains automatically. In an iterative process we manually create timetables for the

passenger trains first. In the second step, the freight trains are calculated automatically with the methods described in Sections 2.1 and 2.2. Afterwards the timetables are adapted and improved iteratively (compare Figure 6 for a sketch of the process).

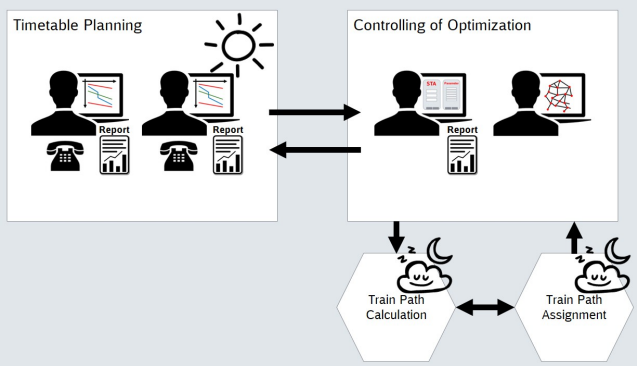


Figure 6: Iterative Process for Annual Timetable: Planning by hand during the day and automated optimization during the night

Numerical Experiments

For this experiment we again consider the data from November, 14th, 2013 regarding infrastructure, blockages and customer requests. For clarity of presentation we restrict the shown results to the German part of Corridor One (EEIG Corridor Rhine-Alpine EWIV (2019)) rather than the entire German network. Corridor One stretches from sea ports of Rotterdam, Zeebrugge, Antwerp, Amsterdam and Vlissingen to the port of Genoa covering Netherlands, Belgium, Germany, Switzerland and Italy. Here only the part in Germany from Emmerich and Aachen to Basel will be considered. We have a test set consisting of 210 requests and create slots on 38 sections (compare Figure 1, left).

The experiment shows, that our approach is very efficient compared to manual planing in 2013, as we are able to create an average of about 3% more slots. Furthermore the creation of $935,814$ slots only takes 2.5 hours compared to several days back in 2013.

Table 2: Key performance indicators (KPIs) for the train path assignment

KPIs	**Result**
Number of Customer Requests	210
computing time	02:57h
percentage train path assignments on slots	80%
average BFQ	1.27

From Table 2 we see that precalculation of slots is beneficial for our approach as 80% of all assigned train paths use slots. 20% of the requests are assigned to individually created train paths. The average BFQ is 1.27 which is more than 5% better compared to the results of the manual process in 2013. As the train path assignment takes about 3 hours, a whole timetable for Corridor One is created in 5.5 hours fulfilling 210 customer requests on $935,814$ slots.

4 Outlook

The presented approach of freight train timetable creation prepares DB Netz AG for the future by making better use of the infrastructure and reducing manual workload. Furthermore, this approach is a way to offer timetables faster and in better quality than nowadays. For the future, we plan to provide a choice set of up to three different timetables within the app, for the customer to decide which fits best. Up to now, we only consider one day for the planning, so a possible limitation might result from the extension to a whole year, e.g. the amount of data generated or the requirement to create homogenous timetables. Nevertheless, we are currently working on this extension to cover multiple days in a simultaneous slot assignment. Additionally we want to use the presented techniques for additional use cases, e.g. the creation of short term timetables and the handling of construction sites. This requires that we extend the algorithms from freight to passenger train scheduling.

Acknowledgements

This work was financially supported by BMVI, Project DigiKap. We also thank Sascha Diekmann for his support with the figures.

References

Feil, M., Pöhle, D., 2014. “Why Does a Railway Infrastructure Company Need an Optimized Train Path Assignment for Industrialized Timetabling”, International Conference on Operations Research, Aachen.

Großmann, P., Hölldobler, S., Mantey, N., Nachtigall, K, Opitz, J., Steinke, P., 2012. “Solving Periodic Event Scheduling Problems with SAT”, In: *Advanced Research in Applied Artificial Intelligence*, Springer Berlin Heidelberg.

Großmann, P., Labinsky, A., Opitz, J., Weiß, R., 2013. “Capacity-utilized Integration and Optimization of Rail Freight Train Paths into 24 Hours Timetables”, In: *Proceedings of the 3rd International Conference on Models and Technologies for Intelligent Transportation Systems 2013*, TUDpress, Dresden.

Li, X., Nachtigall, K., Martin, U., Oetting, A., 2017. “Methodik zur effizienten markt geeigneten Trassenbelegung im spurgeführten Verkehr”, *Eisenbahntechnische Rundschau*, vol. 6.

Nachtigall, K., 1998 “Periodic Network Optimization and Fixed Interval Timetables”, Habilitation thesis, Hildesheim.

Nachtigall, K., Opitz, J., 2014. “Modelling and Solving a Train Path Assignment Model”, International Conference on Operations Research, Aachen.

Nachtigall, K., 2014. “Modelling and Solving a Train Path Assignment Model with Traffic Day Restriction”, In: *Operations Research Proceedings 2015*, Springer, Heidelberg.

Opitz, J., 2009. “Automatische Erzeugung und Optimierung von Taktfahrplänen in Schienenverkehrsnetzen”, Dissertation, Dresden.

Streitzig, C., Nachtigall, K., Martin, U., Oetting, A., 2016. “Anforderungsgerechte Algorithmen zur effizienten marktgeeigneten Trassenbelegung”, *Eisenbahntechnische Rundschau*, vol. 8.

EEIG Corridor Rhine-Alpine EWIV. “https://www.corridor-rhine-alpine.eu/about-us.html” Website by 2019-01-31.

ASSESSMENT of POTENTIAL COMMERCIAL CORRIDORS for HYPERLOOP SYSTEMS

Design of a methodology to select and classify the most suitable corridors for the implementation of a first commercial hyperloop line for passengers

Marc Delas [a,1], Jeanne-Marie Dalbavie [b,2], Thierry Boitier [c,3]
[a] IKOS Consulting Deutschland GmbH
c/o WeWork Sony Center, Kemperplatz 1, 10785 Berlin, Germany
[1] E-mail: mdelas@ikosconsulting.com
[b] Ikos Lab
155 rue Anatole France, 92300 Levallois-Perret, France
[2] E-mail: jmdalbavie@ikosconsulting.com
[c] TransPod Inc
101 College St, Suite HL45, Toronto, Ontario, M5G1L7, Canada
[3] E-mail: Thierry.boitier@transpod.ca

Abstract
This study has led to the elaboration of a proposed methodology used to select and rank the most attractive corridors for the implementation of first commercial vacuum-tube train (or hyperloop) lines for passengers.
From a list of the most populated cities all over the world, it has been possible to sort out the possible transport connections that could be travelled by hyperloop pods without having to build a tunnel or crossing a conflict area.
Then, an evaluation of all selected corridors has been performed on the basis of defined classification criteria. Important parameters characterizing the potential of a corridor have been identified during the research: the number of air passengers on the corridor, the nature of the competitive transport infrastructure, the GDP per kilometre and the topography along the route. Some other minor criteria have also been used, in order to elaborate a robust tool which can be a good help for investors and decision makers.
All selected corridors have been ranked, resulting in a short list of the 250 most attractive corridors for the implementation of first commercial lines.
This study presents a proposal for the ranking of the most promising corridors. It should be followed by proper feasibility studies and ridership calculations.

Keywords
Vacuum-tube train, hyperloop, transport economics, methodology

1 Introduction

In a context where the world population increases rapidly and travels a lot more than before, resulting in a "hyper-mobility" (Crozet (2016)), transport systems are key. Besides, the concerns about a man-caused climate change put the sustainability of our mobility models in question. Indeed, if the Humanity continues to increase its daily mobility with current technologies, it will inevitably conflict with its goal to globally reduce greenhouse gas emissions as agreed in the Paris Agreement signed at the end of the

COP 21 in 2015 in Paris, France.

That is the reason why engineers, scientists, investors and business men all over the planet are imagining "greener" transport solutions that allow the humanity to continue increasing its mobility without putting the survival of the planet and its own existence into danger. The most disruptive solution is the vacuum-tube train, named hyperloop by Elon Musk in 2013 (Musk (2013)). It consists in capsules propelled at very high speed (up to 1,220 km/h) by electromagnetic systems inside low pressure tubes, so that friction and aerodynamic resistance are practically inexistent (Musk (2013)).

Even if the technical feasibility of this idea is still not guaranteed, several companies have started to work on the concept and conduct first tests on their prototypes (Davies (2017)). Three of them are ahead of the market: TransPod, Hyperloop Transportation Technologies and Virgin Hyperloop One. All three are raising funds and negotiating with public authorities to sound out interest in the market. Given that the opportunities of this technology are great and that the first successful implemented system would bring a big advantage to the company that has imagined and produced it, the evaluation of the most attractive places for the development of the first commercial lines is a decisive step.

As no work currently exists on this topic, TransPod and IKOS consulting conducted a study which aims at defining the basis of a methodology for selecting and classifying the most attractive corridors for the implementation of the first commercial hyperloop line for passengers.

2 Appropriateness of conventional methodologies to compare corridors for the implementation of a new commercial hyperloop line

2.1 Cost-Benefit Analysis (CBA)

The most usual methodology to assess the feasibility of a large transportation project and compare it with alternate ones is the Cost-Benefit Analysis (CBA). It uses monetized values (measured in monetary units) to compare total incremental benefits with total incremental costs (Transportation Research Board (n.d.)). To ensure the viability of the project on a long-term perspective, costs and benefits are estimated over a long period of time (20 to 30 years for large transportation projects like railways, roads or airports).

This methodology is mostly used to rank suitable alternatives for new or existing commercial transport lines on a defined corridor. It could therefore be an appropriate tool to compare hyperloop with competitive modes of transport on a specific corridor, based on a financial and economic analysis. As a matter of fact, it would take into account the economic benefit of very high speed, diminution of traffic congestion and possible low emissions with the use of clean energy. On the other hand, it would integrate very high investment costs, concerns about safety and reliability as well as land use (which is a very critical point in dense cities).

Theoretically, for the purpose of this study, which is to select and rank corridors for the implementation of the first commercial hyperloop line, the CBA would also be an appropriate methodology. But as the hyperloop technology is still at an early stage of development, the estimation of decisive parameters for the analysis (cost of a kilometre vacuum-tube, passenger demand, hyperloop users' value of time) is associated with a great margin of error. The evaluation of benefits and costs on each corridor would therefore be very imprecise and the analysis wouldn't be a good basis for investors and

decision makers to determine where to launch the first project of a commercial hyperloop line.

2.2 Multi-criteria analysis (MCA)

The Multi-Criteria Analysis (MCA) is another methodology which can be used to make a choice between several alternative projects, based on an algorithm that combines a set of relevant criteria for the choice and their relative "weights". A scale is defined for the evaluation of the relevant criteria. It can be continuous (e.g. if the evaluation score can take every value between 0 and 5) or discrete (e.g. if the evaluation score can only take the values 0, 1, 2, 3, 4 and 5). Each alternative is given a score on each criterion which is then multiplied by the corresponding weight. At the end, all weighted scores are added, resulting in a representative performance for each alternative. The comparison of those performances gives the most suitable alternative. An example of a Multi-Criteria Analysis is given in Table 1.

Table 1: Example of a Multi Criteria Analysis

Criterion	Weight	Alternative 1		Alternative 2		Alternative 3	
		Evaluation score on the criterion	Weighted score on the criterion	Evaluation score on the criterion	Weighted score on the criterion	Evaluation score on the criterion	Weighted score on the criterion
Time before start of commercial service	1	2	2	3	3	5	5
Safety	7	1	7	3	21	4	28
Environmental impact	3	0	0	1	3	2	6
Social benefit	3	5	15	2	6	3	9
Return on investment	5	2	10	5	25	1	5
Sum	19	10	**34**	14	**58**	15	**53**

According to the MCA presented in Table 1, the alternative 2 is the most suitable because it has the highest sum of all weighted scores (though it does not have the highest sum of all raw scores, which illustrates the importance given to the subjectively attributed weights in this methodology).

The MCA seems to be a well appropriate methodology to select and rank corridors for the implementation of the first commercial hyperloop line. Indeed, it does not require monetizing all benefits and costs like the CBA. However, as the system at stake does not exist yet, and as it exists an immense quantity of corridors in the world, it is not possible to use the MCA as is. A declination and selection by steps had been added to the MCA concept. Moreover, as hyperloop is mixing transportation characteristics of railway and plane, the criteria to be chosen had never been set nor explored in this way.

3 Principles of the methodology for selecting and ranking the most suitable hyperloop transportation corridors

First, the methodology developed in this study selects the routes where a hyperloop system would be technically feasible, economically viable, reliable and safe. To this end,

some requirements are defined to automatically exclude the corridors which are obviously not suitable for the implementation of a hyperloop system (for demographic, economic, geographical or political reasons).

Then, the remaining possible routes are evaluated in a Multi-Criteria Analysis. The ranking of the corridors' interest for the implementation of a hyperloop system are worked out based on the evaluation of specific, scalable and reliable criteria, modulated by margin of error. Each criterion is attributed a conversion method from its initial range of values in its initial unit of measurement to a standardized dimensionless range of values between 0 and 10. The rating 0 indicates a route that is not at all interesting according to the assessment criterion under consideration, whereas the rating 10 is the translation of a most attractive one. After that, criteria are weighted according to their estimated contributions to the attractiveness of a corridor. The more relevant, objective and reliable the criterion is, the greater the weight. Finally, each corridor is rated by summing the weighted ranges of values of each criterion.

The figure 1 illustrates the methodology and its different steps, from the selection of potentially interesting corridors to the ranking of the most suitable ones for the implementation of a first commercial hyperloop line. The following paragraphs get into more detail in the development of the methodology and its application.

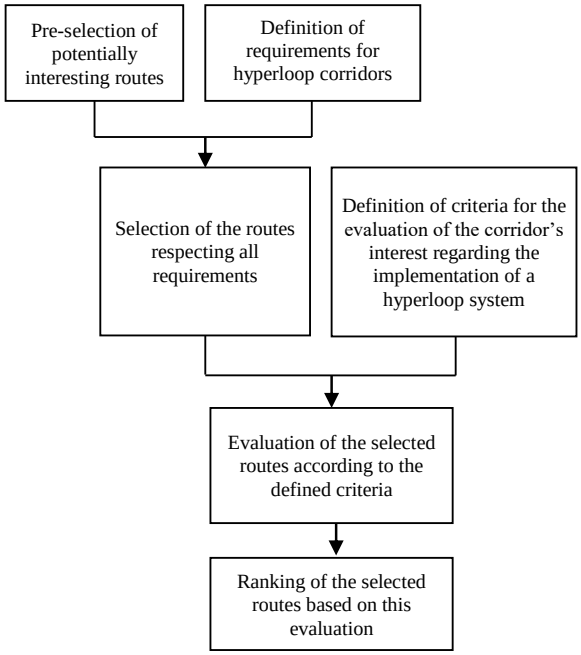


Figure 1: Methodology used to select the most interesting corridors to implement a first hyperloop line for passengers

4 Selection of the most suitable routes

In the first place, following selection criteria are defined and applied with a view to establishing a list possible origin and destination points for hyperloop corridors:

- **Population:** The passenger demand for the use of a hyperloop system has to be high enough. That is why only urban agglomerations with more than 300,000 inhabitants and capital cities with 100,000 to 300,000 inhabitants of countries populated with at least 500,000 inhabitants (this includes Luxembourg City) are considered as possible origin and destination points.
- **Geography:** Crossing a sea or an ocean would drastically increase the infrastructure costs. Only cities connected by land can form an appropriate corridor for a first investment in a hyperloop system.
- **Economy:** Building a hyperloop system requires a huge investment. Low-income economies (GNI per capita lower than $1,006 in 2016) are excluded of the pool of cities.
- **On-going conflicts and tensions**: All of the 28 areas listed by Australian and French Governments as insecure places (war, conflict, high tensions, terrorism, and armed groups presence) are excluded.

After this first selection, 1,488 cities spread over 114 territories remain in the list of possible origin and destination points for Hyperloop corridors. The goal is to reduce this number to 500 cities, in order to limit the amount of data without sacrificing any important population and employment center. The method used to select the final 500 cities is based on the following criteria:

1) No country can have more than 50 cities in the list.
2) The maximum number of cities for a country is determined by:
    - The country's population (for half), available from official statistics
    - The country's GDP (for half), calculated by multiplying the country's population by the country's GDP per capita (cf. United Nations Statistics Division 2017)

Thus, the EXCEL formula used to determine the maximum number of cities for the country j is:

$$M_j = MAX\left(ROUNDUP\left(\frac{1}{2}A\frac{P_j}{\sum_i P_i} + \frac{1}{2}B\frac{GDP_j}{\sum_i GDP_i}; 0\right); 50\right) \quad (1)$$

Where: M_j is the maximum number of cities for country j; P_j is the population in country j ; $\sum_i P_i$ is the total population of all remaining countries in the list ; GDP_j is the GDP of country j ; $\sum_i GDP_i$ is the total GDP of all remaining countries in the list ; A and B are coefficients chosen such that $\sum_j A\frac{P_j}{\sum_i P_i} = \sum_j B\frac{GDP_j}{\sum_i GDP_i}$ and $\sum_j M_j = 500$; A_j is the actual number of cities of country j in the list.

The global distribution of the 500 selected cities can be seen on Figure 2.

Possible connections between those cities are filtered in order to meet the following requirements:

- **Route length:** The hyperloop is a very high speed mode of transport which mostly competes with air and has it greatest interest on middle to long distances. To reach an interesting commercial speed, the route has to be longer than 300km. However, due to the huge investment cost, a first connection over 1,500 km would hardly find the funding. Only city pairs that are 300 to 1,500 km apart are therefore considered

in this study to form a potential hyperloop corridor. The distance between the selected cities is estimated using the GIS software QGIS 2.18.12 (with the tool "Distance matrix") with a margin of error under 5%.

- **Geography:** Corridors that require the construction of a new undersea tunnel or a new bridge over the sea or the ocean are not selected for economic reasons.
- **On-going conflicts and tensions:** Corridors that cross insecure areas are excluded.

After having performed the selection on the origin and destination points and on their connections, 6,167 corridors remain in the list of the most suitable routes for the implementation of a first commercial hyperloop line. With help of a Multi-Criteria Analysis, they are ranked from the most to the less attractive one.

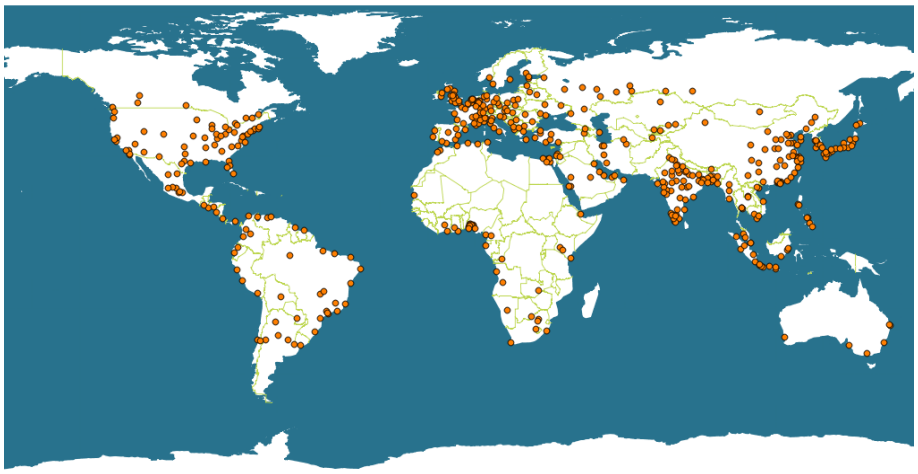


Figure 2: Global distribution of the 500 selected cities

5 Ranking methodology of the selected corridors

5.1 Criteria used for the evaluation

The following criteria are used to perform the Multi-Criteria Analysis:

- **Air traffic**: Air is the closest transport mode to hyperloop in its characteristics: very high speed, limited number of passengers, high users' value of time. That is why the demand for hyperloop will probably be high where the air traffic is significant. Air transportation being well developed and traceable throughout the world, accurate and reliable data is available [protected source].
- **Average load factor per aircraft:** A saturated air traffic indicates that there is a potential for more passenger demand on very high speed transport modes. The average load factor is often given with the passenger traffic on a corridor.
- **GDP per kilometre**: The GDP at a national level does not reflect the disparities of wealth and population between two cities of the same country. That is the reason why we use the GDP of metropolitan areas evaluated by McKinsey (Mc Kinsey & Company (2016)) rather than the GDP of countries. Considering the GDPs of origin and destination, it is logic to consider that the higher the sum of their GDPs, the more profitable the hyperloop line will be. Concerning costs, it is assumed in the first place

that the construction and exploitation costs are directly proportional to the distance d between origin and destination. The two dimensions (costs and benefits) are summed up in one single indicator that we note C:

$$C = \frac{GDP_{Origin} + GDP_{Destination}}{d} \quad (2)$$

- **Trip nature:** The higher demand is on domestic trips (connections between two cities of the same country). Domestic corridors will therefore be preferred to international corridors.
- **Route length:** The ideal compromise between a corridor where the very high speed reached by the hyperloop system brings a significant gain of time and a route whose infrastructure costs stay reasonable is evaluated around 600 km route length. The most interesting corridors for the implementation of a first commercial hyperloop line are the one between 300 and 900 km. Over 900 km, the investment costs are barely sustainable. The distance between the selected cities is estimated using the GIS software QGIS 2.18.12 (with the tool "Distance matrix") with a margin of error under 5%.
- **Natural disasters**: Natural disasters would have a major impact on the hyperloop system. In a document produced by the United Nations, every major city is associated with the number of natural disasters it is exposed to.
- **Topography:** A steep terrain would raise the infrastructure costs and cause too high accelerations for passengers. The data for topography is accessed using Google Earth Pro 7.3.0.3832
- **Available transport modes**: High-speed rail, highway, conventional rail would directly compete with hyperloop. The less available transport modes, the more interesting the corridor. Data on available transport modes is reliable and exhaustive, often furnished by the public authorities themselves, as they have an interest and a duty to communicate on such big infrastructure projects.
- **Country's GDP per capita:** Potential users of a hyperloop system are people with a high purchasing power. The United Nations provide a list of the GDP per capita in every country all over the world.
- **Country's ecological performance:** The hyperloop is potentially a clean transport mode, as it does not use any fossil fuel. Countries interested in reducing their greenhouse gas emissions are more likely to invest in this technology. The "ecological performance" of the country is not easily available, that is why only major countries will be evaluated on this criterion based on the GGEI (Global Green Economy Index).

5.2 Scaling of the evaluation criteria

All criteria are expressed in different physical units and cannot be directly added. Therefore, a conversion method to a standardized dimensionless parameter is required for every criterion. In the end, after conversion of the actual value, a number between 0 and 10 is obtained. The rating 0 indicates a route that is not at all interesting according to the assessment criterion under consideration, whereas the rating 10 is the translation of the most attractive power by only taking this criterion under account.

Different mathematical functions are used to convert each parameter into a

standardized dimensionless parameter between 0 and 10: discrete, linear, quadratic, logarithmic. The goal by developing those conversion functions is to develop a methodology which can discriminate corridors on relevant criteria directly or indirectly related with costs and benefits for the whole system.

5.3 Weights attributed to the standardized dimensionless parameters

The following weights are attributed to the standardized dimensionless parameters:

Table 2: Weights attributed to the standardized dimensionless parameters

Standardized dimensionless parameter	Corresponding classification criterion	Weight
P_1	Air traffic	8
P_2	Average load factor per aircraft	3
P_3	GDP per kilometre	9
P_4	Trip nature	2
P_5	Route length	2
P_6	Exposure to natural disasters	2
P_7	Topography	6
P_8	Available transport modes	8
P_9	Country's GDP per capita	4
P_{10}	Country's ecological performance	2
Total		46

In this evaluation, four criteria are particularly relevant:

- Air traffic
- GDP per kilometre
- Topography
- Available transport modes

By integrating the topography and the competitive situation among the most important parameters, the feasibility of the project and the possibility to get the support of both public and private investors are taken into account. Other criteria relative to the countries involved in the project like their GDP per capita and their interest in developing a greener economy are less important elements, as they are changing and subject to a frequent reevaluation.

6 Results

The weighted sum is calculated with Excel, giving a single score to every corridor. Then, corridors are ranked in decreasing order of this score. According to this ranking, the ten most attractive corridors for the implementation of a first commercial hyperloop line are (in that order):

1) Chicago – New York City
2) Houston – Dallas-Fort Worth
3) Sydney – Melbourne
4) Washington, D.C. – New York City
5) Detroit – New York City
6) Montréal – Toronto
7) Orlando – Atlanta
8) Buffalo – New York City
9) Atlanta – New York City
10) Tampa-St. Petersburg – Atlanta

It can be noted that 8 of the 10 most attractive corridors are domestic routes in the United States. Many reasons can explain this:

- Air traffic is very important in the United States.
- The GDP of many American metropolitan areas is very high as the average GDP per capita is one of the highest on the planet.
- Rail passenger traffic is not very developed over the country and there are still no high speed rail lines operating in the United States (but some are planned).
- The east side of the country is relatively flat.

Thus, the United States of America, and particularly its east side, is a great area to implement the first hyperloop line. The main problem though could be the reluctance of the federal state to invest in a massive transport project like this. It would create disparities between the states that will not be easy to compensate in the future without investing massively in developing a national hyperloop network.

It is surprising that no corridor located in the Middle East appears among the 100 best ranked routes. As a matter of fact, countries in this region of the world are willing to develop and diversify their transport infrastructure and possess the financial capacity to do so. A possible reason why no corridor in the Middle East is ranked among the best ones is an undervaluation of the cities' GDP in the Arabian Peninsula. This raises the question of the relevance and the quality of the ranking methodology.

7 Discussion

A limit to the study is the lack of data on certain classification criteria, such as the GDP of metropolitan areas, for which only one serious source was available and exhaustive. Moreover, the quality of the selected data and the methodology can also be questioned as corridors in the Middle East that intuitively seem attractive for the implementation of a Hyperloop system are not even classified under the top 100.

Some classification criteria are dependent on each other, which raises the question if there was no possibility to combine them in a single indicator. Air traffic, GDP per kilometer and country's GDP per capita to a lesser extent are all linked to the amount of people with a high purchasing power who could be potential hyperloop users. Similarly, the GDP per kilometer depends directly on the route length. But as the conversion function is linear for the GDP per kilometer (the higher the GDP per kilometer the most attractive the corridor, which at constant GDP means: the lower the distance the most attractive the corridor) whereas it is quadratic for the route length (corridors between 300 and 900 km are privileged with an optimum at 600 km), evaluation are interpreted differently for both criteria.

The attribution of a score ranging from 0 to 10 based on a weighted sum results in small differences between consecutive corridors in the ranking. That is why a small change in the conversion method from the initial value in its initial unit of measurement into a standardized dimensionless parameter can totally change the final ranking. Hence, it is more accurate to analyze the ranking by forming groups of corridors with similar scores rather than considering only the first one and leave the rest aside.

Moreover, it is very difficult to reflect in a criterion the political will of a city or region to invest in the installation of a new line of a new transportation mode. Hence this impactful feature is poorly taken into account.

8 Conclusions

This study has led to the elaboration of a methodology used to select and rank the most attractive corridors for the implementation of a first commercial Hyperloop line. This methodology has been based on a Multi-Criteria Analysis, which differs from the most usual assessing methodology for large transportation projects: the Cost-Benefit Analysis.

Starting from a list of the most populated cities all over the world, it has been possible to select and rank the most suitable corridors for an investment on the development of a first vacuum-tube train line. The data used to obtain this ranking have been thoroughly chosen, analysed and weighted in order to integrate multiple dimensions (economic, geographical, environmental, political, safety-related).

At the end, a list of the 250 most attractive corridors for the implementation of a first commercial Hyperloop line has been elaborated. This list is a good starting point for further study. But it can be revised and improved by cross-checking the data with help of complementary databases. Besides, routes located in the United States are overrepresented, whereas some promising connections in the Middle East do not appear on top 20 of the ranking. A reevaluation of the data and the methodology could help correct these inconsistencies.

Now that the most attractive corridors for the implementation of a first Hyperloop line have been identified, it would be interesting to develop a methodology to forecast ridership and revenues on a connection and to apply it to the 10 best corridors in the ranking. To do that, transport models like the Logit Model could be used. This will help refining this study by quantifying precisely the costs and benefits associated with the vacuum-train system on each line. If the profitability of the exploitation is demonstrated, a hyperloop service could then be implemented on the most favorable corridor.

References

Crozet, Y., 2016. *Hyper-mobilité et politiques publiques. Changer d'époque ?*, Economica (Méthodes et approches), Paris.

Davies, A., 2017. "Hyperloop's First Real Test Is a Whooshing Success", *Wired*, Available online at https://www.wired.com/story/hyperloop-one-test-success/, checked on 1/4/2018

Mc Kinsey & Company, 2016. "Urban World" – Android Apps on Google Play. Available online at https://play.google.com/store/apps/details?id=com.mckinsey.urbanworld&feature=nav_result#?t=W251bGwsMSwxLDMsImNvbS5tY2tpbnNleS51cmJhbndvcmxkIl0, checked on 1/2/2018.

Musk, E., 2013. "Hyperloop Alpha.", Available online at https://www.tesla.com/sites/default/files/blog_attachments/hyperloop_alpha3.pdf, checked on 5/15/2017

Transportation Research Board, n.d. "Transportation Benefit-Cost Analysis", The National Academies of Sciences, Engineering, and Medicine, Available online at http://bca.transportationeconomics.org, checked on 1/19/2019

Impact of calibration of perturbations in simulation: the case of robustness evaluation at a station

Marie Milliet de Faverges [a,b], Christophe Picouleau [a], Giorgio Russolillo [a]
Boubekeur Merabet [b] Bertrand Houzel [b]
[a] CEDRIC laboratory, CNAM Paris, France
[b] DGEX Solutions, SNCF Réseau, Saint-Denis, France

Abstract
This paper deals with robustness evaluation at station, and in particular for the train platforming problem (TPP). This problem consists in a platform and route assignment in station for each scheduled train. A classical robustness evaluation is simulation: simulated delays are injected on arriving and departing trains then propagated, and results are averaged on a large number of trials. A robust solution of the TPP aims to limit the total amount of secondary delays. However, a simulation framework at station is difficult to calibrate: it requires a realistic delays generator and an accurate operating rules modeling.
This paper proposes an original simulation framework using classical statistical learning algorithms and calibration assessment methods to model simulation inputs. This methodology is applied on delay data to simulate delay propagation at station. It highlights the importance of delay calibration by showing that even slight miscalibration of inputs can lead to strong deviations in propagation results.

Keywords
Simulation, platforming problem, Calibration, Machine learning, Delay Distribution

1 Introduction

Robustness evaluation is a central topic for both academical and industrial actors in the railway field. Resources are saturated, demand is increasing and the network is congested, while investments are rare and expensive. This leads to strong pressure on infrastructure manager and railways companies to respond to these new problems. The challenge is particularly important at main stations: they form bottlenecks on the railway network, and delays propagate fast due among others to shared infrastructure, rolling stock planning and passenger activity. It is crucial to optimize railway operations robustness at station to limit the impact of perturbations.

The recent availability of delay data is a promising opportunity for that. Delays are recorded at different points of the railway network, allowing to have a better comprehension and analysis of perturbations occurrences and propagation. This is useful to improve railway models accuracy at different levels (delay distributions, operating rules,..) or to imagine new strategies based on these records.

This paper presents preliminary results on possible utilization of Machine learning approaches for robustness evaluation at station. It proposes a simulation framework using classical statistical learning algorithms and calibration assessment methods to model simu-

lation inputs. The learning model estimates individual probabilities of delay of each train based on the context, and the quality of the predicted probabilities is assessed independently of the simulation. These predictions are then used to simulate delay propagation at station. The machine learning approach is compared with other delay models. This experiment highlights the importance of calibration by showing that even a slight miscalibration of inputs can lead to strong deviation in propagation results.

This study is structured as follow: section 2 presents a short overview of existing works on railway simulation for robustness evaluation, section 3 describes the case study and the chosen methodology. Delay modeling work is shown in section 4 and delay propagation algorithm in section 5. Experiments are conducted in section 6 and results are discussed in section 7.

2 Related Work

This research proposes a new way of assessing the calibration of the perturbations generator in a simulation framework. Reviews of related studies conducted on both simulation for railway robustness evaluation and delay modeling are provided in this section.

2.1 Simulation for robustness evaluation:

A robust solution of an operations research problem is in general defined as a solution that will remain feasible when input parameters experience small variations. In railway research, schedules are usually not feasible anymore when disturbances occur, and robustness is more about finding a solution that can be recovered with limited use of dispatching (delay propagation, rescheduling, reordering, etc). In particular for railway station operations, a robust solution generally aims to reduce delay propagation and the amount of secondary delays (Caprara et al. 2010; Armstrong and Preston 2017).

There are two main ways to evaluate robustness of schedules, and in particular at station. The first one is to define reliability indicators based on characteristics of the schedule (headways, residual capacity, margins, etc). For instance Carey 1999 proposes deterministic reliability measures based on headways spreading in station. Performance indicators are easy to compute, but only give a partial vision of the robustness as they do not reflect traffic performances. The second one is simulation. It requires extensive description of the infrastructure, operating rules and perturbations distribution, but gives a more realistic and global evaluation of the ability of the solution to deal with small perturbations in real conditions.

It is however important to calibrate the parameters of the simulation tool correctly, especially the operating rules and the disturbances distribution used for sampling (Koutsopoulos and Wang 2007). Setting operating rules is complicated: real-time dispatching decisions are various (reordering, rerouting, event cancellation, etc) and it may be difficult to anticipate agents' choice in real-time. Moreover, these dispatching actions are not compatible with robustness evaluation concepts (reduced delay propagation with limited use of delay management). It must be decided during the simulation tool design what are the available decisions, and in which conditions they are applied. Carey and Carville 2000 use simulation and delay propagation algorithm to analyze reliability of routing and platforming solutions. Two operating frameworks are studied: one with fixed platform assignment and the other one with the possibility of platform changes, reducing strongly the amount of knock-on delays. On the other side, calibration of the disturbance distribution is also a key

element: simulation aims to estimate the behavior of the solution in real operations. For that, simulated delays must be reasonable, otherwise results will not be relevant. Usually, perturbations are generated according to a given probability distribution and then applied on the solution. For instance, Carey and Carville 2000 generate small delays using a uniform distribution and a beta distribution and apply them to randomly chosen train at each step of the simulation.

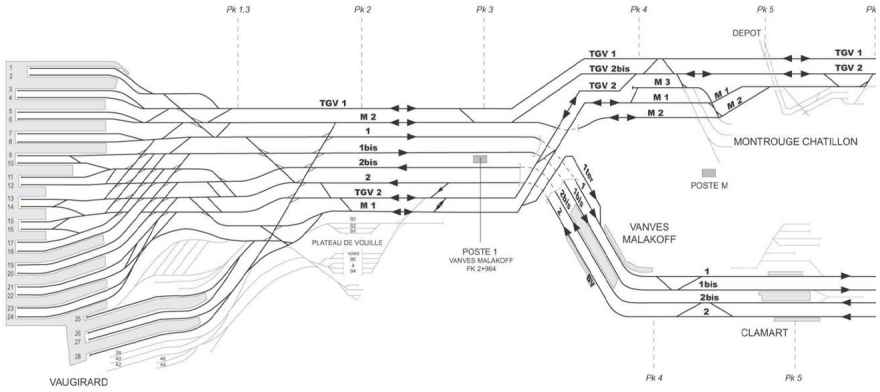
For the past few years, railway data, and in particular historical records of realized circulations have been more available. This is a promising opportunity for reliability measurement, and in particular for the sampling in simulation tools. Indeed, it is difficult to generate by hand a distribution that is concordant with reality, and actual observations of the network may help to find a way to generate reasonable d elays. To deal with this issue, Landex and O. A. Nielsen 2006 calibrate the delay distribution and rules of operations by comparing actual outputs and simulated outputs. They repeat this step several times before using their module to evaluate the robustness of timetables. Büker and Seybold 2012 express the issue of unknown primary delay distribution, since primary and secondary delays are not separated in operational data. Similarly, they compare key performance indicators based on results of simulations with operations records in order to calibrate the distributions. Koutsopoulos and Wang 2007 propose a calibration methodology based on minimization of the error between observed and simulated measurement. Larsen et al. 2014 model dwell times with a Weibull distribution for robustness evaluation using simulation. The location and shape parameters are estimated by maximum likelihood for peak hours and off-peak hours using records of arrival and departure times. Cui, Martin, and Zhao 2016 present an original method using reinforcement learning to automatically calibrate initial delays of simulation tools. Disturbances parameters are updated until convergence of a cost function. They present an application on a real network, where two parameter (mean delay and probability of delay) are tuned per combination type of train/ type of disturbance.

2.2 Delay Modeling

Train delay modeling is a well studied subject in railway research. Many studies have focused on finding an adequate distribution for empirical observations of train delays. Goverde 2005 uses a Kolmogorov-Smirnov test to assess the goodness-of-fit of state-of-the-art distributions (normal or negative exponential) on different types of delays recorded at the Eindhoven station (arrival delays, arrival non-negative delays, departure delays and dwell time excedents). Yuan 2006 evaluates different candidate distributions for delay records from the station the Hague, with one test per train type and direction. The Weibull, Gamma and log-normal distributions fit non-negative arrival and departure delay data well based on the Kolmogorov-Smirnov test. Briggs and Beck 2007 model delays in UK with q-exponential laws. Bergström and Krüger 2012 compute maximum likelihood estimation of the coefficients of lognormal, negative exponential and power-law distributions, and compare them graphically with observations from the Swedish railway network. Wen et al. 2017 show that primary delay durations are better fitted with a log-normal distribution than a Weibull one, even for data from different stations or during different period of the day. Harrod, Pournaras, and B. F. Nielsen 2018 show that delays on the Danish network are beter modeled with mixed distributions of lognormals than with a negative exponential distribution.

However results may depend on a large number of factors, like the type of delay (arrival, departure, dwell time), the range of values, location (station, line, etc), type of train,

Figure 1: Montparnasse station layout



operating rules, etc and may not be transposable from one case study to the other. For high-speed arrival non-negative delay data from the Montparnasse station, Faverges et al. 2018a compare state-of-the art distributions based on the Akaike Information criterion (AIC), and choose the negative binomial and the lognormal distributions to model delays.

3 Problem description

3.1 The platforming problem

The train platforming problem consists in routing trains through station and affecting them platforms. First solutions must be given months before operations, but adjustments can be done until a few days in advance. This problem is known to be NP-complete (Kroon, Romeijn, and Zwaneveld 1997). Finding solutions can be very challenging for main stations due to traffic density and a complex i nfrastructure. The train timetable is given, so arrival and departure time are fixed and solutions must satisfy commercial, security, resources and passenger flow c onstraints. This problem has been well studied with various approaches, for instance with MILP (Mixed Integer Linear Programming) formulation, constraint propagation or greedy heuristic (Sels et al. 2014).

SNCF Réseau, the french infrastructure manager, has recently developed a tool, OpenGOV, to solve the route and platform assignment problem at station. It is based on an extensive description of the station layout (platforms, paths, conflicts between resources) and the description of the different constraints. The problem is solved using MILP. Binary variables match trains with incoming path, platform and outgoing path. Two conflicting resources (crossing paths, tracks, platform, etc) cannot be affected to trains in the same time window whose size depends on the type of trains and the type of resources in conflict.

The case study is the Montparnasse station in Paris, France. This is a terminal station with 28 platforms and about 500 incoming and outgoing paths. There are about 700 sched-

uled trains per day, with suburban trains, high-speed trains and intercity trains. The models of infrastructure and operational constraints implemented in OpenGOV for the Montparnasse station are used in this study. Moreover, only passenger trains are considered for initial delay distribution and delay propagation. Indeed, other trains are more flexible and have a lower priority. They often experience variations in travel time (positive and negative delays) to adapt to other trains. Therefore, observation data of technical trains are unusable and dispatching rules are too complex to be modeled.

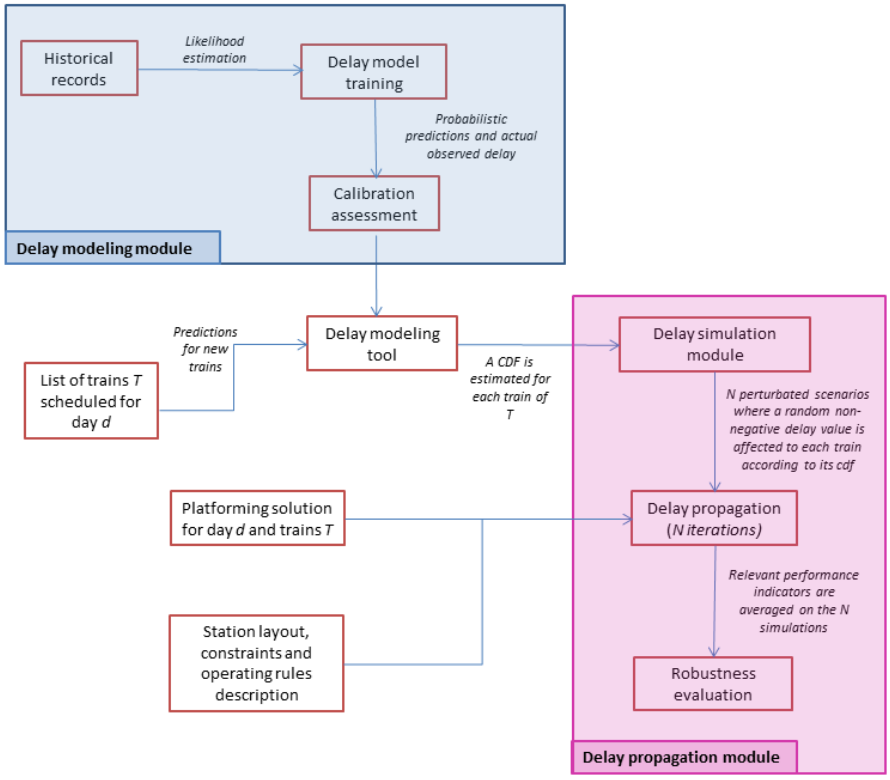
3.2 Proposed simulation methodology

This work takes advantage of the large amount of data collected on the network to build a calibrated and data-driven simulation framework. The methodology is summarized on the figure 2 . Historical records at a station are used to built a probabilistic model for initial delays. This model is then applied to simulate new delay samples (perturbations scenarios) for trains of a given day. N scenarios containing one delay value (non negative and often equal to zero) for each train of the day are obtained. For each of these scenarios, delays are propagated according to a platforming solution and given operating rules. The performance of the solution is then evaluated by averaging results of the delay propagation on the N iterations.

In this approach, the model for perturbations simulation is studied independently of the operating rules modeling, and in particular its adequacy is assessed before the sampling of delays. At the delay model training step, different methods can be tested. In this work, a new approach using Machine Learning is presented, enabling to model more precisely delay distributions by automatically estimating the influence of different context-related factors based on what happened in the past. Indeed, many researches have shown dependencies between the observed delay and the context, e.g train type, hour, line, infrastructure, timetabling, capacity consumption, etc (Olsson and Haugland 2004; Abril et al. 2008). Other state-of-the art delay models are tested as benchmark.

Calibration of the delay distribution is usually done a posteriori by comparing simulated and observed values, however a priori calibration has important benefits. At first, it allows to identify precisely and easily bias in the probability distribution, while with the classical methodology it is difficult to separates errors in the distribution and errors in operating rules when results do not match. Moreover, in the case of simulation for robustness evaluation at station, observations are not concordant with the hypothesis of the simulation framework: some trains experience extreme delays, others are cancelled or modified during operations (e.g new schedule) or capacity may be constrained (e.g limited infrastructure). Observations and simulations can not be compared directly as they do not always include the same events. For instance, if a train is cancelled during operations, it will free infrastructure and reduce the opportunity of conflicts for surrounding trains. However, the robustness must be evaluated by taking into account every scheduled train, and in the simulation framework, every train might experience a delay. If reality and simulation outputs are compared, the scheduled train might have an initial delay or be impacted by other trains, but there are no observations to relate with.

Figure 2: data-driven simulation methodology



The main contributions are the use of Machine Learning to simulate delays and a new a priori calibration assessment methodology that evaluates the quality of the delay distribution before delays are sampled for the simulation. Different delay models with variable quality are tested, and the comparison of these models aims to highlight the impact of the quality of the delay modeling part on performance results.

4 Delay probabilities estimation

4.1 Classical delay models

This paper presents four alternatives to model delay distribution.

The first alternative consists in simulating delay scenarios with a negative exponential distribution. This method is very simple and doesn't require data as the distribution parameter just need to be set at the inverse of the mean delay value, or any other approximation of it. The mean value of the dataset containing all passenger train delays in minutes, excluding outliers (negative values are set to 0 and delays greater than 20 minutes are deleted) is used. This method is not realistic at all as all type of events are expected to follow the same

distribution, but it represents correctly the general behavior of train delays (high probability of small values, skewness, etc).

The second approach is similar but different train profiles are studied. Initial delays are also generated with a negative exponential distribution but the distribution parameter is depending on the type of train (high-speed, suburban,..) and the type of event (arrival and departure). This method is relevant when there are no available data but known statistics on the mean delay value. In this case, the parameters are set to the inverse of the mean value of the corresponding dataset. Table 1 gives main characteristics of the different delay types; it must be noticed that the average delay varies a lot.

The third approach computes empirical distribution based on historical records. The set is divided according to train types and event. For each of these data sets, a discrete probability function is built with the relative frequency of every delay value. This approach requires a database with observed delays, and some features (train types, event), but the calculations are easy. It is more realistic than the other method because it is built on historical observations and separate different cases. However, it doesn't consider more precise separations (line, stopping pattern, density of the traffic, type of day, peak hours, etc). It is possible to increase the number of clusters in order to consider more parameters, but it might affect the precision of the estimated empirical distribution as there will be less elements in each cluster.

The last one uses generalized linear models and is described in the next section.

4.2 A statistical learning approach

The methodology for delay modeling with Machine Learning is explained more precisely by Faverges et al. 2018a. It is based on three main steps: datasets creation, model training and goodness-of-fit assessment.

This approach relies on statistical properties of delay data (choice of a modeling distribution) and on learning aspects. It aims to estimate individual delay probabilities at station by taking into account the potential impact of other features. Moreover, calibration of these probabilistic predictions is evaluated based on the predictions.

Data collection

Historical records of train delays associated with a location and scheduled event time are collected for trains arriving at and departing from Montparnasse station. A data base is created for every train type (high speed, suburban and regional) and event type (arrival and departure). Relevant indicators are added and encoded to obtain a numerical set (e.g. origin, date, stopping pattern, type of day, arrival time, trip duration, etc).

Outliers are excluded from data sets. In practice, delays above a threshold are deleted. There are several reasons for this. At first large delays are rare and unpredictable. They do not have the same causes as small delays and add noise in data. Secondary, this paper focus on robustness to small delays, and simulating large delays will not reflect reality as in real-life large disturbances require specific actions to minimize their impact. Third, the Machine Learning approach used here is based on a maximum likelihood estimation, there is no need to optimize parameter based on unlikely and irrelevant observations. The truncation threshold depends on the type of event and type of trains: for arrivals, suburban trains are usually cancelled when they experience delays above 10 minutes while high-speed trains and intercity are maintained. The Montparnasse station has a high rate of punctual

Table 1: Data sets description

Set	size	truncation threshold	Mean value	Main features
High- speed arrivals	25900	20	3.08	stopping pattern, scheduled stopping time, type of day, time slot, traffic density (on line, at origin and destination), rolling stock
High-speed departures	28700	10	0.48	type of day, time slot, destination, traffic density in station, rolling stock
Suburban arrivals	38900	10	1.03	stopping pattern, scheduled stopping time, type of day, hour, traffic density (on line, at origin and destination), rolling stock
Suburban departures	40600	5	0.18	type of day, hour, destination, traffic density in station, rolling stock, duration
Regional and Intercity arrivals	11400	15	2.26	Origin, type of day, time, traffic density, rolling stock
Regional and Intercity departures	11500	7	0.45	type of day, time, destination, traffic density in station, rolling stock

departure trains due to its terminal station status, so a low threshold is enough. Beside extreme delays, some trains arrive in advance, in particular the high-speed trains. In this model, observations with negative values are set to zero. This is a strong assumption, but at this point, negative values are more complex to model and predict, and they are less relevant than positive delays for the robustness evaluation. Indeed, if a train arrive in advance and create a conflict with another train at the station, it is expected that the early train can wait until its schedule time, without creating new delay. These negative delays are rare (they concern usually only high-speed arrivals) and with small value (one or two minutes).

The data sets are described in table 1. The mean value is estimated among the truncated non negative values recorded in minutes. These data are collected over a year (summer 2017 to summer 2018), and they exclude days of major system failures and following days of recover (13 days), major scheduled works (10 days) and strikes days (32 days). Features are similar in the different sets, but they are processed differently. For instance, time slot is in hour for suburban trains as they have a high frequency, but it is a few hours for high-speed trains, the stopping pattern and scheduled stopping time make sense only for arrivals and not departures, etc.

Finally, each of these sets is separated into two parts: a training set that is used to build a model and a validation set to assess its goodnessof-fit.

Model training

A generalized linear model (GLM) is trained on each of the training sets (high speed arrivals, high speed departures, regional arrivals and regional departures). GLM are convenient in this case as they model a variable with a probabilistic distribution Faverges et al. 2018b; Faverges et al. 2018a. The prediction for each train is not a single value but the probabilities corresponding to every possible outcome. Different train types are separated to improve models performances by reducing heterogeneity: travel time instability has not the same causes for these different cases, and the same features may have different impact.

The R package GAMLSS is used to implement these models. It extends classic GLM by allowing a large variety of probabilistic distributions, by modeling multiple parameters simultaneously and enabling to truncate distributions (Stasinopoulos, Rigby, et al. 2007). On delay data from the Montparnasse station in Paris, the truncated negative binomial distribution (NBI) is chosen (Faverges et al. 2018a). It is the best compromise between complexity (number of distribution parameters to fit) and likelihood of the model. The model is displayed bellow with $\boldsymbol{Y}$ the delays, $\boldsymbol{X}$ the covariate matrix and $\boldsymbol{\beta_\mu}$ and $\boldsymbol{\beta_\sigma}$ estimated parameters. $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ are the NBI distribution parameters.

$$\begin{cases} \boldsymbol{Y} & \sim \text{NBI}_{tr}(\boldsymbol{\mu}, \boldsymbol{\sigma}) \\ \ln(\boldsymbol{\mu}) & = \boldsymbol{X}\boldsymbol{\beta_\mu} \\ \ln(\boldsymbol{\sigma}) & = \boldsymbol{X}\boldsymbol{\beta_\sigma} \end{cases} \tag{1}$$

The figure 3 shows how the negative binomial distribution fits da ta. Observations are separated by type of train and events, and represented by the histograms. Parameters of the NBI are univariate maximum likelihood estimates of the true parameters, and corresponding probabilities are displayed with dots.

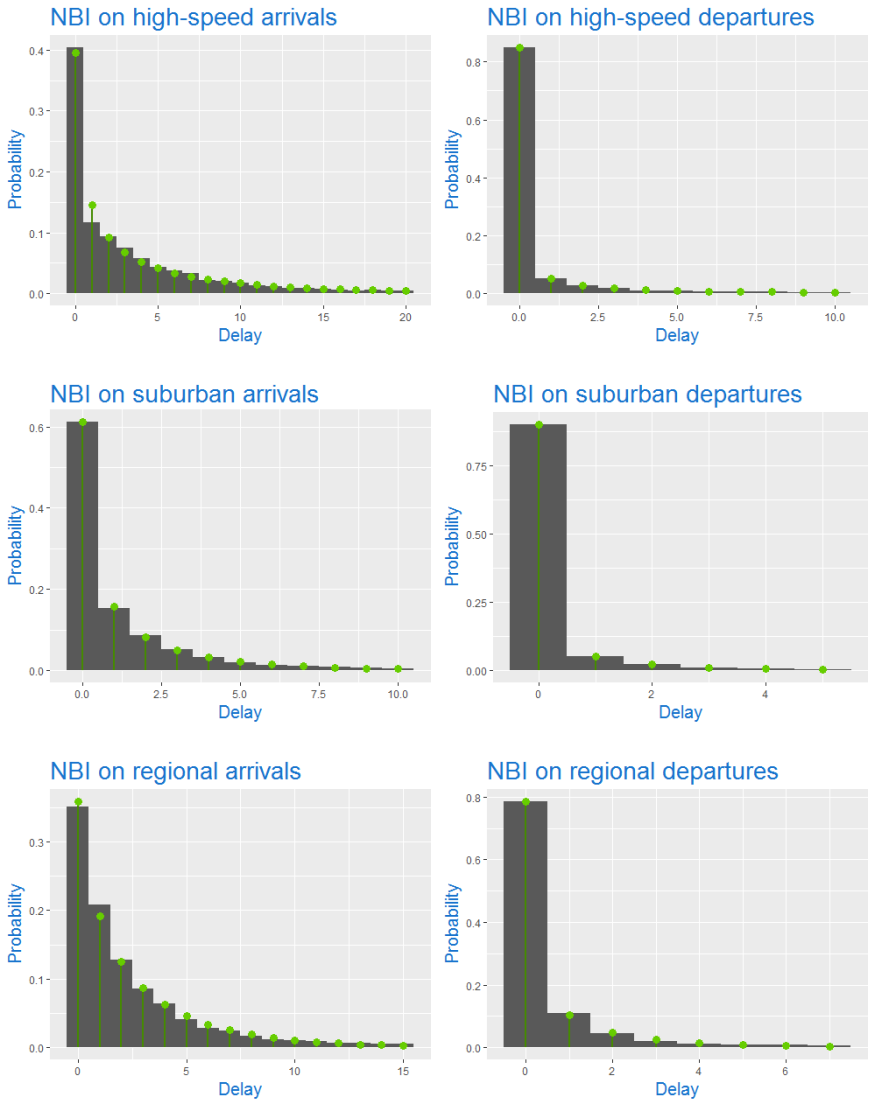
Model evaluation

As estimated individual probability mass functions are used to simulate delays, it is important to evaluate their quality and realism. However, usual residuals-based methods, like the mean absolute error, are not an option in this case because predictions and observations are not homogeneous (probability distribution and an integer).

This paper proposes to evaluate the quality of the model based on its calibration, ie the adequacy between estimated probabilities and observed rates of delay, at different points of the cumulative distribution functions. A graphical diagnostic of the calibration is done using grouping strategy: for a threshold given t, trains are sorted according to their estimated probability of having a delay higher than t and separated in g groups of similar predicted delay risk, then a calibration plot is obtained by displaying for each group the average estimated probability with the observed rate of delays in the group. A model can be considered calibrated when points are close to the diagonal.

This paper compares the different delay models only with calibration plots. An example of application of a statistical test to evaluate the significance of the deviations from the diagonal is given in Faverges et al. 2018a with the Hosmer-Lemeshow test. The plots are convenient as they are easily interpretable and allow to identify bias in the predictions (overestimation of risk for instance).

Figure 3: Negative binomial



5 Delay Propagation algorithm

This section presents the deterministic function that is applied on stochastic samples drawn from distributions computed previously. It aims to approximate the final delays based on the input scenario. These input delays (called *initial delays* here by opposition with *propagated* or *secondary* delays that are created in the station) are set to occur before the entrance of the station for arriving trains or at the platform for departing trains. This propagation function is an approximation of reality as during operations, many modifications are done on the original schedule (changes in the platform assignment, rolling stock management, human resources schedules, etc) based on human decisions.

For this preliminary study, a simple propagation algorithm is considered. It makes the strong assumptions that allocated paths are fixed and that the train sequence can be changed only if it is possible to maintain a train at its original schedule instead of delaying it. The goal is to study the reaction to delays without dispatching measures.

5.1 Station layout and constraints model

The infrastructure and constraints models are the ones implemented in the tool OpenGOV created by SNCF (cf section 3.1). There are two types of routes: arrival and departure, represented by an ordered succession of tracks. An arrival route is composed by an incoming track representing the entrance in the station and the beginning of a main line, three intermediate track sections and one platform track. A departure route has a platform track, three intermediate tracks and one outgoing track. Conflicting paths are defined as two paths that cannot be affected in a too short lapse of time, for instance if they share one or more tracks, or if they are crossing.

A solution of the platforming problem consists in the assignment of an arrival path and a departure path to each train. This assignment must respect rules, as described in section 3.1. The delay propagation algorithm has to take these constraints into account.

Minimal headways to respect between trains are set in OpenGOV according to the different cases of conflict and the station layout modeling: the type of train (high-speed, suburban, technical, etc), the type of event (arrival, departure, platform reoccupation, etc), the position of paths crossing (involved tracks), etc. The value associated with the different configurations corresponds to a security norm used for schedule conception. If a train is delayed, other trains on conflicting paths must wait the time necessary to ensure that the constraint is respected.

5.2 Algorithm

The algorithm is presented below. The notations are:

- $T = (t_1, ..., t_n)$ the list of trains sorted by schedule time $(h_{t_1}, ..., h_{t_n})$, and their corresponding simulated initial delays $(d_{prim,t_1}, ..., d_{prim,t_n})$
- The current delays $(d_{curr,t_1}, ..., d_{curr,t_n})$ correspond to the total delay of each train (initial and secondary). They are initialized at zero and then updated according to the delay propagation of other preceding trains and the initial delay of the corresponding train.

- For each ordered pair of trains (t, t') using conflicting paths, $cst_{t,t'}$ is the minimal headway to respect between the two trains. It depends on the type of train, the paths and the type of conflict
- For each $t \in T$, $CT_{prev(t)}$ is the list of previous conflicting trains with t, ie the list of trains t' that may impact t if they experience a delay higher than the scheduled buffer time $buffer_{t',t}$
- For each $t \in T$, $CT_{foll(t)}$ is the list of following conflicting trains, ie the list of trains t' that are impacted by t if t has a delay higher than the scheduled buffer time $buffer_{t,t'}$

The simulated initial delays are bounded by the truncation threshold, so they also produce bounded secondary delays. Moreover, only delays less than the maximal truncation threshold are considered to build $CT_{prev(t)}$ and $CT_{fol(t)}$.

In this simple algorithm, changes in the sequence are considered only if the train can be maintained at its original slot in order to cancel its secondary delay. These changes are possible only if they are compatible with all the trains originally scheduled before. For instance, two trains arriving at the station from the same track cannot be reordered, and it is naturally forbidden to exchange arrival and departure of the same train if it is delayed for the arrival and not the departure, but this is usually not a problem for a terminal station.

Algorithm 1 Propagation algorithm

Data: list of train $T = (t_1, ..., t_n)$ sorted by schedule time with their scheduled path, and their corresponding initial delays $(d_{t_1}, ..., d_{t_n})$

Result: Values of all secondary delays

initialization: Current delays are set to 0 $d_{curr,t} \leftarrow 0 \; \forall \; t \in T$

for $t \in T$ *sorted by schedule time* h_t **do**
  if $d_{curr,t} > d_{prim,t}$ **then**
    t has a secondary delay higher than its initial delay
    Test to verify if it is possible to maintain t at its original schedule time by changing the train sequence. It must be compatible with all the previous trains
    change = TRUE
    for $t' \in CT_{prev(t)}$ **do**
      if $h_{t'} + d_{curr,t'} < h_t + d_{prim,t}$ **then**
        if $h_{t'} + d_{curr,t'} + cst_{t',t} > h_t + d_{prim,t}$ **then**
          t' passes before t and the headway constraints is not fulfilled
          change = FALSE
        end
      else
        with its delay, t' passes after t. A change is the sequence may be possible
        if *t and t′ correspond to the arrival and departure of same train* **then**
          change = FALSE
        end
        if *t and t′ use the same platform or the same incoming track* **then**
          change = FALSE
        end
        if $h_t + d_{prim,t} + cst_{t,t'} > h'_t + d_{curr,t'}$ **then**
          the headway constraints is not fulfilled if t passes before t′
          change = FALSE
        end
      end
    end
    if *change = TRUE* **then**
      It is possible to change order of trains and maintain t at its original schedule with a potential initial delay but without secondary delays
      $d_{curr,t} \leftarrow d_{prim,t}$
    end
  else
    Current delay is set to initial delay
    $d_{curr,t} \leftarrow d_{prim,t}$
  end
  At this step, current delay of t is known. It is propagated to following trains
  for $t' \in CT_{foll(t)}$ **do**
    Secondary delay of t' is updated based on current delay of t
    $d_{curr,t'} \leftarrow \max\left(d_{curr,t'}, d_{curr,t} - \text{buffer}_{t,t'}\right)$
  end
end

6 Experiments

As described above, this paper presents four delay modeling alternatives for perturbations simulations. The differences between the results obtained by these approaches are studied by experimenting this methodology on a set of platforming solutions of the Montparnasse station. These solutions are the final schedules built by SNCF Reseau before operations. Four weeks are studied (the third week of the month of January, February, July and August). For the simulation part, 5000 iterations are done (delay simulation and propagation).

6.1 Differences between delay models

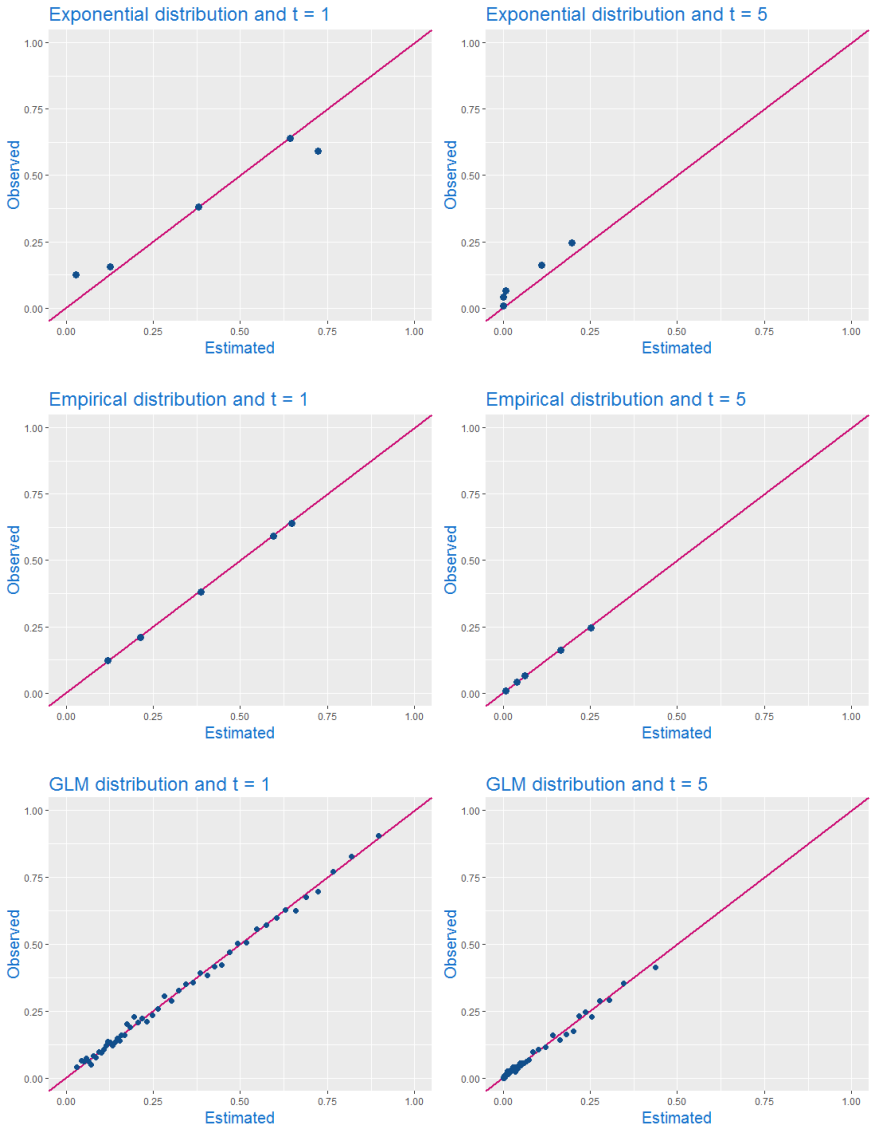
Three delays models are compared using calibration plots in Figure 4: the exponential modeling with one distribution per train and event type, the empirical distribution computed with historical records and the probabilities estimated with a GLM. The plots are build as described in subsection 4.2. A model is calibrated if points are close to the 45 degree: this means that estimated probabilities are concordant with observed delay frequency.

Two plots are displayed for each model: the first one to evaluate calibration of positive initial delay probability $\boldsymbol{P}(Y > 0)$ and the second one to evaluate the calibration of probabilities of delays greater than 5 minutes $\boldsymbol{P}(Y \geq 5)$.

For the exponential and the empirical models, there are only 6 groups possible as estimated probabilities are the same among different clusters *train type/event type*. The GLM model estimates delay probabilities using more features, so the range of predicted probabilities is larger and predictions are individual. It is visible on the graph since points modeling groups spread on the diagonal (50 groups are used). The model is more discriminant because it successfully recognize more punctual trains with low estimated probability from more delayed ones with higher estimated probability. It is also well calibrated.

The empirical model is very well calibrated as points are really close to the identity line. However, these probabilities are not precise, they only have a few values possible. The exponential model shows deviations between observations and estimations. In particular, it overestimates the large values of $\boldsymbol{P}(Y > 0)$ (points under the line) and underestimates small values (points over the line). $\boldsymbol{P}(Y \geq 5)$ is slightly underestimated for all clusters. Samples drawn with this model might differ from reality as certain trains are systematically more (or less) delayed that what is observed i dta.

Figure 4: Calibration plots



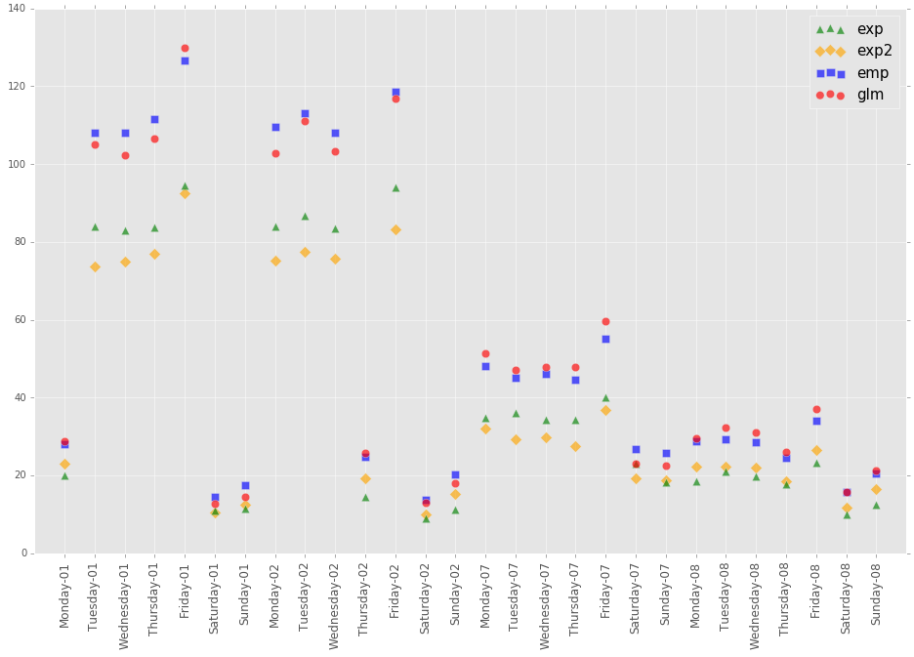
6.2 Propagation results

For each instance and each input delay scenario, two indicators are studied to compare the different initial delays modeling: the number of trains with a positive secondary delay created at station and the mean value of these positive secondary delays. They are computed after the propagation of the initial delays and averaged over the N iterations, considering

only passenger trains on four weeks of January, February, July and August. *exp* stands for the simplest exponential model with an unique delay distribution for all trains, *exp2* for the model with one distribution per cluster, *emp* for the empirical distributions per cluster based on time-stamps data and *GLM* for the learning model.

Results are given on the following plots.

Figure 5: number of secondary delayed passenger trains

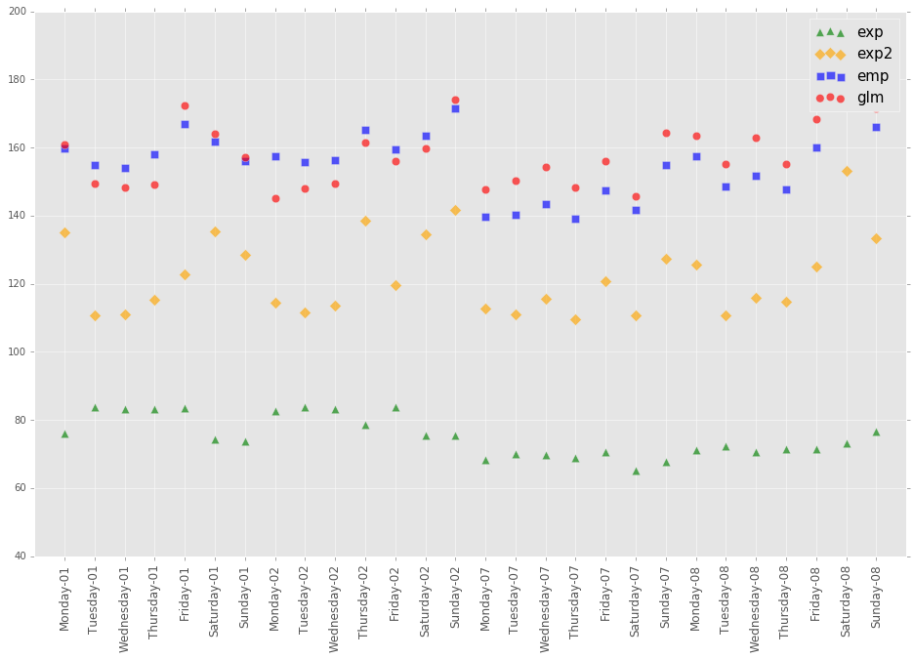


Considering the number of trains with positive secondary delay on Figure 5, all delay models provide highly correlated values with a redundant pattern of less delayed trains with the exponential models (especially *exp2*) and a larger number of secondary delayed trains with the *GLM* model.

Regarding the average positive secondary delays on Figure 6, strong deviations are observed between models. The *exp* model underestimates strongly the mean delay. This is not surprising, as arrivals delays are larger than departure delays in reality but modelled with the same distribution here, this model tends to simulate more initial delays with smaller values than the other models. They propagate less longer. Despite a correct calibration, the *exp2* model also shows important deviations with the other models, probably because it underestimates the probability of larger delays (cf Figure 4).

Finally, the average delay is close between the models *emp* and *glm* on all instances with differences of only a few seconds. They are both calibrated, and differences in discrimination doesn't have a visible impact on the average positive secondary delay.

Figure 6: Average secondary passenger train delay in seconds



7 Discussion

This work presents preliminary results on the perspective of delay modeling with Machine Learning to evaluate and improve robustness of operations at station. In particular, it focuses on the impact of calibration of delay modeling and expresses the difficulty to calibrate operating rules.

A priori calibration: it consists in assessing the goodness of fit of the perturbations distributions before the propagation algorithm. This approaches has several benefits:

- A posteriori calibration requires to compare results with actual observations that includes several outliers (large delays, but also cancelled trains that are not observed and may impact results). These outliers are not relevant for a robustness study, which focuses on small deviations of input parameters.
- A robust solutions must absorb delays with a limited use of dispatching. However, in reality, multiple changes are made on the schedule and the propagation is performed differently. In particular for the routing phase, alternative paths are preferred to propagation during operations. Calibrating probabilities based on results that are obtained with different processes may affect the results.
- Finally, using a priori calibration is promising to develop and test new delay propagation and dispatching strategies.

Delay propagation This paper also raises the issue of realism of propagation algorithm. This algorithm aims to represent real operations, but in this case of delay propagation in station, there are several limits to its realism:

- During operations, it is common to change path assignment, and even platform to avoid delay propagation. However, this is unrealistic to use such algorithm in a simulation framework as this is a complex decision problem. Moreover, it also must be avoided for a robustness study as a solution is not robust if infrastructure managers must perform multiple changes to the schedule in order to avoid delay propagation.
- The security constraints used in this study are sometimes too conservative. They correspond to conception constraints and must be respected when the schedule is conceived. However, in certain cases during operations, the required time between the two trains is less than the security constraint for conception, and the second train can pass before the constraint is respected.

Perspectives

- This methodology should be tested on platforming solutions with different level of robustness to evaluate more precisely the impact of input calibration. This paper shows that a bad calibration can lead to false magnitude in results. Working on same solutions of the same day would help to see the relative impact of the delay distribution when they are compared based on their robustness.
- In addition of delay distribution, more work should be done on propagation algorithm and operating rules calibration.
- Differences between delay distribution modeling should be studied at a more microscopic level, for instance to detect the differences due to systematic delays (trains that systematically experience secondary delays at station, useless buffer times,...). This will help to detect robustness defaults in solution.
- The delay modeling part could be improved with more precise data: delays are recorded in minute in this data set, this lack of precision add noise in results. Moreover, other modeling strategies than GLM should be tested for individual probabilities, like for instance Random Forests.

8 Conclusions

This paper presents a simulation methodology for robustness evaluation at station using statistical techniques (Generalized Linear models to estimate delay distributions according to the context and calibration plots to assess the goodness-of-fit) to provide a more realistic delay model that doesn't require a posteriori calibration. A robustness evaluation framework is used, characterized by truncated delay distributions and simple dispatching measures. A priori calibration is suitable in this case, as these assumptions do not totally reflect reality (large delays require specific dispatching that is not modeled here and trains are sometimes cancelled), comparing simulation results with observations to assess calibration might lead to bias.

The generalized linear model is compared with three other delay modeling techniques (two exponential models and one empirical distribution). The calibration assessment shows that the GLM and the empirical distribution are both well calibrated, unlike the exponential model that shows slight deviations. The GLM is also more precise and achieve to discriminate better the most punctual trains from the most delayed one while other models use the same probabilities among different clusters.

These modeling approaches are used for simulation of operations at station on 28 platforming problem solutions. Based on the number of trains experiencing secondary delays and the value of the average positive secondary delay, the empirical distributions and the GLM distributions give similar results while other models show strong deviations.

Acknowledgements

This work is conducted as part of a CIFRE PhD convention for an industrial agreement between SNCF Reseau and the CEDRIC laboratory - CNAM.

Authors thank Hajar Taleb for her precious help on OpenGOV tool and Rémi Parel for his knowledge on operations at Montparnasse station. Authors also wish to acknowledge the help provided by Antoine Robin, who worked on a preliminary version of the propagation algorithm during his internship at SNCF Réseau.

References

Abril, M et al. (2008). "An assessment of railway capacity". In: *Transportation Research Part E: Logistics and Transportation Review* 44.5, pp. 774–806.

Armstrong, John and John Preston (2017). "Capacity utilisation and performance at railway stations". In: *Journal of Rail Transport Planning & Management* 7.3, pp. 187–205.

Bergström, Anna and Niclas Krüger (2012). "Modeling Passenger Train Delay Distributions: Evidence and Implications for Valuation". In: *The 5th International Symposium on Transportation Network Reliability (INSTR2012), Hong Kong, China, December 18-19, 2012*. Pp. 61–61.

Briggs, Keith and Christian Beck (2007). "Modelling train delays with q-exponential functions". In: *Physica A: Statistical Mechanics and its Applications* 378.2, pp. 498–504.

Büker, Thorsten and Bernhard Seybold (2012). "Stochastic modelling of delay propagation in large networks". In: *Journal of Rail Transport Planning & Management* 2.1-2, pp. 34–50.

Caprara, Alberto et al. (2010). "Robust train routing and online re-scheduling". In: *OASIcs-OpenAccess Series in Informatics*. Vol. 14. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Carey, Malachy (1999). "Ex ante heuristic measures of schedule reliability". In: *Transportation Research Part B: Methodological* 33.7, pp. 473–494.

Carey, Malachy and Sinead Carville (2000). "Testing schedule performance and reliability for train stations". In: *Journal of the Operational Research Society* 51.6, pp. 666–682.

Cui, Yong, Ullrich Martin, and Weiting Zhao (2016). "Calibration of disturbance parameters in railway operational simulation based on reinforcement learning". In: *Journal of Rail Transport Planning & Management* 6.1, pp. 1–12.

Faverges, Marie Milliet de et al. (2018a). “Estimating Long-Term Delay Risk with Generalized Linear Models”. In: *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, pp. 2911–2916.

— (2018b). “Modelling passenger train arrival delays with Generalized Linear Models and its perspective for scheduling at main stations”. In: *ICRE*. IET.

Goverde, Rob MP (2005). “Punctuality of railway operations and timetable stability analysis”. In:

Harrod, Steven, Georgios Pournaras, and Bo Friis Nielsen (2018). “Distribution Fitting for Very Large Railway Delay Data Sets with Discrete Values”. In: *Trafikdage 2018*.

Koutsopoulos, Haris and Zhigao Wang (2007). “Simulation of urban rail operations: application framework”. In: *Transportation Research Record: Journal of the Transportation Research Board* 2006, pp. 84–91.

Kroon, Leo G, H Edwin Romeijn, and Peter J Zwaneveld (1997). “Routing trains through railway stations: complexity issues”. In: *European Journal of Operational Research* 98.3, pp. 485–498.

Landex, Alex and Otto Anker Nielsen (2006). “Simulation of disturbances and modelling of expected train passenger delays”. In: *Timetable Planning & Information Quality. Delft University of Technology, WIT Press Publishing, The Netherlands*, pp. 85–93.

Larsen, Rune et al. (2014). “Susceptibility of optimal train schedules to stochastic disturbances of process times”. In: *Flexible Services and Manufacturing Journal* 26.4, pp. 466–489.

Olsson, Nils OE and Hans Haugland (2004). “Influencing factors on train punctuality—results from some Norwegian studies”. In: *Transport policy* 11.4, pp. 387–397.

Sels, Peter et al. (2014). “The train platforming problem: The infrastructure management company perspective”. In: *Transportation Research Part B: Methodological* 61, pp. 55–72.

Stasinopoulos, D Mikis, Robert A Rigby, et al. (2007). “Generalized additive models for location scale and shape (GAMLSS) in R”. In: *Journal of Statistical Software* 23.7, pp. 1–46.

Wen, Chao et al. (2017). “Statistical investigation on train primary delay based on real records: evidence from Wuhan–Guangzhou HSR”. In: *International Journal of Rail Transportation* 5.3, pp. 170–189.

Yuan, Jianxin (2006). “Stochastic modelling of train delays and delay propagation in stations”. In:

Optimal Real-time Line Scheduling for Trains with Connected Driver Advice Systems

Ajini Galapitage [a,1], Amie R. Albrecht [a], Peter Pudney [a,b], Peng Zhou [a]
[a] Scheduling & Control Group, University of South Australia
[b] Future Industries Institute, University of South Australia
[1] E-mail: ajini.galapitage@mymail.unisa.edu.au

Abstract
On most rail networks, if a train is delayed then following trains will not know about the delay until they encounter a trackside signal that tells the driver that the next section of track is still occupied. The train will usually have to slow significantly, which causes delays to propagate back along the track. By using in-cab Driver Advice Systems connected to centralised scheduling systems, train delays can be detected as they happen, and new schedules can be calculated and issued to following trains so that additional delays are avoided. It is impossible to re-schedule the whole rail network at once in real time as the problem is too large. An alternative, more practical approach is micro-scheduling to independently optimise small sections of the network.
We describe and illustrate a method that can be used to ensure adequate and energy-efficient train separation. The method can be used during timetable planning to ensure robust timetables or can be used in real time to prevent trains from encountering restrictive signals, smoothing the flow of trains along a corridor.

Keywords
Optimal train control, dynamic rescheduling, line scheduling

1 Introduction

Energy-efficient driving strategies are often disrupted by train separation constraints, particularly when there are short time headways between trains and when some trains are delayed. When a train encounters a restrictive signal it will usually have to slow significantly, which disrupts efficient driving and introduces delays that can propagate back through the network.

Driver Advice Systems (DAS) can help trains follow a schedule precisely, and save energy at the same time [Scheepmaker et al., 2017, Panou et al., 2013, Albrecht et al., 2016a,b]. Connected Driver Advice Systems (C-DAS) extend this capability by adding communication with a central control system, which can provide real-time updates to individual train schedules in response to disruptions on the network.

Previous work has described how C-DAS can be used in real time to smooth the flow of trains through junctions, by adjusting the target arrival times of trains approaching the junction to avoid conflicts [Galapitage et al., 2018, Chen et al., 2015]. The on-board DAS ensures that the revised targets are achieved.

Luan et al. [2018a,b] discuss the integration of real-time traffic management and train control. Part 1 gives a good overview of various approaches, and develops mixed integer

programming solutions to the problem of determining optimal sequences, routes and arrival times for trains. Part 2 discusses optimal scheduling and energy efficiency, but use simplified speed profiles and assume constant gradient, curve and speed limits on each block section.

In this paper, we show how measurements of train movements can be used to identify locations and times where trains are delayed along a line without junctions, and we use examples from a long-haul freight line and from an intercity passenger line to show how small adjustments to train schedules can be used to ensure safe separation of trains while minimising energy use.

2 Measuring Train Delays

Many railways in the UK use the Energymiser[1] driver advice system, developed by Australian company TTG Transportation Technology and based on train control methods and software developed by the Scheduling and Control Group at the University of South Australia [Albrecht et al., 2016a,b]. As well as giving train drivers advice on how to drive efficiently, these units collect data that includes the position and speed of each train at 10-second intervals. This data can be used to analyse the performance of a railway. In this section we use journey logs from Chiltern Railways to investigate delays on the rail network. In particular, we use data collected on trains travelling from Princes Risborough to London Marylebone via High Wycombe during August 2016. Figure 1 shows two of the Chiltern routes to the west of London Marylebone. Princes Risborough is three stations south of Aylesbury, just south of a junction where trains from London can either head north to Aylesbury or continue west.

Our data from August 2016 includes 2172 "up" trips from Princes Risborough to London Marylebone via High Wycombe. Figure 2 shows the measured speed profiles of trains for the "up" direction, highlighting both the different stopping patterns and the considerable variations in speed.

2.1 Variation in Section Durations

Energymiser journey logs can be used to determine how long it took trains to drive between stops, and how much variation there was in these section durations.

Table 1 shows the durations of stop-to-stop journey sections for the measured train journeys. The columns are:

- the origin of the trip section
- the destination of the trip section, which is not necessarily the next station along the route
- the number of trips that did this section
- the median section duration, in seconds; half the trips had a section duration less than this value, and half had a section duration greater than this value
- the first quartile section duration, in seconds; one quarter of the trips had a section duration less than this value

[1] `http://www.ttgtransportationtechnology.com/energymiser`

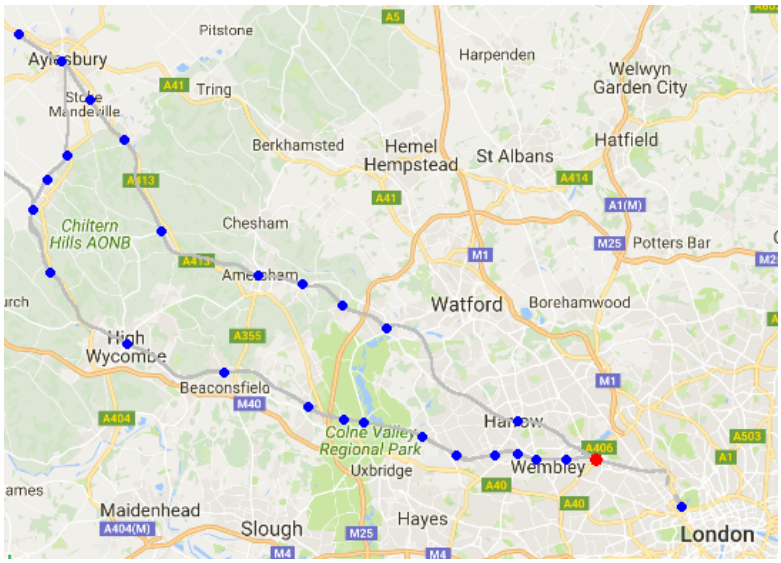


Figure 1: The Chiltern rail network west of London Marylebone. The background map is from Google.

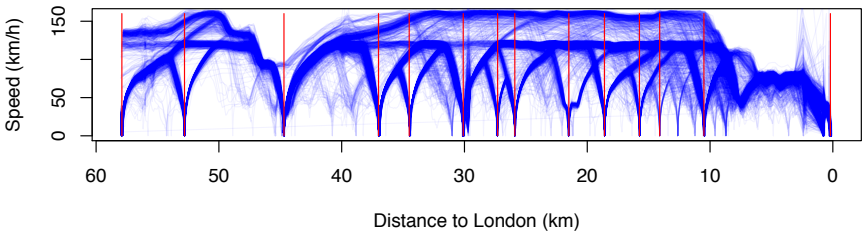


Figure 2: Speed profiles of trains travelling from Princes Risborough to London Marylebone via High Wycombe, August 2016.

- the third quartile section duration, in seconds; one quarter of the trips had a section duration greater than this value
- the inter-quartile range (IQR), in seconds; this is the difference between the third quartile value and the first quartile value, and is a measure of the variation in section durations.

Sections with wide variation, where the duration IQR is more than 10% of the median section duration, are indicated with a '*'.

Some of the variation in section running durations could be due to variations in driving advice provided by the Energymiser units. However, many trains run late; for these trains, the advice is to drive as quickly as possible and so the variations are due to other factors.

2.2 Slowing Between Scheduled Stops

We are particularly interested in times and locations where trains are slowed between stops by the signalling systems. We did not have access to signalling data, but if a train slows significantly between scheduled stops then it is almost certainly because of traffic issues.

Figures 3 and 4 show times and locations where trains travelling in the "up" direction slowed to less than 40 km/h. The darker dots indicate speeds less than 20 km/h. The horizontal black lines represent train station locations. We expect trains to travel slowly when arriving at a stop and departing from a stop, but low speeds away from stops indicate a traffic problem.

The delays at 8 km are near Neasden Junction; trains travelling in the "up" direction through the junction are often delayed by train movements on other paths through the junction. Galapitage et al. [2018] describe a method for real-time rescheduling of trains at junctions.

There were several other sections where trains slowed to less than 40 km/h. There are no junctions on these route sections.

3 Line Scheduling

A train following another train along a track will be delayed if it gets too close to the leading train. Once a train has encountered a restrictive signal, it will have to slow; this can introduce further delays on the corridor.

In this section we describe how small adjustments to individual train schedules can be used to ensure adequate separation between trains to avoid encounters with restrictive signals. The method can be used in the timetable planning stage to develop robust timetables, or in real time to ensure smooth running on a corridor.

We will illustrate the method using four simulated but realistic examples.

3.1 Example 1: Long-haul Freight

The Dedicated Fast Freight Corporation in India is building two new rail corridors, in the east and west of the country. These corridors will each carry a mix of freight train types over long distances, with headways between trains as low as six minutes. Crew change locations are fixed on each corridor. To maximise capacity, the running time between any given pair of adjacent crew change locations will be the same for every train. However, differences

Table 1: Section durations for trains travelling from Princes Risborough to London Marylebone via High Wycombe, August 2016.

			section duration				
origin	destination	trips	median	Q1	Q3	IQR	
Princes Risborough	Saunderton	433	203	197	211	14	
Princes Risborough	High Wycombe	466	454	436	465	30	
Saunderton	High Wycombe	430	279	273	284	11	
High Wycombe	Beaconsfield	1207	270	262	281	19	
High Wycombe	Seer Green & Jordans	25	338	330	347	17	
High Wycombe	Gerrards Cross	72	490	477	504	28	
High Wycombe	West Ruislip	11	745	728	774	46	
High Wycombe	Wembley Stadium	75	943	884	1066	182	*
High Wycombe	London Marylebone	165	1545	1440	1679	239	*
Beaconsfield	Seer Green & Jordans	524	100	96	106	10	*
Beaconsfield	Gerrards Cross	644	247	241	254	13	
Beaconsfield	London Marylebone	16	1247	1174	1347	173	*
Seer Green & Jordans	Gerrards Cross	547	162	157	169	12	
Gerrards Cross	Denham Golf Club	197	99	93	102	9	
Gerrards Cross	Denham	208	145	139	149	10	
Gerrards Cross	West Ruislip	86	289	280	302	22	
Gerrards Cross	South Ruislip	223	362	354	369	16	
Gerrards Cross	Wembley Stadium	166	613	604	628	24	
Gerrards Cross	London Marylebone	221	1180	1119	1257	138	*
Denham Golf Club	Denham	196	51	47	54	7	*
Denham	West Ruislip	120	179	174	184	10	
Denham	South Ruislip	173	251	247	257	10	
Denham	Wembley Stadium	64	502	495	510	15	
Denham	London Marylebone	28	1041	994	1062	68	
West Ruislip	South Ruislip	68	145	140	150	11	
West Ruislip	Northolt Park	63	241	233	247	15	
West Ruislip	Sudbury Hill Harrow	13	297	295	315	20	
West Ruislip	Wembley Stadium	63	405	392	417	26	
South Ruislip	Northolt Park	144	119	116	123	7	
South Ruislip	Wembley Stadium	220	297	290	309	19	
South Ruislip	London Marylebone	49	841	817	871	54	
Northolt Park	Sudbury Hill Harrow	24	69	67	73	6	
Northolt Park	Wembley Stadium	154	198	191	207	16	
Northolt Park	London Marylebone	11	705	699	816	117	*
Sudbury Hill Harrow	Wembley Stadium	12	137	136	148	12	
Sudbury Hill Harrow	London Marylebone	12	841	814	897	83	
Wembley Stadium	London Marylebone	409	563	523	610	87	*

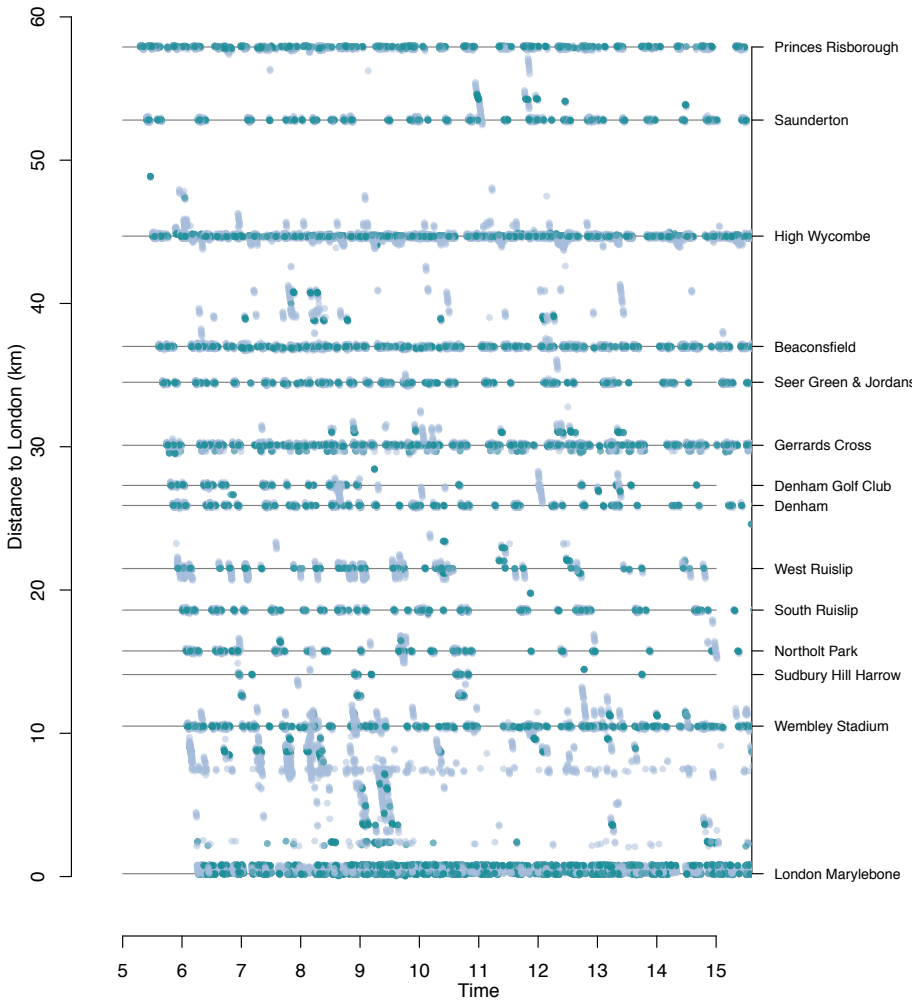


Figure 3: Times and locations where up trains slowed to less than 40 km/h, between 05:00 and 15:00. The darker dots indicate speeds less than 20 km/h.

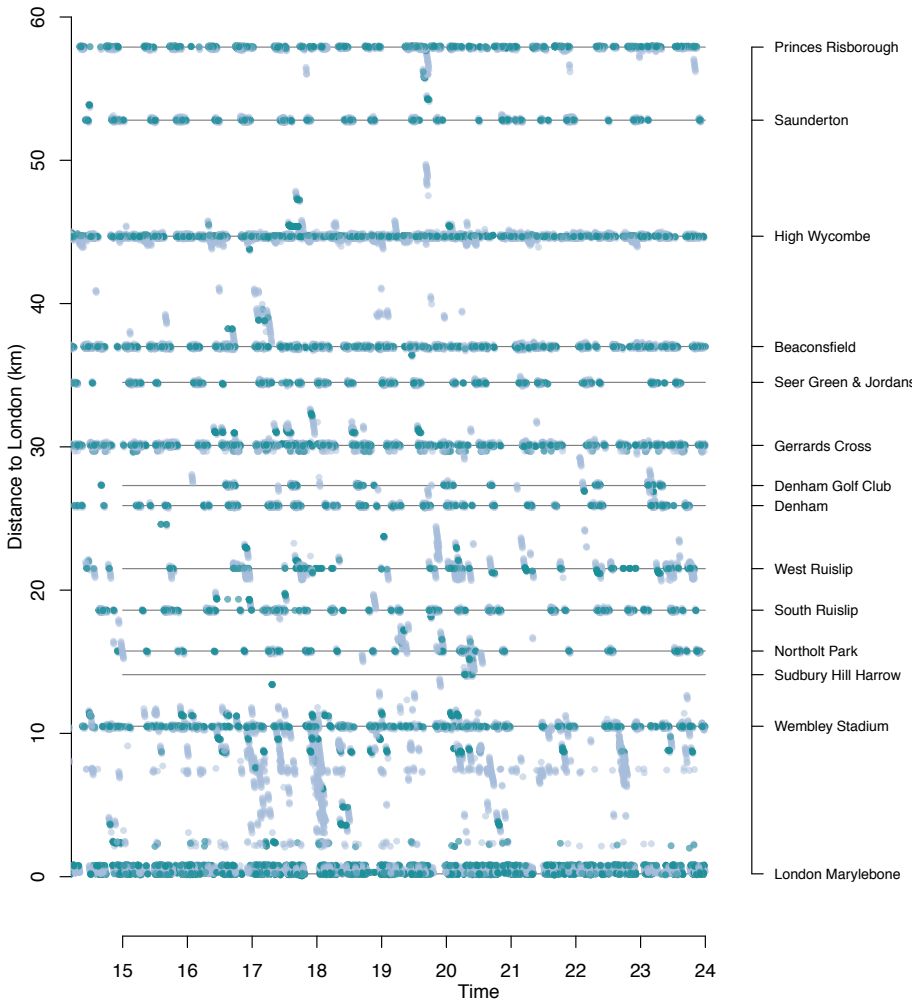


Figure 4: Times and locations where up trains slowed to less than 40 km/h, between 15:00 and 00:00. The darker dots indicate speeds less than 20 km/h.

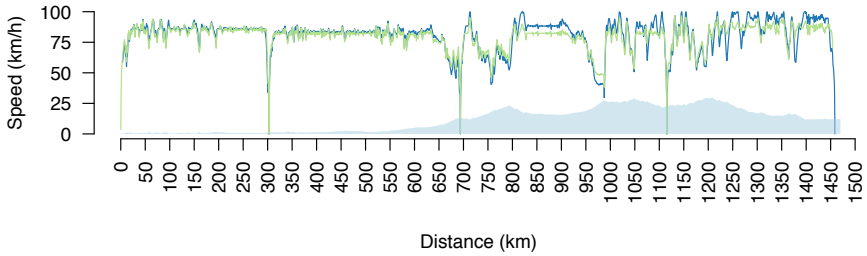


Figure 5: Speed profiles for a bulk train (blue) and a container train (green).

in train performance will mean that the separation between trains will vary as trains drive between crew change locations. For example, a heavy train may be slowed on hills more than a following light train.

The track will have three-aspect signalling with a spacing of 1.5 km. This means that a following train must be at least 3 km behind a leading train, otherwise it will encounter a yellow signal and have to slow.

The separation between two trains traveling along a line depends on the relative speeds of the two trains, which in turn depends on the locomotive performance and trailing load of each of the trains, and on the gradients, curves and speed limits. Figure 5 shows energy-efficient speed profiles for a bulk train (blue) with a trailing mass of 6500 tonnes and for a container train (green) with a trailing mass of 4500 tonnes. The optimal speed profiles were calculated using our Energymiser software. The shaded region at the bottom of the graph indicates the elevation profile of the track. Both trains have the same section running time for each of the four sections of the trip, but the speed profiles are different because of the different train characteristics. For example, the laden bulk train is slowed more by the hills near 985 km than the lighter container train, and so has to travel faster elsewhere in the journey to make up time.

Trains will normally follow each other with a headway of six minutes at each stop. Figure 6 shows the two journey paths with time on the horizontal axis. The container train starts six minutes behind the bulk train.

Figure 8 shows speed profiles $v_1(t)$ and $v_2(t)$ for the bulk train and container train on the third section of the route. Each speed profile has been optimised independently to meet the overall section duration of 5H45M with minimum energy. The heavier bulk train is slowed more by the hills than the lighter container train.

The distance between the two trains at any time is the train *separation*. Figure 7 shows separation as a function of time. We can see from Figure 7 that the trains are too close near times 03:53, 09:01, 10:45, 13:16 and 15:01. The low separation near times 03:53, 09:01 and 15:01 occur because the leading train is stopping for crew changes, and the following train catches up while the leading train is slowing to a stop. In these situations we can allow the following train to get close because it is also going to stop at these locations, and there is space at the crew change locations for more than one train. We are more interested in low separations that occur between stops, at times 10:45 and 13:16.

One way to prevent a following train from getting too close to a leading train is to insert timing points for the following train that will slow it at certain places on the track. The lowest separation occurs at time 10:45, where the distance between the trains is 1438 m.

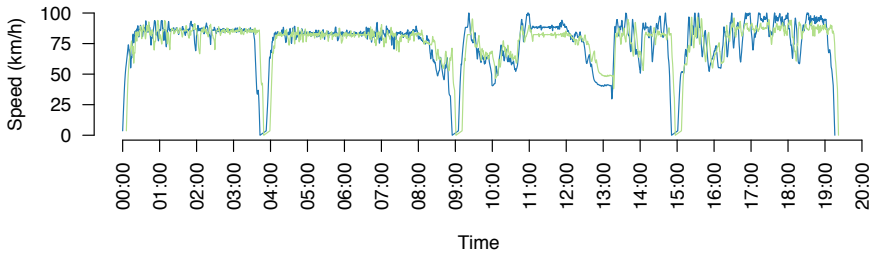


Figure 6: Speed against time for the bulk train and the container train.

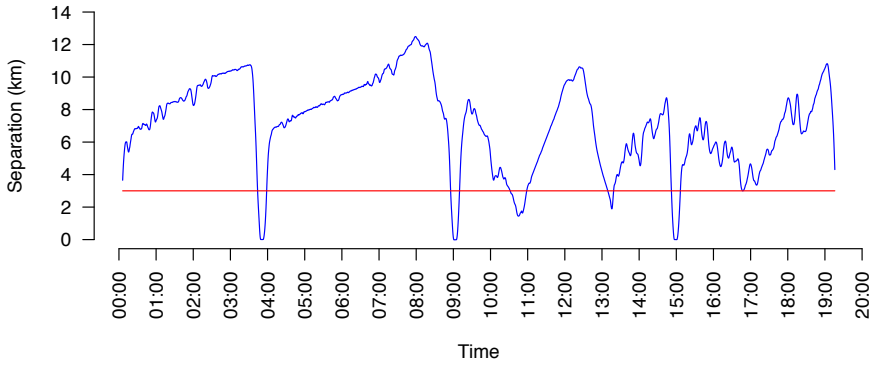


Figure 7: Separation between the two freight trains.

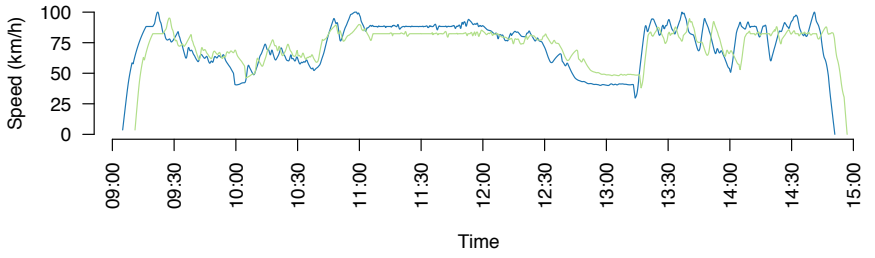


Figure 8: Speed profiles of the two freight trains on the third journey section.

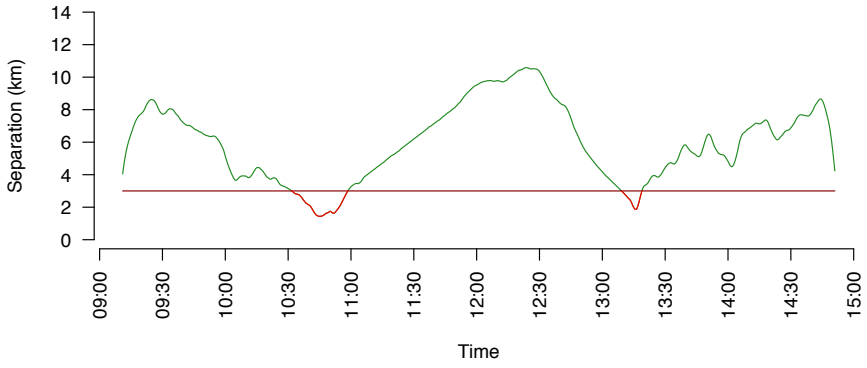


Figure 9: Separation between the two freight trains on the third journey section.

The minimum separation required is 3 km. Let $x_1(t)$ and $x_2(t)$ be the locations of the leading and following trains at time t. The separation between the trains at time t is $\delta(t) = x_1(t) - x_2(t)$, and the rate of change of this separation is $\delta'(t) = v_1(t) - v_2(t)$. Figure 9 shows the separation δ. The green regions are where the separation between the trains is less than 3 km.

One approach to resolving the separation violations is to place a timing constraint at times with minimum separation. In each of the regions with low separation, we search for the time $\tau = \arg\min_t \delta(t)$ at which the separation between the two trains is minimum, then set a timing constraint

$$t_2^*(x_1(\tau) - h) \geq \tau$$

for the rescheduled Train 2, where h is the minimum allowable distance between the trains. This constraint ensures that, at time τ, Train 2 will not have passed the location that is distance h behind the location of Train 1. The path of the rescheduled Train 2 is described by the distance profile x_2^* and the speed profile v_2^*.

The original train paths have minimum separation of 1.438 km at time $\tau =$ 10:45:10, with zero derivative. The new profile for Train 2 has $x_2^*(\tau) = x_1(\tau) - h$, because the timing constraint is active, and $v_2^*(\tau) \leq v_2(\tau)$, because Train 2 is now travelling slower at time τ. The new separation is

$$\delta^*(\tau) = x_1(\tau) - x_2^*(\tau) = h$$

with

$$\delta^{*'}(\tau) = v_1(\tau) - v_2^*(\tau) \geq v_1(\tau) - v_2(\tau) = 0$$

and so $\delta^*(\tau - \epsilon) \leq h$ for small ϵ; that is, the new separation is slightly less than h immediately prior to time τ. Figure 10 shows more detail around time τ.

The separation constraint is still violated after introducing a single timing constraint. To resolve this, instead of adding one timing point at the minimum separation point we can add timing constraints throughout the journey. In practice, we add timing constraints at closely spaced discrete points in regions where the minimum separation dips below 3 km. We use

$$t_2^*(x_1(k\Delta t) - h) \geq k\Delta t, \qquad k \in \{0, 1, \ldots\},\ \delta(k\Delta t) < h$$

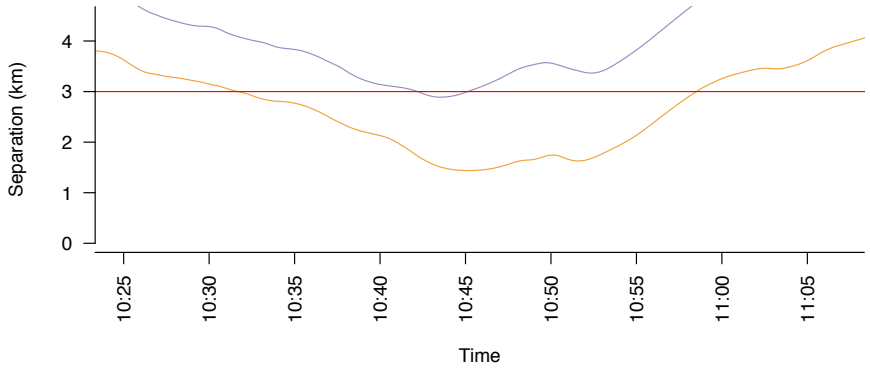


Figure 10: Detail of the original freight train separation δ (orange), and separation after rescheduling the second train (purple), around time τ.

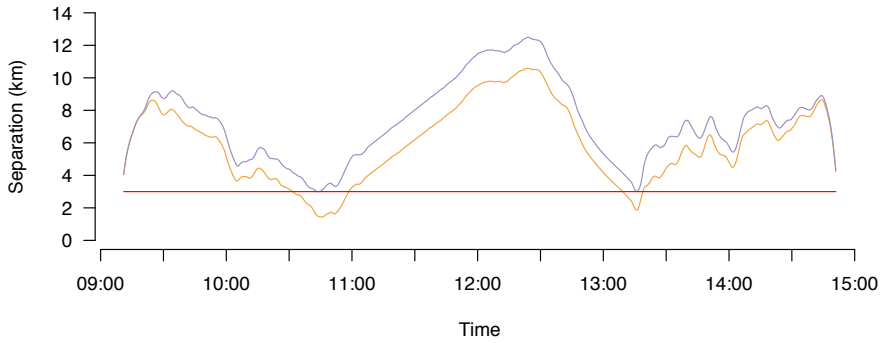


Figure 11: The original freight train separation δ (orange), and separation after rescheduling the second train (purple).

where Δt is a time step chosen to suit the problem; in this case, we used $\Delta t = 25$ seconds. Figure 11 shows the original separation δ (orange), and separation after rescheduling the second train (purple) to increase the separation throughout both regions where the separation drops below 3 km.

The rescheduled container train is still able to finish its journey on time. We expect it to use more energy, since its path has been constrained. In this case, the extra energy use is negligible—just 0.0133% more than without the extra timing constraint.

The particular example does not require a trade-off between speeding up the leading train and slowing down the following train, as suggested by Albrecht et al. [2018].

3.2 Example 2: Express from London

In this next example, we simulate the motion of two express passenger trains running from London Marylebone to Princes Risborough. This is a 58 km journey taking 25 minutes. The

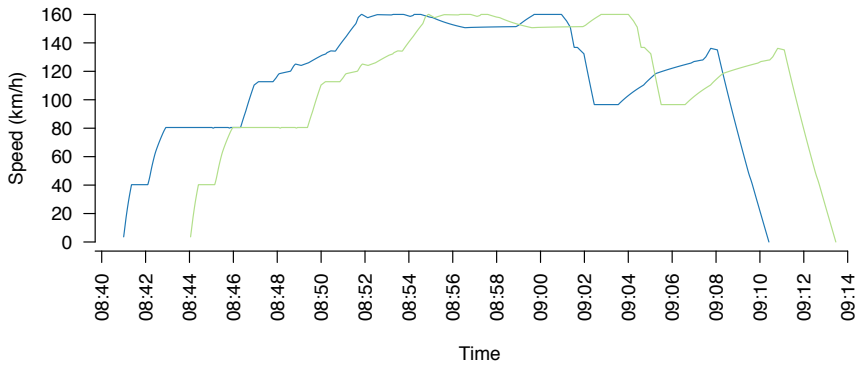


Figure 12: Speed profiles for London – Princes Risborough trains.

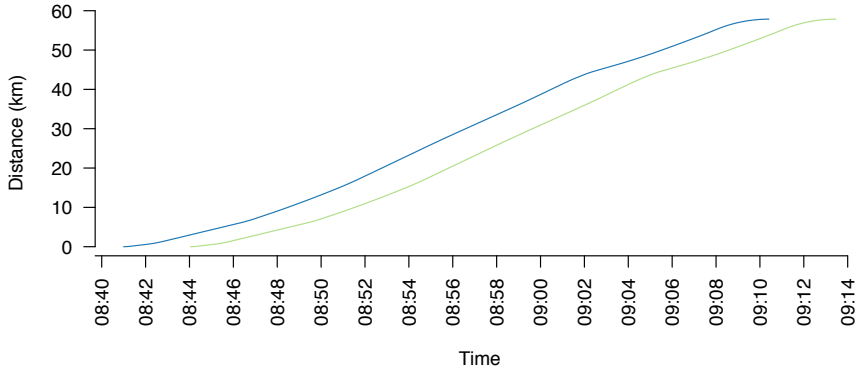


Figure 13: Train graph with two identical passenger trains running express from London.

two trains have identical characteristics and identical optimal journey profiles. Figure 12 shows the optimal speed profiles.

As with the previous example, we assume that the required minimum separation is 3 km. We start the second train as soon as the first train has travelled 3 km, to maximise the likelihood of interaction between the trains. Figure 13 shows the distance that each train has travelled at a given time.

Figure 14 shows the separation between the two trains. The relatively low speed limits leaving London mean that the first train speeds up while the second train is still travelling slowly. This increases the separation between the trains, and the separation remains above the critical 3 km for the remainder of the journey. There is no need to intervene.

3.3 Example 3: Approaching London

Our example trains travelling away from London never got too close due to the initial speed limits. Next we simulate two trains running towards London, from Wembley Stadium to

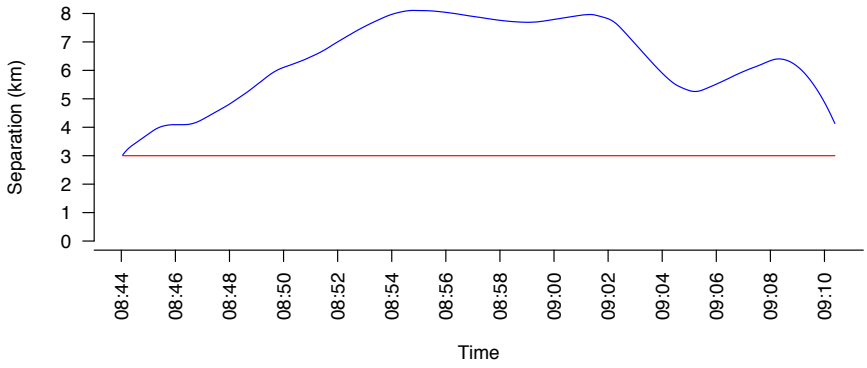


Figure 14: Separation of the two trains from London.

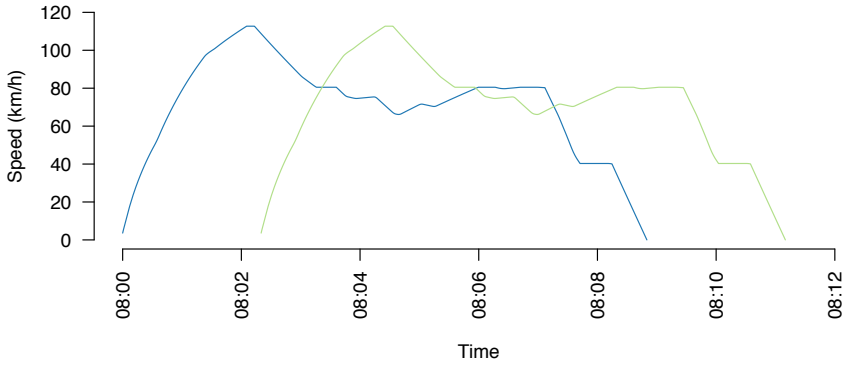


Figure 15: Optimal speed profiles of two trains approaching London.

London Marylebone. This is a 10.5 km journey which takes just under 9 minutes. Once again, we start Train 2 as soon as Train 1 has travelled 3 km. Figure 15 shows the optimal speed profiles of the two trains when separation is not considered.

Figure 16 shows the separation between the two trains (orange). The two trains are closest together at 08:08:49 when Train 1 is arriving at London Marylebone. We add timing constraints for Train 2 to keep the 3 km separation from Train 1, using the same method as in our first example. Figure 16 shows the separation after adding the timing constraints for Train 2 (purple).

The rescheduled Train 2 finishes its journey 48 seconds late, and consumes 1.76% less energy than the original journey.

If we want Train 2 to arrive on time then we need to speed up Train 1. So next we run the Train 1 fast as possible and adjust Train 2 to meet the minimum separation requirement. Figure 19 shows the separation after adding timing constraints for both trains (purple). After making Train 1 as fast as possible, it arrives 29 s early and Train 2 is still 17 s late at the destination. Together, the trains use 23% more energy than the optimised journeys. Because of the low speed limits near the end of the journey, it is not possible to meet the separation

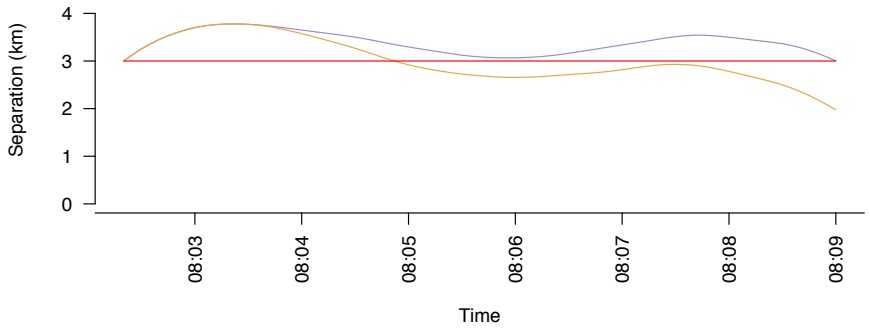


Figure 16: The original separation (orange) for the two trains approaching London, and separation after rescheduling Train 2 (purple).

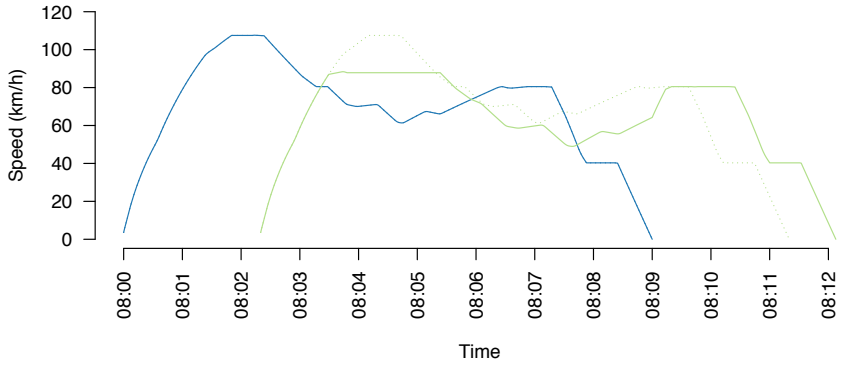


Figure 17: Original speed profiles of the two trains approaching London (blue and dotted green lines) and speed profile of the rescheduled second train (green).

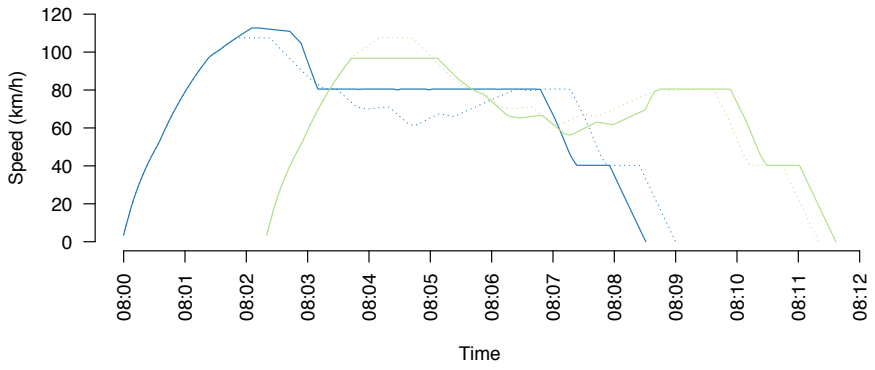


Figure 18: Original speed profiles of the two trains approaching London (dotted lines) and speed profiles of the rescheduled trains (blue and green).

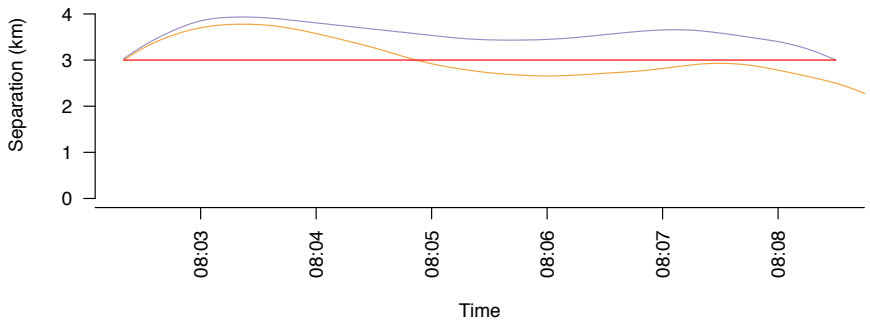


Figure 19: The original separation of the two trains approaching London (orange), and separation after rescheduling both trains (purple).

constraints without changing the time between arrivals at London Marylebone.

3.4 Example 4: Mid-journey Speed Restriction

None of the examples so far demonstrate a scenario where both trains arrive on time and the optimal strategy is a compromise between speeding up Train 1 and slowing Train 2. Our final example does this, using a scenario where the minimum separation occurs in the middle of the journey.

In this example, we simulate two trains running from London Marylebone to Princes Risborough, which is a 57.86 km journey with a duration of 36 minutes. The trains get closer together in the middle of the journey as they encounter a low speed limit that we have imposed to demonstrate the principle.

Train 1 starts its journey at 08:41:00 and finishes at 09:17:00. Train 2 starts and finishes three minutes after Train 1. Each train consumes 630 MJ energy. Figure 20 shows the speed profiles of the two trains, and figure 21 shows the separation between the two trains.

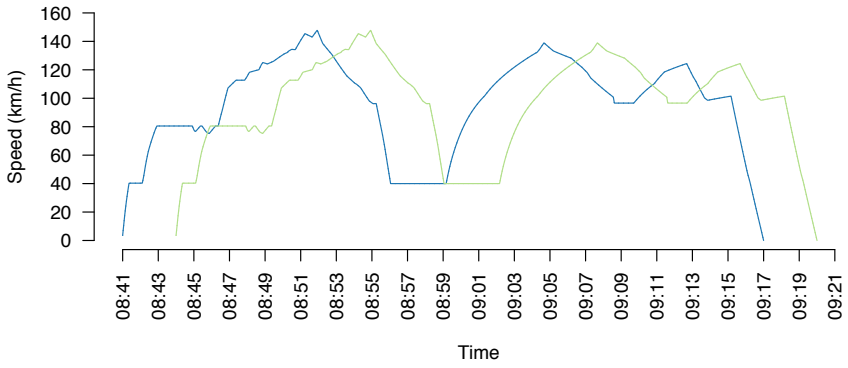


Figure 20: Original speed profiles of the two trains with a mid-section speed restriction.

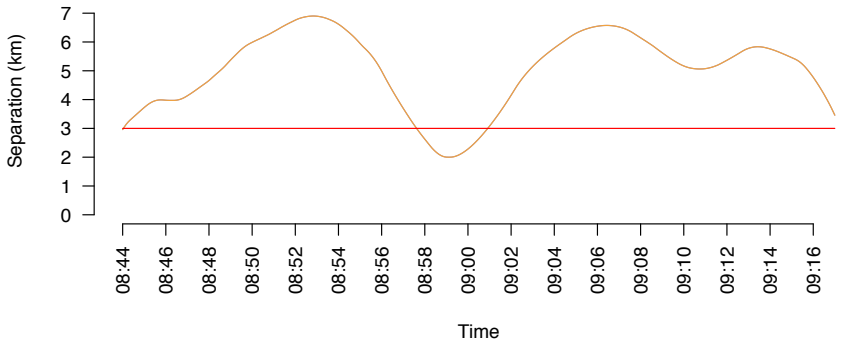


Figure 21: Original separation of the two trains with a mid-section speed restriction.

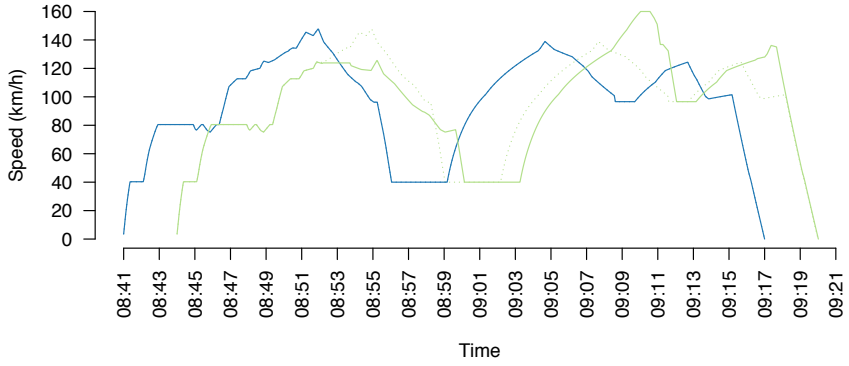


Figure 22: Speed profiles of the two trains with a mid-section speed restriction, after rescheduling Train 2.

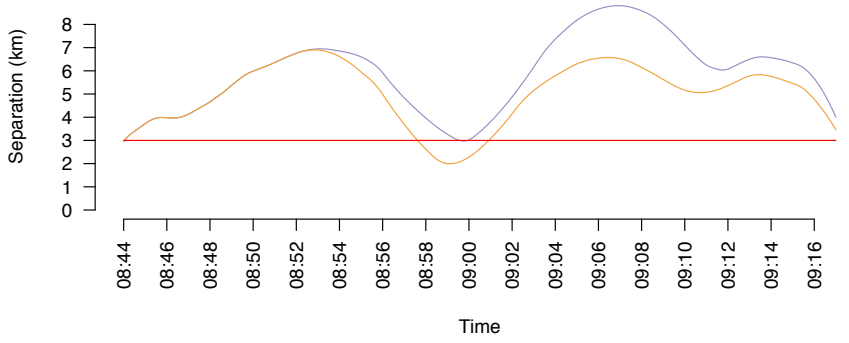


Figure 23: The original separation of the two trains with a mid-section speed speed restriction (orange), and separation after rescheduling Train 2 (purple).

The minimum separation occurs at 08:59:05 when Train 1 is at 27.022 km and Train 2 is at 25.022 km. Since the minimum separation occurs near the middle of the journeys, we can change the speed profiles of either train without compromising their ability to finish on time. We can either speed up the first train or slow down the second train to meet the separation requirement, or do a combination of both.

Slowing Down the Second Train

First, we simulate the optimal journey for Train 1 and slow down Train 2 to meet the separation constraint near 08:59:05. Figure 22 shows the speed profiles of the two trains after adding a timing point for Train 2. The dotted green line represents the original speed profile of Train 2. Both trains still arrive at the destination on time, but together consume 6.8% more energy than the original optimal journeys. Figure 23 shows the separation before and after rescheduling Train 2.

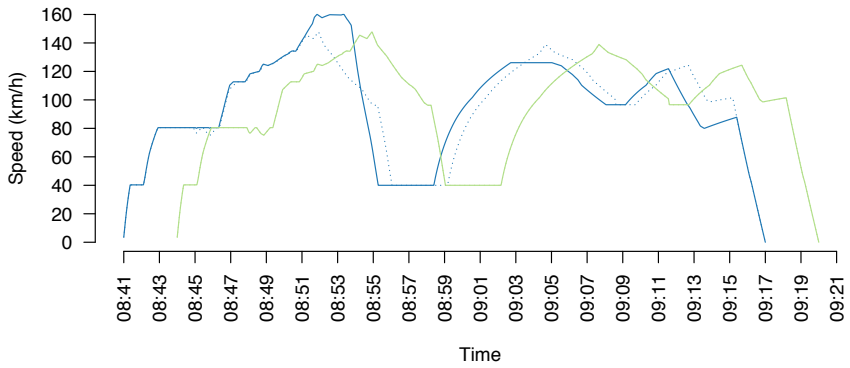


Figure 24: Speed profiles of the two trains with a mid-section speed restriction, after rescheduling Train 1.

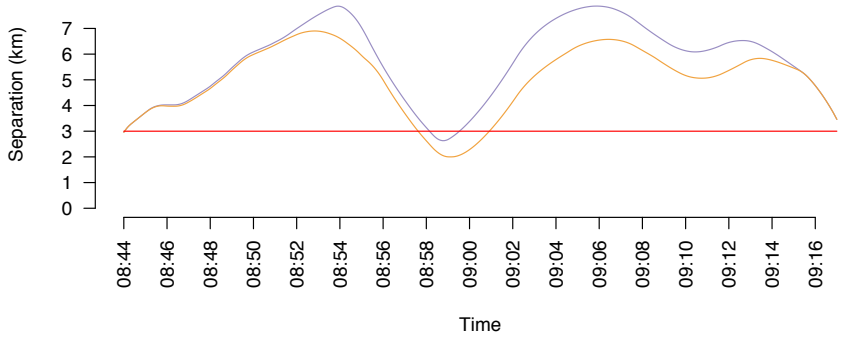


Figure 25: The original separation of the two trains with a mid-section speed restriction (orange), and separation after rescheduling Train 1 (purple).

Speeding Up the First Train

Next, we simulate the optimal journey for Train 2 and speed up Train 1 to meet the separation constraint near 08:59:05. Figure 24 shows the speed profiles of the two trains after adding a timing point for Train 1. The dotted blue line indicates the original speed profile of Train 1. Both trains still arrive at the destination on time, but together consume 5.1% more energy. Figure 25 shows the separation before and after rescheduling Train 1.

Adding Timing Points to Both Trains

We can find the optimal compromise between slowing Train 2 and speeding up Train 1 by imposing a latest arrival time τ for Train 1 at $x_1 = 28.022$ km, and then driving Train 2 to avoid getting too close to Train 1.

The earliest that Train 1 could arrive at the timing point x_1 is $\tau =$ 08:59:20 and the latest it could arrive at the timing point is $\tau =$ 09:00:06. We vary the time τ in this interval then run Train 1 with this constraint and Train 2 to avoid Train 1. We calculate the total energy

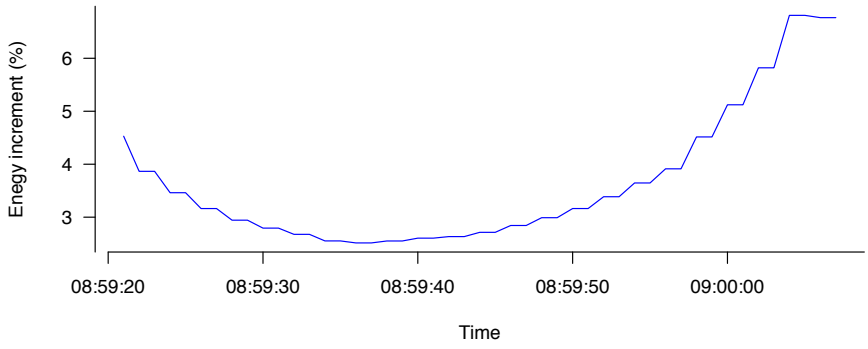


Figure 26: Overall energy increment for the two journeys with a mid-section speed limit, for different values of the timing constraint τ.

consumption for each τ. Figure 26 shows the total energy increment for each value of τ.

The graph is "lumpy" because of the numerical precision of the Energymiser software used to calculate optimal journeys. Nevertheless, the graph shows that we can meet the separation constraints and minimise the overall energy use by setting $\tau \approx$ 08:59:37.

4 Conclusion

Trains will be delayed if they get too close to the train ahead. These types of delay can be reduced by designing robust timetables with adequate train separation, and then by using Driver Advice Systems to ensure that trains are driven to the timetable.

Nevertheless, when a train is delayed, the delay can propagate to following trains if they encounter restrictive signals, which introduces further delay.

We have described a method that can be used while planning timetables to ensure adequate separation between trains, but also in real time to make small adjustments to individual train schedules so that restrictive signals are avoided. The method can simply adjust the schedules of following trains to maintain the required separation, or can find the energy-optimal trade-off between speeding up a leading train and slowing down a following train.

References

A. Albrecht, P. Howlett, P. Pudney, X. Vu, and P. Zhou. The key principles of optimal train control-Part 1: Formulation of the model, strategies of optimal type, evolutionary lines, location of optimal switching points. *Transportation Research Part B: Methodological*, 94:482–508, 2016a.

A. Albrecht, P. Howlett, P. Pudney, X. Vu, and P. Zhou. The key principles of optimal train control—Part 2: Existence of an optimal strategy, the local energy minimization principle, uniqueness, computational techniques. *Transportation Research Part B: Methodological*, 94:509–538, 2016b. ISSN 0191-2615. doi: http://dx.doi.org/10.1016/j.trb.2015.07.024.

A. Albrecht, P. Howlett, P. Pudney, X. Vu, and P. Zhou. The two-train separation problem on non-level track—driving strategies that minimize total required tractive energy subject to prescribed section clearance times. *Transportation Research Part B: Methodological*, 111:135 – 167, 2018. ISSN 0191-2615. doi: https://doi.org/10.1016/j.trb.2018.03.012.

L. Chen, C. Roberts, F. Schmid, and E. Stewart. Modeling and solving real-time train rescheduling problems in railway bottleneck sections. *Transactions on Intelligent Transportation Systems*, 16(4):1–9, 2015.

A. Galapitage, A. R. Albrecht, P. Pudney, X. Vu, and P. Zhou. Optimal real-time junction scheduling for trains with connected driver advice systems. *Journal of Rail Transport Planning and Management*, 8(1):29 – 41, 2018. ISSN 2210-9706. doi: 10.1016/j.jrtpm.2018.02.003.

X. Luan, Y. Wang, B. D. Schutter, L. Meng, G. Lodewijks, and F. Corman. Integration of real-time traffic management and train control for rail networks–Part 1: Optimization problems and solution approaches. *Transportation Research Part B*, 115:41–71, 2018a. doi: 10.1016/j.trb.2018.06.006.

X. Luan, Y. Wang, B. D. Schutter, L. Meng, G. Lodewijks, and F. Corman. Integration of real-time traffic management and train control for rail networks–Part 2: Extensions towards energy-efficient train operations. *Transportation Research Part B*, 115:72–94, 2018b. doi: 10.1016/j.trb.2018.06.011.

K. Panou, P. Tzieropoulos, and D. Emery. Railway driver advice systems: evaluation of methods, tools and systems. *Journal of Rail Transport Planning & Management*, 3:150–162, 2013.

G. M. Scheepmaker, R. M. Goverde, and L. G. Kroon. Review of energy-efficient train control and timetabling. *European Journal of Operational Research*, 257(2):355–376, 2017.

Energy Savings with Enhanced Static Timetable Information for Train Drivers

Thomas Graffagnino [a], Roland Schäfer [b], Matthias Tuchschmid [b], Marco Weibel [b]

[a] Timetable and Network Design, Infrastructure, Swiss Federal Railways
Hilfikerstrasse 3, 3000 Bern 65, Switzerland

[b] Energy, Infrastructure, Swiss Federal Railways
Industriestrasse 1, 3052 Zollikofen, Switzerland

Abstract

On the network of the Swiss Federal Railways (SBB) there is huge variability in the energy consumption for comparable train runs. Consequently, there is a significant potential to achieve energy savings in the context of improved driving strategy, which can be influenced by providing useful information to the train driver. As part of the smartrail programme operated by the Swiss railway industry, several energy savings measures are due to be implemented. As a first step in the smartrail energy measures, SBB conducted a pilot test in summer 2018. This pilot involved 473 test runs on two important passenger trains in Switzerland: the long-distance train IC5 and the local train S12. For each train run, based on effective routing, train composition, speed restrictions and timetable fixed points, a speed profile and new service times for each station were calculated early each morning for all the train runs of the day.

A survey among the test train drivers showed that more than 80% of them would welcome the rollout of the additional information in the near future. A comparison of the accompanied journeys against the 'baseline', i.e. same trains in the same period without additional information, shows a significant reduction in energy consumption without affecting punctuality: depending on the train journey, the accompanied runs consumed between 1.4% and 13.3% less energy per gross tonne-kilometre.

The high levels of acceptance by the train drivers combined with the significant energy savings achieved without affecting punctuality is very promising. For this reason, a system-wide rollout is currently being investigated and could be started by late 2019.

Keywords

Energy consumption, Timetable, Train control, Traffic-Management System, Train Driver

1 Introduction

Swiss Federal Railways (SBB) operates one of the most dense-running mixed traffic networks in the world. There is huge variability in the energy consumption of similar train runs. On train runs with a comparable duration on the same line, energy consumption can vary by approximately 50% (see Figure 1). Part of this variability can be linked to driving strategy. This illustrates that there is significant potential to generate energy savings through improved driving strategy, which involves providing useful information to the train driver.

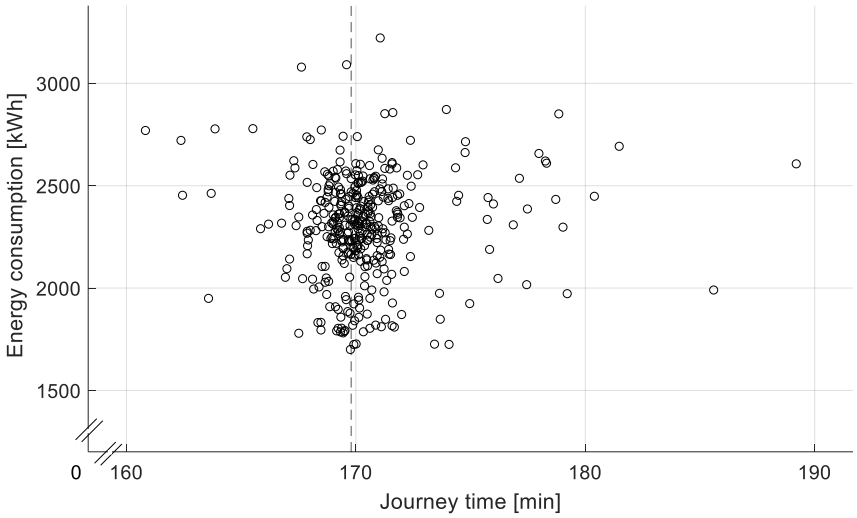


Figure 1: Energy consumption vs journey time on the line Zurich-Geneva Airport from December 2015 to August 2016. The dashed line indicates the nominal journey time.

Driving strategy is usually improved using driving advisory systems (DAS). During a preliminary pilot project by BLS (BLS (2019)) in spring 2017 (Studer et al (2017)), two different DAS with coasting capabilities were compared to an initial version of a static speed profile solution provided by SBB. For this pilot project, SBB provided the static speed profile solution and was responsible for comparing the energy consumptions of the three systems. Surprisingly, even though – from a theoretical point of view – the solutions with coasting capabilities should need less energy for a given running time, the static solution showed comparable energy savings in practice.

In 2017, SBB, together with other railway companies in Switzerland, started a digitalisation programme named smartrail (smartrail (2019)). As part of this programme, a range of different measures will be implemented with the aim of achieving energy savings and improving energy consumption. As a preliminary step for smartrail, SBB decided, based on the findings of the study with BLS, to enhance the traffic management system RCS with static speed profiles. This paper explains how the static speed profiles are calculated and presents the results of the first operative tests. Thanks to the relative simplicity of the static profile, a rollout on the SBB network could be realised within a short amount of time.

2 Methodology and Calculation of Speed Profiles

SBB conducted a pilot test in summer 2018, from the 20th of August to 22nd of September. Overall, 473 test runs were performed on two important passenger trains in Switzerland: the long-distance train IC5 on the line Zurich–Olten–Biel–Geneva and the local train S12 on the line Brugg–Zurich–Winterthur. For the tests, the regular train drivers were

accompanied by a representative from the project, who presented and explained the new timetable information.

2.1 Fixed Points and the Algorithm for the Speed Profile

Typical timetable planning in Switzerland begins with a calculation of the minimal running time between stations. Based on these minimal running times, linear time margins are added to the running times of passenger trains (typically 7%). Then, based on knowledge of actual traffic situations and expected delays, the time margins are changed to ensure higher levels of punctuality and traffic stability. After these steps, there is no related speed profile that considers the arrival, departure and passing time for all stations. The time margins are dimensioned in the form of percentages or absolute values without linking back to any speed profile for train driver.

In 2006, based on the real-life experiences of train drivers, SBB developed an algorithm that can reconstruct a feasible speed profile for a given train timetable. The key element in this calculation lies in identifying the stations where the times must absolutely be respected and the stations where a slightly adapted time has no significant negative effect. The stations where times must be respected are called fixed points. For the SBB pilot, we conducted interviews with planners and train dispatchers to identify the fixed points. Fixed points are typically stations with train conflicts, train connections and journey start or end points.

Knowing the target running time between the fixed points, an algorithm reduces the maximum speed in increments until the target running time is achieved. This algorithm considers only acceleration, braking and running at a constant speed, without factoring in coasting capabilities. It is important to mention that the braking phase of the static speed profiles is calculated with the use of regenerative braking, as SBB trains run on 15 kV AC. Working between each pair of fixed points, the algorithm can compute the new static speed profile for each train run, ensuring that the planned times are complied with the fixed points. At this stage, we also allow for slight time deviations from the annually planned times in day-to-day operations at stations which were not identified as fixed points. The small deviations from annually planned times is not a problem, because SBB doesn't communicate planned times to passengers. We communicate commercial times to passengers which are set so early that, the trains cannot depart earlier than them. The algorithm is configured so that the results are very easy to achieve for a train driver thanks to restricting speed changes to well-known positions on the track. Therefore, an additional train positioning system is not needed.

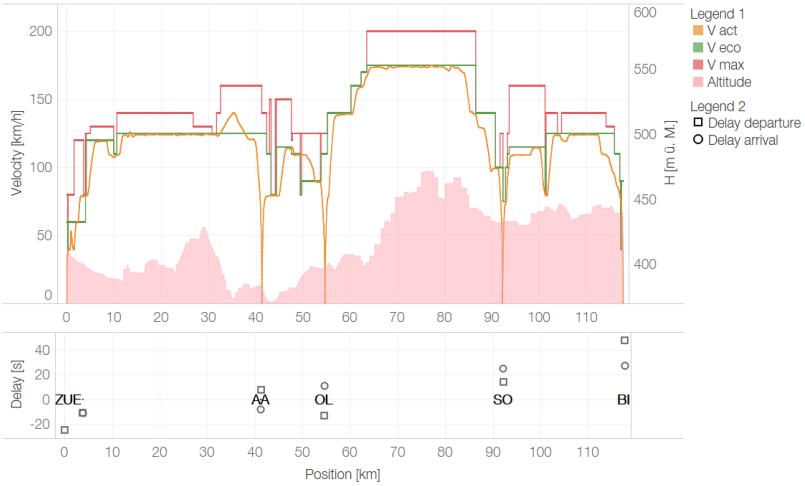


Figure 2: Example of energy-saving train run of IC5 between Zurich and Biel; the green line shows the daily computed static speed profile and the yellow line shows the actual speed profile for a train run.

As can be seen in Figure 2, the green line of eco speed did not factor the option of coasting into our daily computations of speed profiles. This is because we chose a static solution without train positioning and live delay calculation; without positioning, it is quite difficult to precisely determine when to coast for a punctual arrival. Furthermore, the train drivers are still allowed to coast with this static system and should consider the new eco speed profile information as the mean speed to achieve the target times.

2.2 Daily Calculation of Speed Profile

The daily calculations were conducted early each morning for about 100 test train runs using a special extended version of the system RCS. For each train run, a speed profile and the corresponding service times were calculated, based on effective routing, train composition, daily speed restrictions and timetable fixed points.

As shown in Figure 3, instead of annual timetable information without speed profiles, the daily computation provided the new information to be used for the tests. Within the RCS system, the exact routing, all speed restrictions due to maintenance on the network and all daily rolling stock information are provided for all trains. Enhanced with fixed point information, this daily computation delivers feasible and easily comprehensible timetable information for the train drivers:

- For each train run, there are timetable fixed points which must be respected to ensure that the operation remains conflict-free.
- Based on these identified fixed points, an algorithm creates a static speed profile which respects the fixed timing points and temporary as well as static speed restrictions.

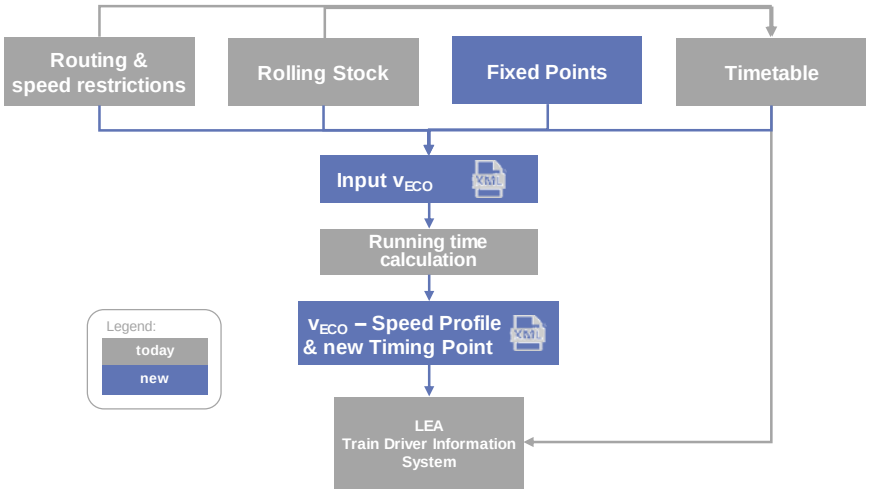


Figure 3: System overview for daily calculation of speed profiles.

2.3 Presentation to the Driver and Advice for the Test

The improved and extended timetable was provided to the train drivers using an adapted version of their usual train driver information system LEA (LEA (2013)). The new information for each test train run was:

- Planned timing point shown to an accuracy of 10 seconds instead of the usual 1 minute (Point 1 on Figure 4)
- For each station/stop daily calculated service times were shown instead of the annual commercial times (Point 1 on Figure 4)
- Important stations for conflict-free operation were marked down (Point 2 on Figure 4)
- The static speed profile for punctual operation was recalculated every day (Point 3 on Figure 4)

The train driver was asked to follow the suggested speed whenever the departure was on time. In the event of a delay, the train drivers were free to choose their own strategy to get the train back on time, as it is the case today on any train run. The conflict detection system of the traffic management system RCS (Rail Control System) (RCS (2019)) was active during the tests. This system is known as ADL (Adaptive Steering) (RCS ADL (2019)) and in the event of any conflict, the train driver was obliged to follow the advice given by the system as usual.

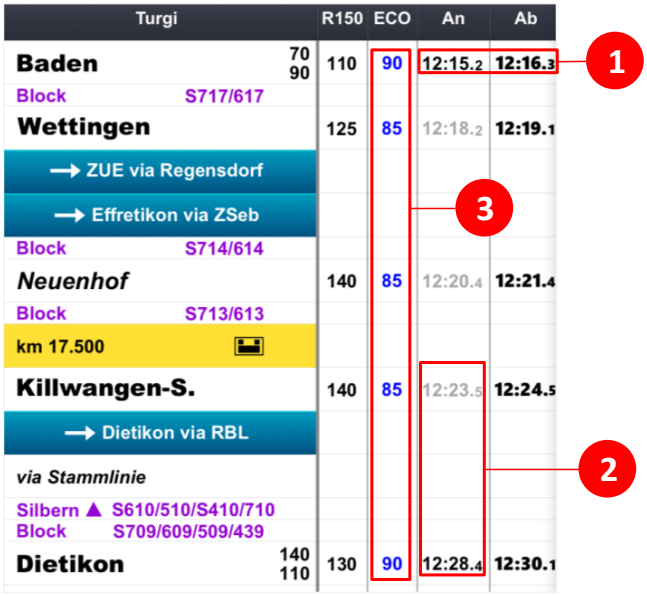


Figure 4: Example of enhanced train driver information.

3 Results

3.1 User Experience

The regular train drivers completed a questionnaire on the acceptance of the newly displayed information, focusing primarily on the optimised driving profile. In total, 242 train drivers responded to the questions, which represents 92% of the accompanied drivers from the test.

- 93% of train drivers state that they can implement the optimised driving profile well to very well.
- On average, 80% of the train drivers would welcome a rollout of the new timetable information in the near future.

There are some differences in the responses depending on the experience of the train driver and the type of rail traffic: Experienced train drivers tended to state that they already knew the static speed profile based on personal experience. Less experienced train drivers welcomed it more readily, viewing it as a shortcut to build up their own experience. In regional traffic, acceptance was generally higher than in long-distance traffic. We assume that the demands placed on regional drivers are greater and the workload of these drivers is higher, so any assistance is more appreciated.

3.2 Energy Consumption

Most of the trains were equipped with energy measurement devices with a temporal resolution of one second, which allowed us to perform a precise analysis on the train runs. The total amount of consumed energy for an individual train run was determined by

summing up the energy consumption from the start to the end time of the train run, as provided by the traffic management system RCS.

In the following section, energy consumption with or without application of the static profile is compared for the different tracks and directions. 'Eco' refers to cases where the static profile was provided, whereas 'baseline' refers to normal cases without any additional input provided to the train drivers. Statistical significance tests were performed for all the comparisons carried out.

To obtain comparable values, the energy consumption of every single train run was converted to a specific energy consumption value in Wh/Gtkm, with additional correction applied to cover the difference in altitude between the start and end stations, i.e. subtraction of the corresponding gravitational potential energy (referred to below as potential energy).

Local train between Brugg and Winterthur (S12)

A total of 276 runs were conducted during the test period. 159 of these runs were 'baseline' runs and 117 were 'eco' runs. Table 1 provides an overview of the test setup.

Table 1: Overview of test setup for S12

RABe 511 (Regio-Dosto)	**Distance**	**Average weight**	**Δ Potential energy**	**Journey time**
Brugg→Winterthur	56.6 km	306.8 t	72.5 kWh	55m 36s
Winterthur→Brugg	56.6 km	306.8 t	-72.5 kWh	54m 42s

Table 2 shows the specific, altitude-compensated energy consumptions of single train runs for the S12 in both directions. The reduction in energy consumption in the direction Brugg–Winterthur is more pronounced than in the other direction. We suppose that this is the case because the timetable for the direction Brugg–Winterthur allows for more scope for optimisation.

Statistical Significance of the Differences

The statistical significance of the differences between the 'baseline' and 'eco' runs was estimated using the null hypothesis that there is no difference between 'baseline' and 'eco'. Table 2 provides an overview of the results. Numbers in Wh/Gtkm denote median specific, altitude-corrected energy consumptions. Percentages denote relative differences between the 'baseline' and 'eco'. Bold-type percentages indicate significant results based on Wilcoxon rank-sum tests with significance level of 5% and p-values $p<0.01$. Statistically significant differences were obtained for both directions.

Table 2: Overview table of results for S12 energy consumption

Brugg–Winterthur	**-13.3%**
'baseline'	26.3 Wh/Gtkm
'eco'	22.8 Wh/Gtkm

Winterthur–Brugg	**-7.6%**
'baseline'	26.4 Wh/Gtkm
'eco'	24.4 Wh/Gtkm

Long-Distance Train between Zurich and Geneva (IC5)

For the evaluation of the long-distance IC5 trains, the analysis was sub-divided into segments.

- For both directions, the track was split in Biel, where there is often a change of train driver or train composition (single-unit to double-unit or vice versa)
- We differentiated between single-unit and double-unit trains due to the increased efficiency of double-unit trains observed on tracks with high maximal allowed speed (as compared to local trains with lower maximal allowed speed).

A total of 1406 runs were completed in the test. 1079 of these runs were 'baseline' runs and 327 were 'eco' runs. Table 3 provides an overview of the test setup for the four segments.

Table 3: Overview of test setup for IC5, direction Zurich–Geneva, with two segments.

RABDe 500 (ICN)	**Distance**	**Average weight**	**Δ Potential energy**	**Journey time**
Zurich→Biel	117 km	365.6 t	29.27 kWh	62min 18s
Biel→Geneva	152 km	365.6 t	- 44.9 kWh	81min 18s
Geneva→Biel	152 km	365.6 t	44.9 kWh	80min 42s
Biel→Zurich	113.2 km	365.6 t	-22.27 kWh	62min 48s

Statistical Significance of the Differences

The statistical significance of the differences between the 'baseline' and 'eco' runs was once more estimated using the null hypothesis that there is no difference between 'baseline' and 'eco'. Table 4 provides an overview of the results. Numbers in Wh/Gtkm denote median specific, altitude-corrected energy consumptions. Percentages denote relative differences between the 'baseline' and 'eco'. Bold-type percentages indicate significant results based on Wilcoxon rank-sum tests with significance level of 5% and p-values $p<0.025$. Significant differences were obtained for five out of eight sets.

Table 4: Overview and comparison of median specific, altitude-corrected energy consumptions. Percentages denote relative differences between 'baseline' and 'eco'.

	Single-unit train	**Double-unit train**
Zurich - Biel	**-3.0%**	*-1.4%*
'baseline'	134 runs: 23.0 Wh/Gtkm	133 runs: 21.9 Wh/Gtkm
'eco'	55 runs: 22.3 Wh/ Gtkm	29 runs: 21.6 Wh/ Gtkm

Biel - Genf	**-2.0%**	*-2.6%*
‘baseline’	175 runs: 19.7 Wh/Gtkm	92 runs: 19.0 Wh/Gtkm
‘eco’	60 runs: 19.3 Wh/Gtkm	22 runs: 18.5 Wh/Gtkm
Genf - Biel	*-2.0%*	**-4.2%**
‘baseline’	171 runs: 19.9 Wh/Gtkm	99 runs: 19.1 Wh/Gtkm
‘eco’	41 runs: 19.5 Wh/Gtkm	41 runs: 18.3 Wh/Gtkm
Biel - Zurich	**-3.4%**	***-7.4%***
‘baseline’	174 runs: 23.3 Wh/Gtkm	101 runs: 21.5 Wh/Gtkm
‘eco’	63 runs: 22.5 Wh/Gtkm	16 runs: 19.9 Wh/Gtkm

Note that the specific energy consumption is much higher for the segment Zurich-Biel (and vice versa) as compared to the specific energy consumption between Biel and Geneva (and vice versa). This is probably due to the high-speed segment (max. speed 200 km/h) between Solothurn and Olten.

3.3 Punctuality

While the tested system had no negative impact on punctuality, a more detailed look at the data produces a picture that is somewhat clearer.

In Switzerland, punctuality is measured based on a threshold of three minutes (in percent) on arrival at 53 major stations. As seen in Figure 5, the system compared to the baseline had no negative impact on this threshold of 180 seconds. Where it becomes more complicated is when we analyse the delay upon arrival between 0 and 60 seconds. The aim of the static speed profile is to use the running time margin in order to reduce energy consumption. In doing so, we expect to reduce the number of trains arriving at the stations early; this is clearly observable in the results. The discussion then turns towards what is acceptable within the timeframe of 0 to 60 seconds and if some trains should arrive slightly in advance by between -30 and 0 seconds. At the time of writing, discussions on this trade-off between punctuality and energy savings are still ongoing.

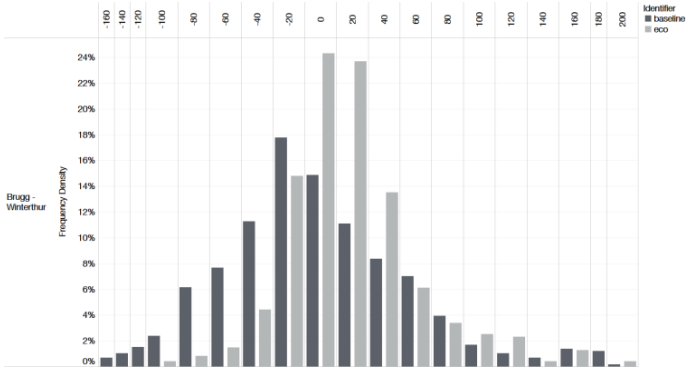


Figure 5: Histogram of delay upon arrival at fixed points for S12

Furthermore, for some test runs we measured an increase in arrivals with more than 60 seconds delay. These delays beyond 60 seconds are clearly not acceptable and we are

analysing the reasons. An initial analysis of the causes of these delayed arrivals identified the following factors: driver difficulties in knowing the exact delay of the train at any time, which impacts on the ability to review the run strategy with respect to delay; some quality problems in running time calculations and a lack of margin to counterbalance train driver reaction times.

In summary, the comparison of the accompanied runs with the 'baseline' (same trains in the same period) shows a significant reduction in energy consumption achieved without affecting punctuality at 3 minutes: depending on the train run, the accompanied runs consumed between 1.4% and 13.3% less energy per gross tonne-kilometre. In general, the reduction on local trains is higher than on long-distance trains.

4 Discussion and Next Steps

Most of the train drivers were astonished to discover how well suited the speeds of the static speed profiles are and stated that the figures were confirmed in practice. Furthermore, following the eco speed profile is practicable and the modifications as shown in figure 4 are understood within a few minutes. These high levels of acceptance by train drivers combined with the significant reduction in energy consumption without affecting punctuality based on the three-minute criterion is promising. For this reason, a system-wide rollout scheduled for late 2019 is currently ongoing. The central topics for implementation in late 2019 are the automatic generation of fixed points, the trade-off solution for energy consumption vs. punctuality and the training of all train drivers in how to use the new system.

The implementation of this system represents a first step for the future development of the RCS ADL system towards ATO. It is also a component of the larger project smartrail (smartrail (2019)), which aims to reduce global system costs while increasing safety and capacity. The next step for reducing energy consumption on train runs will be the introduction of coasting speed profiles with future ATO systems. Furthermore, the smartrail project is also developing a new timetable planning system, which needs to factor energy-saving considerations into calculations for running times between fixed points, based on the work of Prof. T. Koseki (Koseki (2015)), to ensure the lowest possible energy consumption.

The authors would like to thank all members of the team and all other SBB collaborators involved in this multidisciplinary project.

References

Studer T., Schäfer R., Graffagnino T., 2017, "Fahrempfehlungen im S-Bahn-Betrieb: Pünktlich und energiesparend am Ziel", Eisenbahntechnische Rundschau, Issue 010/2017

BLS, 2019, website https://www.bls.ch (check on 30.01.2019)

RCS, 2019, website https://www.sbbrcs.ch/en/ (check on 30.01.2019)

RCS ADL, 2019, website https://www.sbbrcs.ch/en/family/rcs-adl/ (check on 30.01.2019)

LEA, 2013, example on website https://stories.sbb.ch/mit-lea-im-fuhrerstand/2013/10/15/ (check on 30.01.2019)

Smartrail, 2019, website https://smartrail40.ch/ (check on 30.01.2019)

Koseki T., Watanabe S., 2015, "Energy-saving train scheduling diagram for automatically operated electric railway.", Journal of Rail Transport Planning & Management., 2015 Vol.5, No.3, p.183.

Multi Objective Optimization of Multimodal Two-Way Roundtrip Journeys*

Felix Gündling [1], Pablo Hoch [1], Karsten Weihe [1]
[1] Department of Computer Science, Technische Universität Darmstadt
Darmstadt 64289, Hochschulstraße 10, Germany
E-mail: {guendling, hoch, weihe}@cs.tu-darmstadt.de, Phone: +49 (0) 6151 16-20874

Abstract
Multi modal journeys often involve two trips: one outgoing and one return trip, as in many cases, the traveller would like to return to her starting point. If a car or bike was used in combination with public transportation (i.e. park & ride), this introduces a dependency between outward and return trip: both must include the same parking place. Optimizing both trips independently may yield suboptimal results. We consider the multi modal two-way roundtrip problem and propose several algorithms. All proposed algorithms compute journeys that are optimal regarding multiple criteria. We present a variant that supports price optimization (including driving and parking costs) as a Pareto criterion in addition to travel time and the number of transfers. Our study with realistic scenarios based on real data shows promising results.

Keywords
2-way round trip, multi modal, multi criteria optimization, park and ride, routing

1 Introduction

Many journeys do not consist of one-way trips. On the contrary, in many cases travelers return to the starting point (home for private, office for business trips) (Baumann et al., 2004). We consider a very common practical use case of multi modal routing: optimizing outward and return trip of a journey involving both private (e.g. a private car or bike) and public transportation (e.g. busses, trains, etc.). Planning a journey with commonly available online systems, that calculate optimal one-way trips, becomes quite cumbersome: finding the optimal P&R parking place (or an optimal place to park the bike) is not trivial. A parking place that was optimal for the outward trip might yield a suboptimal or infeasible return trip and vice versa. Considering multiple optimization criteria such as the number of transfers and travel time (accumulated for both trips), there might even be multiple optimal solutions. The reason for this is the time-dependent and directed nature of public transportation: an optimal route on the return trip does not necessarily include the parking place used in the outward trip. Combining independently optimized journeys may thus yield suboptimal or infeasible journeys. Consequently, optimizing both trips in a combined manner is required to compute optimal journeys for this use case. The fastest journey may not always be the most attractive one for everyone: in addition to a short travel time, some users prefer a cheap and/or convenient travel that minimizes the number of transfers. Since priorities of those

*This work was partially supported by Deutsche Bahn.

optimization targets differ from traveler to traveler, our approaches compute a complete Pareto set considering convenience (number of transfers) and travel time as criteria. One variant additionally considers the price of the journey as optimization criterion: the total price is the sum of the costs of private transportation including time dependent costs for parking as well as the public transport ticket price. For the public transport ticket price, our price model assigns a separate milage price per means of transportation (high speed trains are more expensive than local public transport). Parking prices are based on the parking duration. Our evaluation is based on real data: the public transport timetable is provided by Deutsche Bahn and covers Germany including trains (long distance as well as local), metro, busses and streetcar services. Street routing is based on Project OSRM (Luxen and Vetter, 2011) using an OpenStreetMap dataset covering the same geographic area as the public transport timetable. To the best of our knowledge there are neither scientific publications nor commercial systems offering this functionality. The presented algorithms are suitable for use in online routers and mobile routing applications.

The paper is organized as follows. Section 2 discusses previous work in the area. Section 3 outlines our contribution to the topic. Section 4 introduces basic notation and the formal problem definition. Section 5 describes the different approaches. Section 6 shows how to extend these approaches to optimize price as an additional Pareto criterion. Section 7 presents the results of our experiments and Section 8 concludes our results and gives an outlook to future research directions.

2 Related Work

To the best of our knowledge, there are no publications that solve the described problem in a Pareto-optimal way optimizing multiple criteria. The first solution solving the problem (Bousquet et al., 2009) considers a single optimization criterion: travel time. An improved bi-directional shortest path algorithm to solve the problem is described in (Huguet et al., 2013). Another approach based on access node routing (Delling et al., 2009) is presented in (Spinatelli, 2015). All three publications optimize travel time as single criterion and apply their algorithm to datasets covering a single city: Paris including its suburbs (Huguet et al., 2013), the rural area around Lyon (Bousquet et al., 2009), and Milano (Spinatelli, 2015). Recent advances in public transport routing and multi modal routing such as RAPTOR/MCR[1] (Delling et al., 2012, 2013), CSA (Dibbelt et al., 2013), TripBased (Witt, 2015) were not extended to compute Pareto-optimal journeys for the multi modal park and ride two-way roundtrip problem.

3 Contribution

In this paper, we present various algorithms to solve the two-way park and ride roundtrip problem optimizing multiple criteria in a Pareto-optimal way. We compare different solutions based on a time-dependent graph model (Disser et al., 2008) with an algorithm based on connection scanning (Dibbelt et al., 2013) and another algorithm which is based on Trip-Based routing (Witt, 2015). All approaches optimize travel time as well as the number of transfers. Furthermore, we propose a variant that additionally optimizes prices.

We evaluate all algorithms on a realistic nationwide network: a complete public trans-

[1] Round bAsed Public Transit Optimized Router, Multimodal Multi Criteria RAPTOR

port schedule for all of Germany including all modes of public transportation (e.g. busses, street cars, all kinds of trains) kindly provided by Deutsche Bahn. Our computational study shows that our algorithms are suitable to be deployed in online or mobile multimodal routing systems.

4 Preliminaries

This section describes the problem definition, static and dynamic/user inputs, and how they are preprocessed to be used as input for our core routing algorithms.

4.1 Problem Definition

We consider computing Pareto optimal solutions to the problem $\alpha \xrightarrow{t_{\text{out}}} \omega@[t_1, t_2] \xrightarrow{t_{\text{ret}}} \alpha$ where we call the outward trip t_{out} and the return trip t_{ret}. α and ω are locations (addresses / geographic coordinates). α might be the user's home address and ω the office address. The time interval $[t_1, t_2]$ is the minimal time range to stay at ω (e.g. office hours). Thus, our journeys have one of the following two structures:

$$\alpha \xrightarrow{\text{car}_1} p \xrightarrow{\text{walk}_1} s_w \xrightarrow{\text{pt}_1} s_x \xrightarrow{\text{walk}_2} \omega@t_1 \quad \dots \quad \omega@t_2 \xrightarrow{\text{walk}_3} s_y \xrightarrow{\text{pt}_2} s_z \xrightarrow{\text{walk}_4} p \xrightarrow{\text{car}_2} \alpha \quad (1)$$

$$\alpha \xrightarrow{\text{walk}_1} s_w \xrightarrow{\text{pt}_1} s_x \xrightarrow{\text{walk}_2} \omega@t_1 \quad \dots \quad \omega@t_2 \xrightarrow{\text{walk}_3} s_y \xrightarrow{\text{pt}_2} s_z \xrightarrow{\text{walk}_4} \alpha \quad (2)$$

The first one is most interesting to us. However, enabling the approach to find journeys with the second structure is necessary to avoid presenting unreasonable journeys to the user: it is not reasonable to use the car[2] if the trip between α and s_w over p ($\alpha \longleftrightarrow p \longleftrightarrow s_w$) takes longer than walking directly between α and s_w ($\alpha \longleftrightarrow s_w$). By allowing both structures, the journey involving the unnecessary car leg (structure 1) will be superseded by the walking journey (structure 2).

We minimize the combined travel time sum of t_{out} and t_{ret} as one Pareto criterion and the combined number of transfers of t_{out} and t_{ret} as another. The travel time includes the time from the start with the car at α until t_1 for the outward trip and the time from t_2 until α is reached again for the return trip. This includes waiting times at ω.

Note that the stations s_y and s_z as well as the stations s_w and s_w do not need to match but the parking place p is required to be the same for outward trip t_{out} and return trip t_{ret}. The user specifies α, ω, t_1, t_2, maximum driving distance d_{max} and maximum walking distance w_{max}. This naturally limits the number of parking places (candidates for p) and stations (for journey structure 2) reachable from α (car_1 and car_2 / walk_1 and walk_4), the number of candidate stations for s_w and s_z reachable from a parking place (walk_1 and walk_4), and the number of candidate stations for s_x and s_y reachable from ω (walk_2 and walk_3).

4.2 Inputs

Basically, an algorithm to solve the problem described above requires information about the road network, the locations of suitable parking places P, and the public transport timetable.

[2]Bad weather or mobility impairments could be reasons to use the car regardless of longer travel time. However, weather dependent routing and routing for handicapped persons is not addressed in this paper.

The road network as well as the locations of parking places are extracted from OpenStreetMap. The public transport timetable consists of a set of stations S where each is associated with a geographic coordinate and a transfer time, *trips* (a vehicle visiting a stop sequence with associated departure and arrival times) and a set of *footpaths* that connect stations which are in close proximity so that walking between them is feasible. Furthermore, the timetable data contains information such as track names, service head signs, train category and service attributes like wireless internet availability or bicycle carriage. All presented algorithms require the trips to be grouped into *routes*: all trips in a route share the same sequence of stations. Additionally, trips in a route are not allowed to overtake each other. Otherwise, the route needs to be split into two separate routes. Grouping into routes is done as a preparation step.

4.3 Preprocessing

Since driving and walking is only available for the first and the last leg of both trips t_{out} and t_{ret}, we do not need to integrate both networks (timetable and road network). This allows us to use specialized models and algorithms for each network: contraction hierarchies for the road network and Time Dependent/CSA/TripBased routing for public transport (cf. Section 5). Consequently, we can split the procedure to compute optimal roundtrip journeys into two parts without losing optimality: the preprocessing step computes all possibilities for the first and last leg of t_{out} and t_{ret}. This is the input for the actual core routing algorithm described in Section 5.

Procedure `preprocess_roundtrip()` shown in Listing 1 computes three sets W, C, and D: these enumerate all possibilities to reach a public transport station from α (sets W and C) and ω (set D) respecting the journey structure and user supplied driving and walking limits d_{max} and w_{max}. The preprocessing makes use of the following data structures and procedures:

- The procedures `car_route` and `foot_route` compute shortest paths (optimizing travel time) on the car/foot street network. They return the required time. Our `car_route` routine makes use of (Luxen and Vetter, 2011). The `foot_route` routine is a specialized implementation based on OpenStreetMap data. Routes by foot are computed by a specialized algorithm that considers stairs, crossing roads, elevators, and many more elements.

- The table `dist` contains precomputed foot path durations between parking places and nearby stations. `dist[`p`][`s`]` is the time it takes to walk from parking place p to station s (and vice versa).

- `get_stations` and `get_parkings` are geographic lookup functions taking a coordinate and a radius. They return all stations/parkings where the distance to the given coordinate is less than the provided radius. The functions can be efficiently implemented using a spacial data structure such as an R-tree or a quadtree.

C contains all possibilities to get to a public transport station from α and vice versa (required to find journeys with Structure 1). Note that the entries store also the parking location. This is important because the core routing algorithm needs to match parking locations from t_{out} and t_{ret}. W contains all possibilities to walk between α and nearby public transport stations within w_{max} distance. W is required to find journeys with Structure 2. D

Listing 1: Preprocessing Procedure: computes edge sets W, C, and D to connect α and ω with the public transport network.

```
dist[p ∈ P][s ∈ S]

fn car_route(from, to) do ... return driving_time done
fn foot_route(from, to) do ... return walking_time done
fn get_parkings(coordinate, radius) do ... return parking_set done
fn get_stations(coordinate, radius) do ... return station_set done

fn preprocess_roundtrip(α, ω, d_max, w_max) do
  W := ∅ // possibilities for α ⟷ s ∈ S via foot
  walking_candidates := get_stations(α, w_max)
  foreach s ∈ walking_candidates do
    walking_time = foot_route(α, s)
    W := W ⋃ {(α → s, walking_time), (s → α, walking_time)}
  done

  C:= ∅ // possibilities for α → p ∈ P → s ∈ S and s ∈ S → p ∈ P → α
  Π := get_parkings(α, d_max) // parking candidates
  foreach p ∈ Π do
    c_out := car_route(α, p) // driving time outward
    c_ret := car_route(p, α) // driving time back
    station_candidates := get_stations(p, w_max)
    foreach s ∈ station_candidates do
      w := dist[p][s] // walking time between parking and station
      C := C ⋃ {(α → p → s, c_out + w), (s → p → α, c_ret + w}
    done
  done

  D := ∅ // set of possibilities ω ⟷ s ∈ S via foot
  destination_station_candidates = get_stations(ω, w_max)
  foreach s in destination_station_candidates do
    walking_time := foot_route(s, ω)
    D := D ⋃ {(s → ω, walking_time), (ω → s, walking_time)}
  done

  return (W, C, D)
done
```

contains all possibilities to walk between ω and nearby public transport stations within $w_{\max}$ distance.

The code in Listing 1 can be improved by computing the routes to all targets in one step. Since Dijkstra-like algorithms (like contraction hierarchies employed in `car_routing`) are inherently "multi target"-algorithms, we calculate the walking/driving times to all candidates in one single step instead of running one one-to-one query for each target in a loop. Thus, we change the interface of `route_car` and `route_foot` to take a single location and a set of targets as input and return the travel time to each target as result.

5 Approaches

This section describes the different approaches for the core routing procedure. Each algorithm takes the same input computed in the preprocessing phase (in addition to the public transport timetable): the sets W and C which connect α with the public transport network through walking/driving, and D which connects ω with the public transport network through walking.

5.1 Time Dependent Graph

One established way to compute multi criteria shortest paths on public transport timetable networks are label correcting algorithms on graph data structures representing the timetable (i.e. time expanded and time dependent graphs). The time dependent graph is more compact (as compared to the time expanded graph) and therefore better suited to cope with large timetables containing not only trains but also streetcars and busses. So our first approach is based on the time dependent graph as described by Disser et al. (2008).

The model presented in (Disser et al., 2008) does not support backward search (latest departure problem) because it is not consistent, meaning that the path lengths $\ell(u, v)$ and $\ell(v, u)$ in forward and reversed graph differ for at least one optimization criterion. An example and the new graph layout can be found in Appendix A. So from now on, finding the journey with the latest departure (starting with a fixed arrival time) is analogous to finding the journey with the earliest arrival (starting with a fixed departure time) with reversed edges. This does also apply to the multi criteria case.

Baseline
In this section, we will describe an algorithm that is purely based on an unchanged base algorithm: the time dependent earliest arrival problem. We extend the graph model so that it fits the problem. Basically, the sets W, C, and D can be seen as edges which extend the time dependent graph. Consequently, we need to add nodes to represent α and ω. In the following, α and ω refer to those additional nodes if they are used in the graph context. Routing t_{out} and t_{ret} independently with all additional edges at once could yield suboptimal or unfeasible journeys due to non-matching parking places.

Since the edges from the set D (connecting ω with public transport stations and vice versa) do not introduce any dependencies, they are added for every search. It is sufficient to add those that match the search direction ($\omega \to s \in S$ for t_{ret} and $s \in S \to \omega$ for t_{out}). However, to prevent the interference between parking places, we conduct one multi-criteria search for each parking place separately for t_{out} and t_{ret}: the graph gets extended by all edges (one for each station that is reachable from p) that lead over the selected parking. These are earliest arrival problems $\omega@t_2 \to p_i$ in case of t_{ret} and latest departure problems $p_i \leftarrow \omega@t_1$ for t_{out} (for every parking i). This generates all optimal trips $T_{\text{out}}^{p_i}$ for t_{out} and $T_{\text{ret}}^{p_i}$ for t_{ret} for every potential parking place p_i. Waiting time (arriving earlier than t_1 or departing later than t_2 at ω) is considered travel time and is therefore minimized as described in Section 4.1. Note that every overall optimal roundtrip needs to be a combination of optimal trips t_{out} and t_{ret} for one of those potential parking places. Otherwise (if an optimal roundtrip would not be a combination of optimal individual trips), it could obviously be improved by an optimal one. Consequently, the combination of all computed trips $\bigcup_{p \in P} T_{\text{out}}^{p} \times T_{\text{ret}}^{p}$ contains all optimal roundtrips. Removing all roundtrips that are superseded by others (including

duplicate ones) yields the final set of optimal roundtrips.

Edges from the set W representing all options to walk from α to a public transport station (for t_{out}) and vice versa (for t_{ret}) are added in a separate search (to enable the system to find journeys of Structure 2). Since there are no constraints to use the same parking place in t_{out} and t_{ret}, they can all be used in one search.

This approach requires $2(|\Pi| + 1)$ invocations of the basic time dependent routing routine (earliest arrival / latest departure) where Π is the set of all considered parking places: for each direction one invocation for every parking place candidate and one with all edges from W. This is certainly not optimal regarding computational effort (compared to the approaches presented later on). However, this approach is still useful for the practical verification of other approaches.

Parallelization

Since all searches are independent, they can be trivially parallelized. In theory, if the number of parallel processors is equals to two times the number of parkings, this can reduce the overall calculation time to the time it takes to respond to one routing query. However, since most systems (besides super compute clusters) cannot provide this level of parallelism, this is not a feasible approach, either.

Combined Search for t_{out} and t_{ret}

The basic Dijkstra algorithm computes shortest paths not only to the target node but to all nodes in the graph. Since the basic algorithm presented by Disser et al. (2008) makes use of goal direction and domination by terminal labels[3], this property does not apply anymore: when the algorithm terminates, only the labels for one destination node will be correct (in the sense that they necessarily represent the non-extensible Pareto set).

The first step of the combined search approach is to compute the shortest paths from/to every single parking place like in the baseline approach described in Section 5.1 for one direction t_{out} or t_{ret}. For the opposite direction, we now can make one combined search: instead of adding just the edges for only one parking, we add all parking edges but combine the edge cost with the criteria computed for the opposite direction in the first step. Since the first routing can yield more than one optimal trip for one parking, we have one additional edge for each optimal trip. Assuming we chose t_{out} in the first step, we add one edge from $s \in S \rightarrow \alpha$ for each optimal t_{out} journey using parking $p_i \in P$ for each station reachable from p_i. The edges carry the following costs: $(\text{dist}[p][s] + \text{car\_route}(p, \alpha) + \text{tt}_i, \text{ic}_i)$ where ic_i is the number of transfers and tt_i is the travel time for journey i in t_{out}. If t_{ret} was chosen for the first step, the approach works analogously.

This approach allows us to reduce the complexity of the baseline approach from $2(|\Pi| + 1)$ invocations of the time dependent routing routine to $|\Pi| + 1$ invocations: in one direction (outward or return) we need to route to/from every parking. In the return direction, only one query is required. The invocation with all edges from W in the first step stays the same. The search in the opposite direction is conducted with all edges in W.

As with the baseline approach, this approach can also be parallelized. However, this approach has one constraint on the ordering: the invocation for the opposite direction requires all results from the first step.

[3] labels are partial journeys that are used in the routing algorithm

No Terminal Domination and Worst Bounds

Applying domination by terminal labels (in combination with lower bounds and goal direction) in the time dependent graph routing is a very effective speedup technique for queries to a single target. However, in this setting (preventing interference of labels that use different parking places), domination by terminal labels as implemented in our basic time dependent routing algorithm demands a high number of invocations as we have seen in the previous three sections. Now, we want to further reduce the number of invocations by computing all optimal journeys over all parkings in one run of the algorithm (as opposed to $|\Pi|$ invocations in the first step of Section 5.1). Simply disabling the domination by terminal labels and routing with all additional edges (W, C, and D) would be one option.

However, domination by terminal labels can be replaced by a different technique that still allows us to discard labels early in the search process: those that are worse or equal to the combined worst (i.e. numerically greatest assuming optimization criteria are minimized) value of each optimization criterion over all parking places ("worst bound") cannot contribute a new optimum. Consequently, this requires at least one terminal label for each parking place. Until this precondition is met, we cannot discard any label.

To implement this, we have a list of parkings that were not yet reached. This list is initialized with all parkings (identified by a unique index) reachable from α. Every time a label reaches the target node, the used parking is removed from this list if it is the first to use this parking. If the list is empty (i.e. every parking was reached), this means that domination by worst bounds can be applied. To track the worst bounds, one variable per optimization criterion is introduced and updated every time a label is created on the target node. If every parking was reached, every newly (through edge extension) created label is compared to the stored combined worst bounds and discarded if its criteria values are equal or greater. The same check will be applied upon queue extraction since the worst bounds can change between queue insertion and extraction. Labels that were created through expansion of edges carrying different parking indices are deemed incomparable to prevent domination of options that may be part of an optimal round trip but are not optimal for this search direction. Walking options from W are always added. They can be implemented as "virtual" parking directly at α. Thus, no driving is required.

The routing for the opposite direction can be implemented as described in Section 5.1 and therefore benefit from unconditional dominance by terminal labels. This approach cannot be parallelized. Altogether this approach further reduces the number of invocations to two, albeit more complex calls: one for each direction t_{out} and t_{ret}.

Concurrent

In this section, we present an algorithm that handles the search in both directions (for t_{out} and t_{ret}) in an interleaved manner. Basically, we still have two multi criteria Dijkstra algorithms with the addition that they exchange information at runtime. Thus, every data structure (such as the priority queue, lower bounds, etc.) is redundant: one for each search direction.

Instead of the standard domination by terminal labels, we maintain a list of complete roundtrips: every time a new label has reached the target node in one direction, it is combined with each terminal label of the opposite direction that has a matching parking place. The resulting valid round trips are then added to the list of complete round trips if they are Pareto optimal. Previously added roundtrips that are worse than the newly added roundtrip are removed. These complete round trips can then be used to dominate labels in both search directions at the creation of new labels and after queue extraction: if a partial roundtrip is

already worse (in the Pareto sense) than a complete roundtrip, it can be discarded. Instead of comparing the label values (here: travel time and the number of transfers) directly with those of the terminal label, we can employ lower bounds to discard suboptimal labels as early as possible: a label with travel time t in the t_{out} routing takes at least $t + lb_{t_{\text{out}}}[n] + lb_{t_{\text{ret}}}[\omega]$ minutes for the complete roundtrip where $lb_{t_{\text{out}}}$ and $lb_{t_{\text{ret}}}$ are precomputed lower bounds for every node in both search directions. This is analogous for the t_{ret} search: $t + lb_{t_{\text{ret}}}[n] + lb_{t_{\text{out}}}[\omega]$.

All in all, we reduced the number of invocations from $2(|\Pi| + 1)$ for the baseline approach to just one. This comes with an increased complexity of the queries. However, the combination of information (instead of separate invocations of the basic time dependent routing procedure) as described here reduces the total number of steps required to compute all optimal round trips.

5.2 Connection Scan

In this section, we present an algorithm that is based on the Connection Scanning Algorithm (CSA) by Dibbelt et al. (2013). As opposed to the time dependent routing algorithm, it does not require a graph to represent the timetable, neither does it depend on a priority queue. The timetable model is a simple array of all elemental connections (departure and arrival of a trip with no intermediate stops in between) of the timetable sorted by departure time. The algorithm iterates through the array and updates earliest arrival times at the stations visited by the iterated connections accordingly. The algorithm also handles footpaths between stations and transfer times between transport services.

As the basic variant of CSA just iterates "through time" (sorted connections) it is not directed towards a specific target station. Therefore, it is well suited to be adapted as a multi target algorithm without a performance penalty. This can be utilized: in the first step, we ignore the actual driving and walking times from D and W that connect α with the public transport timetable. Instead, we search from all stations in D to all stations in W and C.

The original publication does not describe a multi-criteria version of the earliest arrival problem or journey reconstruction for this type of search nor does it describe the latest departure problem or multi source and multi destination routing. Consequently, we need a specialized version of the CSA algorithm for our use case:

- *Multiple Start Stations*: In the basic version, only one station is initialized with the desired start time. In our use case, every station in D is initialized with the walking time (between ω and $s \in S$) as offset that is added to t_2 (for t_{ret}) and subtracted from t_1 (for t_{out}).
- *Multiple Destination Stations*: Basically, this is what the algorithm does anyway if we omit the early termination mechanism which stops when the departure time of the currently iterated connection exceeds the earliest arrival at the destination.
- *Latest Departure Problem*: For t_{out}, we need to solve the problem $(s \in W \cup C) \leftarrow \omega @ t_1$. This can be done analogously to the forward search. For example, the connection array is sorted by descending arrival time and footpath walk times are subtracted instead of added.
- *Multi Criteria*: To support the optimization of the number of transfers as additional Pareto criterion, we do not only store a single earliest arrival time for each station but instead one for each number of transfers. The same applies to the array T which

indicates whether a trip can be reached or not: instead of single reachable bit, one bit per number of transfers is stored. The n^{th} bit indicates whether the trip can be reached with n transfers.

- *Reconstruction*: Since additional journey pointers (which would need to be maintained for every number of transfers) as described in (Dibbelt et al., 2013) slow down the search (scan running time), we chose to adapt the version that works without them. As our implementation of the algorithm supports the optimization of the number of transfers as Pareto criterion, we need to reconstruct one journey for each optimal number of transfers. The recursive call with n transfers at the next interchange stop continues with $n - 1$ transfers. Similarly, the trip reachable array needs to be looked up at the bit referring to the current number of transfers. Not knowing where the journey may have started imposes additional complexity: we need to iterate every possible station and check whether the travel time matches the walk ($\omega \leftrightarrow s \in D$) for this station.

Now that we have a variant that handles multiple departure stations, multiple destinations and multiple criteria in both search directions (earliest arrival / latest departure), we can utilize it to find optimal round trips: For each direction t_{out} and t_{ret}, we execute one search. Both searches are independent and can therefore be executed in parallel. We execute one latest departure query (starting at $t_1 @ \omega$) for t_{out} and one earliest arrival query (starting at $t_2 @ \omega$) for t_{ret}. The results of those queries are then merged to complete roundtrips by iterating every parking place and combining all journeys from t_{out} and t_{ret}. Since not every roundtrip is necessarily optimal, we remove all that are not Pareto optimal. This yields the full set of optimal roundtrip journeys.

5.3 TripBased

As with the Connection Scanning Algorithm, TripBased routing as presented by Witt (2015) is inherently a multi target routing algorithm: it can be seen as a breadth first search on a graph-like data structure consisting of trip sections and transfers between those trip sections. Similarly to the RAPTOR algorithm (Delling et al., 2012), it operates in iterations/"rounds" where the n^{th} iteration computes all optimal connections with n transfers. Each round updates the trip sections that are reachable through one additional transfer from previously reachable trip sections.

We adapt the algorithm to be able to compute optimal journeys to multiple targets. Therefore, we need to keep one result set J for each target station. Additionally, the earliest arrival time τ_{min} needs to be kept separately for each target station to check whether a trip reaching the target is optimal. A new trip segment needs to be added to the queue only if its arrival time does not exceed the maximum earliest arrival time τ_{min} over all target stations. Otherwise, it can be discarded because it cannot be optimal for any target station anymore: every slower connection with less transfers was already discovered in a previous iteration.

The additional footpaths between ω and nearby public transport stations can be handled analogously to those already contained in the basic static timetable.

In addition to the changes required to compute optimal journeys to multiple targets, the basic TripBased algorithm needs to be adjusted to compute connections for the latest departure problem, not just the earliest arrival problem. Since the preprocessed transfers (transfer reduction step) differ for the forward (earliest arrival) and reverse (latest departure)

direction, we need to have one transfer set T for each search direction. This doubles the preprocessing workload. Otherwise, the latest departure computation is analogous to the earliest arrival computation described in (Witt, 2015).

As we now have an algorithm with properties similar to the adapted CSA algorithm (multi criteria, multi source, multi target, earliest departure, earliest arrival), we can use it to compute optimal roundtrip journeys as described in Section 5.2.

6 Price as an Additional Optimization Criterion

In this section, we present a version that optimizes not just travel time and the number of transfers but also the price. The price of the complete roundtrip is comprised of the costs for parking at p (depending on the parking duration), the driving costs (of car_1 and car_2), the public transport ticket price (of pt_1 and pt_2) and an hourly wage to eliminate cheap but exceedingly long journeys that are unattractive from a practical perspective.

Since public transport pricing models are very complex, constantly changing and different for every area, we decided to use two artificial pricing models. Both are milage and vehicle class based: a high speed train (such as a German ICE or French TGV) costs \$0.22 per kilometer, a local train costs \$0.18 per kilometer and short distance transports such as busses and trams cost \$0.15 per kilometer. Additionally, we introduce an hourly wage of \$4.80 (converted to the atomic timetable time unit, minutes). The first model computes just the sum of those costs. The second, more advanced model, introduces a special ticket that allows the passenger to use arbitrary local transports (local trains, busses, trams) for a flat price (\$42.00 here). All mentioned values are freely configurable.

All algorithms need an updated route definition which takes the vehicle class into account because otherwise later departures (which will not be considered by the algorithms) may yield a cheaper connection. In the following, we describe the extensions to the approaches presented in Section 5 that enable price optimization for the two price models described above.

6.1 Graph Based

Extending the graph based approaches (Baseline, Parallelized Baseline, Worst Bound and Concurrent) to support price as additional Pareto criterion is mostly straightforward: the edge weight vector as well as the individual labels carry the price as additional entry. However, we need to also adjust the label comparison. Before, a label a dominated label b if and only if every criterion value of a was less than or equal to the corresponding criterion value in b. The criteria were $(\text{arr}_i, -\text{dep}_i, \text{transfers}_i)$ where arr_i and dep_i are the arrival and the departure time of label i.

Instead of just adding the price to this comparison, the hourly wage requires special treatment to retain correctness of the search. Assume we compare two labels a and b where a has a higher ticket price than b but a lower total price because it arrived earlier and did accumulate less costs due to the hourly wage. Consequently, from a Pareto perspective a dominates b (lower price, earlier arrival with same departure time). Since b arrived later, it now has to wait less for the next departure. Due to the hourly wage, the edge costs less for b than it does for a. So after edge expansion, a does not dominate b anymore which implies that we lost an optimal connection. To prevent this, we need to add the hourly wage price of the travel time difference to the price of a when comparing a with b. This way, the waiting

time disparity is compensated.

The additional edges derived from the set C (connecting α with public transport stations through a parking) now carry the according kilometer based price for the car route. Additionally, the parking itself can be modeled as a time dependent edge: coming back later to the parking increases the costs by \$2.00 per hour (staircase function).

6.2 Connection Scanning

Data Structures Extending the Connection Scanning algorithm to support price as an additional optimization criterion in the Pareto sense requires more effort than for the baseline approach because the data structures were designed with only travel time and number of transfers in mind. Before, the data structures holding the earliest arrival time for each station (S - note that we use the nomenclature of the CSA publication in this section) and the trip reachable bits for each trip T were both two-dimensional arrays with one entry for each number of transfers. This was sufficient for two criteria (number of transfers and travel time) because for each number of transfers only the fastest journey was relevant. Now, when additionally optimizing prices, there can be an arbitrary number of optimal journeys for each number of transfers (all optimal trade-offs between travel time and price). Thus, each entry of S[station][transfers] now maintains an array with all Pareto optimal travel time / price tuples for this station instead of just the minimal travel time for this number of transfers.

The array T holds a bit (for each number of transfers n) that indicates whether a trip is reachable with n transfers. However, this is not sufficient because it is not known at which cost the trip can be reached. Note that the price to reach the trip is not the same for each section of the trip. Consequently, we need to maintain the cheapest price for each trip section for each trip for each number of transfers. This is necessary to compute the correct journey price when iterating the connections array in the main loop of the Connection Scanning Algorithm.

Algorithm When initializing S with the offsets from D (foot routes between ω and nearby public transport stations) the price (incurred by the hourly wage) needs to be initialized, too. Furthermore, the main loop of the algorithm needs to be adjusted: if a connection is reachable through the station and the trip reachable flag is set (i.e. it has a price entry for the corresponding trip section), the cheaper solution is selected. If entering the trip at this station is the cheaper solution, the price of the following trip segments in the T array needs to be updated with the cheaper price including the hourly wage. Only Pareto optimal entries (travel time / price tuples) are added to S[station][transfers] removing superseded ones. Footpaths also incur costs due to the hourly wage.

Reconstruction Naturally, the journey reconstruction step also needs to be adapted to the new data structures: when looking up S and T entries, not only the travel time and the number of transfers but also the price of the entry needs to match.

6.3 Trip-Based

Preprocessing The preprocessing to eliminate unnecessary transfers was aimed at transfers and travel time as optimization criteria. Thus, transfers that lead to cheap connections may be discarded. To prevent this, the transfer reduction step is omitted. Even transfers to

later trip sections of the same trip (in case the trip visits a station two or more times) and other trips of the same line can save money. U-turn transfers are still being removed.

Data Structures To track the price of each journey, queue entries now also carry the current journey price (in addition to the trip segment and the number of transfers). Similarly to the CSA extension, the cheapest price to reach each trip segment is maintained: the data structure $R(t)$ which previously maintained for each trip the first reachable stop now holds the cheapest price to reach each stop of the trip with the corresponding trip.

Algorithm The algorithm now tracks the price of each queue entry. Updating trip t entails maintaining the cheapest prices in $R(t)$ as well as the cheapest prices of all trips of the same route with later departure times.

Pruning We extend the implementation to track the latest arrival time and most expensive price over each target station. Journeys exceeding these limits are discarded and therefore not added to the queue to be processed in the next iteration.

7 Computational Study

Our C++ implementation (compiler: LLVM/Clang 6 with "-O2" optimizations) of the presented algorithms was evaluated on a computer with an Intel Core i7 6850K (6x 3.6GHz) CPU and 64GB main memory. The public transport timetable was provided by Deutsche Bahn and covers all services (busses, trams, trains, etc.) operated in Germany. For foot and car routing the complete OpenStreetMap dataset of Germany was loaded.

The timetable spans the 27th and 28th of November 2018. It contains approximately 30M departure and arrival events (60M events total) that take place in 1.7M trips on 224,832 routes.

Queries are generated by choosing a random t_1 and a random t_2 30min to 4h after t_1. To generate coordinates that yield a high chance of non-empty result sets, we randomly select a public transport station that has at least one arrival event in the time interval $[t_1 - 60\text{min}, t_1]$ and at least one departure event in the time interval $[t_2, t_2 + 60\text{min}]$. Then, a random coordinate in a radius of w_{max} around this station is selected as ω. α is a random coordinate located in a 200km radius around destination. Both coordinates need to be within Germany which is checked with the help of a polygon that resembles Germanys borders.

7.1 Preprocessing

Extracting all parking places and calculating optimal foot paths between public transport stations and nearby parking places takes 41 minutes and 44 seconds. However, this needs to be done only once for every dataset. At runtime, a fast lookup table with the precomputed foot path times is used. Our OpenStreetMap dataset contains 319,361 parking places. On average, 5.18 stations are reachable from a parking place (median 4, 99% quantile 23).

The execution of the preprocessing step described in Section 4.3 takes place at query runtime. Street routing between α and all parking places in the selected radius takes 383ms. The lookup times for stations/parkings in a specified radius around a coordinate are negligible (below 1ms). Lookup of precomputed foot routes between parking places and nearby

public transport stations takes 3.68ms. Since ω is a user input, foot paths between ω and nearby public transport stations (set D) cannot be precomputed. Computing W takes 27.4ms at runtime. The sets W, D and C can be computed in parallel. So in total, 387ms of the runtime are due to preprocessing. The next sections report runtimes including preprocessing times. Therefore, to obtain the total core routing runtime, approximately 0.4s need to be subtracted from the runtimes reported below.

7.2 Baseline Algorithms

Table 1: Runtimes of Baseline Algorithms without Price Optimization in Milliseconds

	avg	Q(99)	Q(90)	Q(80)	Q(50)
Baseline	659 700	2 732 816	1 676 236	924 110	398 985
Combined	399 353	1 455 144	868 975	630 930	258 519
Parallel	276 666	761 288	561 843	444 319	215 487
Comb. Par.	193 793	624 920	402 261	301 287	159 545

As depicted in Table 7.2, parallel execution of the baseline approach yields a reasonable 2.4x speedup on average. The "trick" of an integrated optimization for one of the two directions (including parallel execution for the non-integrated search direction) yields another 2x speedup on average. Nonetheless, the baseline approach and its variations described in Section 5.1 (parallel implementation) and Section 5.1 (combined search) are not really of any practical use because they require many invocations of the time dependent routing routine. Users of online services are not eager to wait more than three minutes for their routing result. However, due to their simplicity those approaches are useful for validation of the other implementations.

7.3 Advanced Algorithms

Table 2: Runtimes of Advanced Algorithms in Milliseconds

	avg	Q(99)	Q(90)	Q(80)	Q(50)
No Terminal Dominance	4800	11 430	7969	6615	4403
Worst Bound	4762	11 268	8042	6564	4363
Concurrent	3573	10 305	6439	5000	3156
CSA	1697	3384	2322	1999	1577
CSA SIMD	908	2453	1353	1113	806
TripBased	816	2644	1302	975	689

In this section, we present the results of the advanced algorithms: different Dijkstra-based algorithms (with worst bounds, without terminal dominance and the interleaved / concurrent approach) on the time dependent graph model (introduced in Section 5.1), Connection Scanning (Section 5.2), and TripBased routing (Section 5.3). Additionally, we have implemented a CSA version that makes use of SIMD instructions.

As we can see in Table 7.3, the concurrent (interleaved) search in both directions (outward and return trip) brings the runtimes on the time dependent graph down from more than three minutes to 3.5 seconds. However, one percent of the queries take more than 10 seconds to answer. All non-graph-based approaches (CSA and TripBased) yield better runtime performance: the CSA SIMD variant as well as the trip based routing have average runtimes under one second. The data-parallel SIMD implementation of the CSA algorithm yields nearly a 2x speedup compared to the basic CSA version. Note that the CSA SIMD version is even faster than the TripBased approach when it comes to the 99% quantile (2.45 seconds vs 2.64 seconds) indicating that it has a more predictable performance profile.

Parking Radius In this section, we analyze the relation of the runtime of the approaches presented in this paper with the $d_{\max}$ parameter which essentially determines the number of parkings to consider. We analyze this relation for two and three optimization criteria.

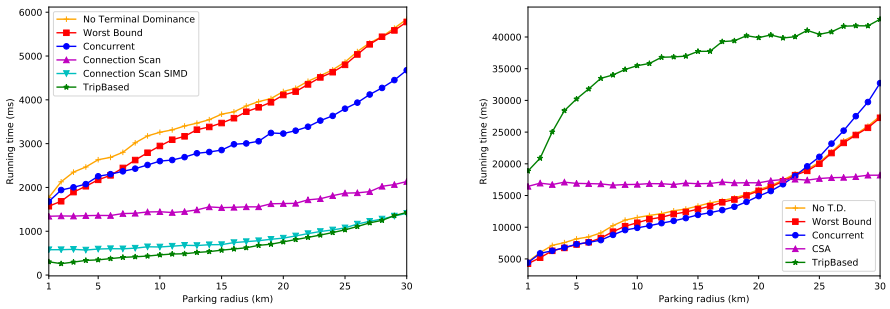


Figure 1: Runtime Subject to Parking Radius Distance $d_{\max}$: the left figure shows the basic optimization with travel time and number of transfers. The right figure shows the runtime for optimization with all three criteria: travel time, number of transfers and price.

As we can see in Figure 1, the runtime scales mostly linearly with the parking radius for all approaches regardless of the optimization criteria. However, the increase in runtime is different for the presented approaches without price optimization: while the Connection Scan SIMD and TripBased runtimes rise minimally with an increased parking radius and stay below one second, Concurrent and Worst Bound runtimes have a steep increase with a growing parking radius.

Note the different ordinate scale of the right graph of Figure 1: price optimization imposes a heavy toll on query runtime. When optimizing prices, the runtimes of the graph based approaches (Worst Bound and Concurrent) for short distances are better than those of CSA and TripBased. However, this changes for $d_{\max}$ values greater than 23km where CSA delivers the fastest (almost constant) runtimes. A mixed approach could pick a graph based algorithm for smaller radii and switch to CSA for larger radii. As the TripBased algorithm is tailored to two optimization criteria (travel time and number of transfers), the runtimes with three optimization criteria lack behind the other approaches.

Distance Analysis In Table 7.3 we see that the runtimes of the CSA and TripBased approaches are insensitive to changing distances between α and ω. All runtimes of graph

Table 3: Runtimes for Different α/ω Distances in Milliseconds

	50km	200km	900km
No Terminal Dominance	2236	4800	5308
Worst Bound	2234	4762	5263
Concurrent	1544	3573	4876
CSA	1778	1697	1656
CSA SIMD	952	908	853
TripBased	878	816	730

based approaches grow with larger distances. Note that for short distances, the average runtimes of the Concurrent approach are lower than those of the basic CSA approach. This is not the case anymore for higher distances.

7.4 Price Optimization

Table 4: base scenario, no price vs. simple price vs. regional price

	no price	simple price	regional price
No Terminal Dominance	4800	19 505	17 492
Worst Bound	4762	19 249	17 316
Concurrent	3573	17 718	15 141
CSA	1697	18 314	23 245
TripBased	816	40 670	39 542

In this section, we analyze the impact price optimization has on the runtimes of the different approaches. We evaluated both public transport price models introduced in Section 6: one is based only on distance and vehicle class (called "simple" in this section), the other model adds a special regional ticket with a flat price (called "regional price"). The changed route definition (described in Section 6) leads to 0.61% more routes. As we can see in Table 4, this additional search criterion increases the runtimes of all algorithms between 4x and 50x compared to the two criteria implementations (regardless of the concrete pricing model).

Since the price optimizing implementation of TripBased disabled most of the speedup it gained through the preprocessing (transfers reduction), it switched from being the fastest implementation to being the slowest implementation. Note that while Worst Bound and Concurrent could gain more than 10% speedup through the regional price model, CSA was slowed down by it (by more than 25%).

8 Conclusion

We presented several novel approaches to compute Pareto optimal solutions to the 2-way park and ride roundtrip problem. In addition to two criteria optimization (travel time and number of transfers), we introduce variants of the approaches which additionally optimize

the journey price. Since many journeys follow this pattern (e.g. for commuters), the developed algorithms are useful in practice. The approaches are based on state-of-the art algorithms for public transport routing such as TD (Disser et al., 2008), CSA (Dibbelt et al., 2013) and TripBased routing (Witt, 2015). Our evaluation on a dataset covering all of Germany shows that the approaches offer query runtimes below 2 seconds which makes them suitable for use in online or mobile app information systems.

References

Baumann, D., Torday, A., and Dumont, A.-G. (2004). The importance of computing intermodal roundtrips in multimodal guidance systems. In *Proceedings of the 4th STRC Swiss Transport Research Conference*, number LAVOC-CONF-2008-029.

Bousquet, A., Constans, S., and El Faouzi, N.-E. (2009). On the adaptation of a label-setting shortest path algorithm for one-way and two-way routing in multimodal urban transport networks. In *International Network Optimization Conference*.

Delling, D., Dibbelt, J., Pajor, T., Wagner, D., and Werneck, R. F. (2013). Computing multimodal journeys in practice. In *Proceedings of the 12th International Symposium on Experimental Algorithms (SEA'13)*, volume 7933 of *Lecture Notes in Computer Science*, pages 260–271. Springer.

Delling, D., Pajor, T., and Wagner, D. (2009). Accelerating multi-modal route planning by access-nodes. In *European Symposium on Algorithms*, pages 587–598. Springer.

Delling, D., Pajor, T., and Werneck, R. F. (2012). Round-based public transit routing. In *2012 Proceedings of the Fourteenth Workshop on Algorithm Engineering and Experiments (ALENEX)*, pages 130–140. SIAM.

Dibbelt, J., Pajor, T., Strasser, B., and Wagner, D. (2013). Intriguingly simple and fast transit routing. In *International Symposium on Experimental Algorithms*, pages 43–54. Springer.

Disser, Y., Müller-Hannemann, M., and Schnee, M. (2008). Multi-criteria shortest paths in time-dependent train networks. In *International Workshop on Experimental and Efficient Algorithms*, pages 347–361. Springer.

Huguet, M.-J., Kirchler, D., Parent, P., and Calvo, R. W. (2013). Efficient algorithms for the 2-way multi modal shortest path problem. *Electronic Notes in Discrete Mathematics*, 41:431–437.

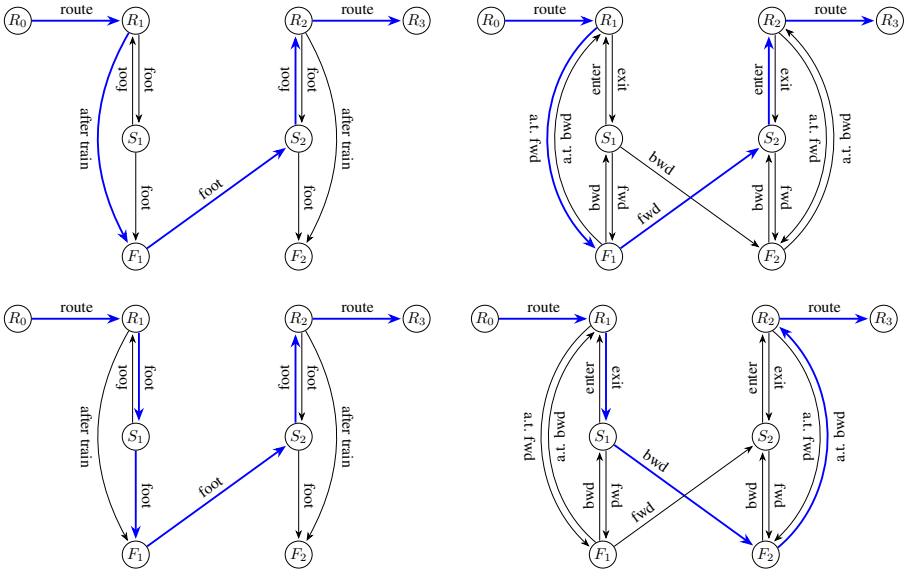
Luxen, D. and Vetter, C. (2011). Real-time routing with openstreetmap data. In *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '11, pages 513–516, New York, NY, USA. ACM.

Spinatelli, S. (2015). Minimal effective time two-way park and ride problem. Master's thesis.

Witt, S. (2015). Trip-based public transit routing. In *Algorithms-ESA 2015*, pages 1025–1036. Springer.

A Changes to the Time Dependent Graph Model by Disser et al. (2008) to Support Latest Departure Queries

Figure 2: Changes to the Time Dependent Graph Model to Support Backward Search: The new model (right side) fixes the inconsistency (different costs for forward and backward search) of the old model (left side).



As we can see in Figure 2 (edge costs are listed in Table 5), the basic time dependent model presented in (Disser et al., 2008) is not consistent (i.e. equal graph costs for the same journey in forward and backward search) for routes containing walks between nearby stations: in the backward search the path includes the transfer costs of S_2 while in the forward search no transfer costs are included (which is the desired behavior). In the fixed model, a walk between two stations has the same costs in forward and backward search direction.

Table 5: Edge Type Costs for Forward and Backward Search in the Time Dependent Graph as (Travel Time, Transfer Count) tuples: costs marked with a star "*" are not feasible at edge expansion if the corresponding label did not use a route edge before. The symbol $\oslash$ indicates that the edge is not feasible in this search direction. ic_s is the transfer time for interchanges at station $s \in S$.

Edge Type	Forward Search	Backward Search
enter	$(0, 0)$	(ic_s, 1)*
exit	(ic_s, 1)*	$(0, 0)$
after train forward	$(0, 1)$*	$\oslash$
after train backward	$\oslash$	$(0, 1)$*
fwd	(x, y)	$\oslash$
bwd	$\oslash$	(x, y)

A train timetabling and stop planning optimization model with passenger demand

Weining Hao[a], Lingyun Meng[a,1], Francesco Corman[b], Sihui Long[a,2], Xi Jiang[a,3]

[a] School of Traffic and Transportation, Beijing Jiaotong University
No.3 ShangYuanCun, HaiDian District, Beijing 100044, China
[1] Email: lymeng@bjtu.edu.cn, Phone: (86)-10-51688520
[b] Institute for Transport Planning and Systems (IVT), ETH Zurich
Stefano-Franscini-Platz 5, 8093 Zurich, Switzerland

Abstract

Train timetabling plays an important part in train management, not only for passengers, but also for train operators. In a highly dynamic transportation market, train timetabling is an essential bridge connecting the service supplier with transportation demand. However, in present operations, train scheduling without considering passenger demand can reduce competitive advantages of railway in the multimodal transportation market and will further lead to passenger dissatisfaction. Therefore, it's important to schedule trains responding to passenger demand in the train planning process. In this paper, we focus on the problem of train timetabling with passenger demand, specifically deciding train stop plan based on different origin-destination passenger demand pairs. Taking the stop indicators as important decision variables, a mixed integer linear programming model is proposed to address this train timetabling and stopping plan integration issue, with minimizing total train travel time and maximizing the number of transported passengers. The weighted-sum method is used to find the Pareto optimal solutions for the proposed bi-objective mathematical model. A set of numerical tests is presented based on Beijing-Jinan high-speed railway line (part of Beijing-Shanghai high-speed railway line) by Cplex optimization solver to validate the model.

Keywords

Train timetabling, Stop planning, Passenger demand, Mixed integer programming, Pareto optimization

1 Introduction

In the rapidly changing multimodal transportation market with intense competition, various transportation modes make efforts to enlarge their own service scope. Providing punctual and flexible service considering passenger demand is especially essential for railway transportation to improve its competitiveness and increase market share in such a situation. An effective train operation plan can provide better service for passengers who choose railway transportation to complete their trips. Due to the growing passenger demand of railway, train operators incline to plan train schedule considering the nature of passengers instead of assuming that passengers will adjust their behaviours to the provided train service. Hence, the scheduling process for railway system has been more and more significant for ensuring punctuality of train operation and for guaranteeing passenger satisfaction.

To provide passenger oriented train service, the key of train scheduling is to meet

passenger demand while reducing the cost of operation and management. This complex task requires a comprehensive consideration of passenger demand patterns and train unit resources. For scheduling with passenger demand, urban rail operation under passenger demand concentrates on minimizing passengers' waiting time at metro stations instead of highlighting the origin and destination of passengers, since metro train always stops at each station. While railway pays closer attention to whether there are enough trains to take these passengers at the station as many as possible and how to schedule these trains in an economic way, such as determining stopping plan and frequency. Therefore, from this point, the train timetable and train stops are both determined according to the passenger demand.

Train timetabling and stop planning are regarded as two critical parts in train scheduling. In tradition, these two parts are separated because there is a sequential planning process that is divided in several steps when schedule trains, as Fig.1 shows. Generally, each previous step is taken as an input of the latter one. After a demand analysis, line planning determines train service frequency and different stopping plans of each train to meet passenger demand, also constrained by infrastructure. Then, based on line plan, the train timetable is given to determine the departure time and arrival time of each train at each station, and provides a foundation of rolling stock schedules and crew schedules. At the same time, the latter two process may need to adapt the departure/arrival times of the obtained train schedule. However, in the real world, the adjusted stopping plan and train timetable might not be the best solution for train operators as well as might not meet passenger demand.

In this work, we focus on the integrated optimization problem of train timetabling and train stop planning (ITTSP), which embeds the train stopping planning constraints, based on potential passenger flow for different origin-destination pairs, into the train timetabling stage. To solve this ITTSP problem, a bi-objective mixed integer linear programming model is formulated, in which passenger demand with different origin and destination stations, train stop planning, train routing and train timetabling are included in the model formulation. A weighted-sum method solution approach is then used to solve the resulting integrated optimization problem, where both the objective functions are directly optimized proportionally to the assigned weights.

The reminder of this paper is organized as follows. Section 2 provides a literature review on demand oriented train timetabling and on the integration of train stop planning and timetabling. Then, a detailed problem statement and model assumptions are given first, followed by a bi-objective model that formulates the ITTSP problem based on passenger demand in Section 3. Next, a weighted-sum method is introduced to solve the resulting ITTSP problem. To evaluate the effectiveness of bi-objective model, a case study based on Beijing-Jinan high-speed railway line (BJ railway line) is tested in Section 4. Finally, conclusions and future research are presented in Section 5.

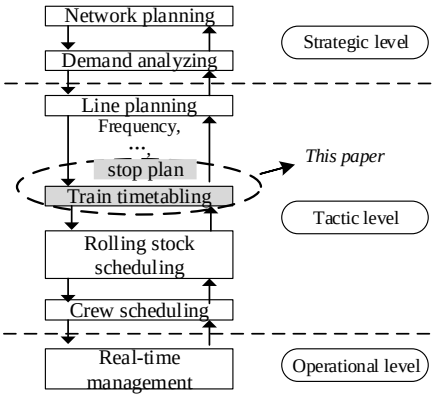


Fig.1: The sequential railway planning process

2 Literature review

In this section, we review the state of the art in two directions: 1) train timetabling considering passenger demand; 2) integrated optimization of train timetabling and stop planning.

Train timetabling plays an important role in train scheduling operation. Several researchers (Szpigel, 1973; Higgins, Kozan and Ferreira, 1996; Caprara et al., 2002; Caprara et al. 2006; Zhou and Zhong, 2007) made some related investigations on this problem. Based on the job-shop scheduling problem, Szpigel (1973) firstly modeled the single-track train scheduling problem to determine the location of crossing and overtaking. Higgins, Kozan and Ferreira (1996) scheduled trains optimally on a single line track and presented a lower bound to reduce the search space in the branch and bound tree. Caprara et al. (2002) and Caprara et al. (2006) gave a graph description of the train timetabling problem with fixing train running time and headway time, while it is not practical due to train acceleration and deceleration. Zhou and Zhong (2007) used branch-and-bound solution algorithms to solve a single-track train timetabling problem and generalized station headway capacities-constrained scheduling formulation.

In recent years, studies on demand-sensitive train timetabling have attracted more and more attention, in order to provide higher level of the train service for transportation demand, especially for the passenger demand. Sun et al. (2014) provided a demand-driven timetable for metro services, which adjusted service frequency dynamically instead of setting it for peak/off-peak time respectively, aiming at minimizing the total passenger waiting time. Canca et al. (2014) put variable demand in a long period into their timetabling model, where vehicle capacity was considered to generate effective solution quality. Barrena et al. (2014a) proposed three exact linear formulations and the branch-and-cut algorithm to design train timetables consistent with dynamic demand. In order to solve large-scale instances, Barrena et al. (2014b) presented an adaptive large neighborhood search (ALNS) meta-heuristic that was able to solve larger and more realistic instances. Niu, Zhou and Gao (2015) proposed quadratic and quasi-quadratic objective functions to formulate total passenger waiting time based on time-varying origin-to-destination demand. Wang et al. (2018) concentrated on time-varying passenger demand of the segment between two adjacent stations to integrate

train scheduling and rolling stock circulation planning on an urban rail transit line. Robenek et al. (2018) formulated a passenger centric train timetabling problem under elastic passenger demand and used a logit model to reflect the unknown demand elasticities. Researchers who studied the problem of passenger demand oriented train scheduling were mostly concerned with adjusting train timetable, but line planning is another essential part reflected by passenger demand. Optimizing both line planning and train timetable can better adapt to passenger demand in practice. In the stage of line planning, train stop planning is of particular importance.

In the literature, most existing researches focused on planning train stop plan. Lan (2002) explained that different stopping programs should be included when designing operation plans for Beijing-Shanghai high-speed railway. Besides, Cheng and Peng (2014) developed a 0-1 bi-level mathematical programming model for urban rail transit special stop schedule scheme, considering elastic passenger demand. Yue et al. (2016) optimized train stopping patterns and schedules for high-speed passenger trail corridors and developed an innovative methodology using a column-generation-based heuristic algorithm to simultaneously consider passenger demand and train scheduling. Yang et al. (2016) proposed a new collaborative optimization method for train scheduling and stop planning problem and handled it through linear weighted method, where the model considered the satisfaction of macro demands on each station. Qi, Cacchiani and Yang (2018) emphasized uncertain passenger demand and aimed to determine both train timetable as well as stop plan.

Different from metro rail with all-stop operations, in railway operation plan, passenger demand has a straightforward influence on train stop patterns. Although the all-stop operation is obviously the simplest way for satisfying passenger demand, it may take long-distance passengers' travel time as an extra cost. Therefore, the integration of passenger demand oriented train timetabling and stop planning is a hot researching direction. Nevertheless, the integrated optimization of train timetabling and stop planning with passenger demand could put stress on the computation time and model difficulty.

In this paper, we take passenger demand into train timetabling and highlight the relationship between stop plan and passenger demand pairs with different origin-destination stations, in order to design a train timetable consistent with demand. This paper proposes the following contributions:

- This work embeds passenger demand into the timetabling phase by choosing the train stop on its travel route. A bi-objective linear programming model is proposed, rigorously considering passenger demand constraints.
- We combine train scheduling and stop planning with passenger demand to generate a train timetable and stop plan simultaneously. The objectives of the model we proposed are to minimize total train travel time, in order to reduce the management costs for rail operators, and to maximize the number of transported passengers, in order to better satisfy passenger demand.

3 Problem statement and model

3.1 Problem statement

Before the mixed-integer linear programming model is described, the problem statement and model assumptions are given sequentially. First, inputs of this problem are explained as below:

(1) A railway network

A railway network is given with a number of stations and segments between adjacent stations, in which the segment between two adjacent stations is set as one section, the station is set including specific siding tracks.

(2) Train information

For each train, we know its origin and destination station, the earliest starting time at its origin station, running time between two adjacent stations, minimum and maximum dwell time at intermediate stations, headway time of two consecutive trains, train carrying capacity, characteristics (i.e. train type).

(3) Passenger demand

We consider passenger demand, in this paper, as different sets of passenger pairs who have different origin and destination stations. For each passenger pair, we know its origin station, destination station and volume.

The ITTSP problem has three decision variables:

(1) For each train, its stopping plan needs be determined, that is whether the train chooses to stop and how long it will stop at this station.

(2) For each train scheduled, we need to determine its departure time at the origin station, the arrival, dwell time, and departure time at intermediate stations, as well as the arrival time at the last station.

(3) For each passenger pair, we need to determine which train it is assigned to and how many of passengers in the passenger pair are assigned.

In our model, we make the following assumptions:

(1) In this paper, our purpose is to provide a train schedule to satisfy passenger demand from the view of train operators, so the response of passenger behaviours to the resulting train service is not included.

(2) In our proposed model, the station dwell time occurs only if the train is required to stop due to the passenger demand, and minimum dwell time is fixed without changing with passenger flow variation at a station.

(3) We assume that the station can accommodate enough trains, which means that the station capacity is not considered.

The general subscripts and input parameters of the proposed formulations are introduced in Table 1 and 2, respectively, and the decision variables are given in Table 3.

Table 1: General subscripts

Symbol	**Description**
i, j, k	Physical node index, $i, j, k \in N$, N is the set of nodes in a railway network.
e	Physical cell index, $e \in E$, E is the set of cells in a railway network.
t, t'	Time index, $t, t' \in \{1...T\}$, T is the planning horizon, e.g. 3 hours.
p	Passenger origin-destination (OD) pair index, $p, p' \in P$, P is the set of passenger OD pairs. One passenger OD pair refers to a group of passengers who have the same origin and destination stations.
f	Train index, $f \in F$, F is the set of all trains which need to be scheduled.
m, m'	Station index, $m, m' \in S$, S is the set of all stations in a railway network.

Table 2: Input parameters

Symbol	Description
E_f	Set of cells train *f* may use, $E_f \subset E$.
E_c	Set of cells of sections between two adjacent stations, $E_c \subset E$.
E_m	Set of cells of station *m*, $E_m \subset E$.
E_i^{o}	Set of cells starting from node *i*.
E_i^{s}	Set of cells ending at node *i*.
$\delta_{f,i,j}$	Free-flow running time for train *f* to drive through cell (*i*, *j*).
$\vartheta_{f,i,j}^{\min}$	Minimum dwell (waiting) time for train *f* on cell (*i*, *j*).
$\vartheta_{f,i,j}^{\max}$	Maximum dwell (waiting) time for train *f* on cell (*i*, *j*).
$g_{f,i,j}$	Safety time interval between train *f*'s occupancy and arrival on cell (i, j).
$h_{f,i,j}$	Safety time interval between train *f*'s departure and release on cell (i, j).
$c_{i,j,t}$	Flow capacity on cell (i, j) at time *t*.
O_f	Origin node of train *f*.
S_f	Destination (sink) node of train *f*.
O_m	Origin node of station *m*.
S_m	Destination (sink) node of station *m*.
m_p^o	Origin station of passenger pair *p*.
m_p^s	Destination station of passenger pair *p*.
EST_f	Predetermined earliest starting time of train *f* at its origin node.
η_p	Volume of passenger OD pair *p*.
C_f	The capacity of train *f*
α	The maximum passenger carrying coefficient of the scheduled trains.

Table 3: Decision variables

Symbol	Description
$a_{f,i,j,t}$	0-1 binary train arrival variables, =1 if train *f* has already arrived at cell (*i*, *j*) by time *t*; =0 otherwise.
$d_{f,i,j,t}$	0-1 binary train departure variables, =1 if train *f* has already departed from cell (*i*, *j*) by time *t*; =0 otherwise.
$u_{f,i,j,t}$	0-1 binary infrastructure usage variables, =1, if train *f* occupies cell (*i*, *j*) at time *t*; =0 otherwise.
$x_{f,i,j}$	0-1 binary train routing variables, =1, if train *f* selects cell (*i*, *j*) on the network; =0 otherwise.
$y_{f,p}$	Passenger assignment variables, passenger volume of passenger OD pair *p* that is assigned to train *f*.
$z_{f,p,i,j}$	Passenger assignment variables on cell (*i*, *j*), passenger volume of passenger OD pair *p* on cell (*i*, *j*) that is assigned to train *f*.
$r_f^{m,m'}$	0-1 binary train stopping variables, =1, if train *f* stops at both station *m* and *m'*, =0 otherwise.
$k_{p,i,j}$	0-1 binary passenger travel route variables, =1, if passenger pair *p* travel on cell (*i*, *j*), =0 otherwise.
$TT_{f,i,j}$	Travel time for train *f* on cell (*i*, *j*).

3.2 Formulation of the mathematical model

The objective function consists of two parts: one is to maximize the number of transported passengers that are carried by planned trains.

$$Z_{passenger} = \sum_{p \in P} \sum_{f \in F} y_{f,p} \tag{1}$$

Another one is to minimize total train travel time from its origin station to destination station.

$$Z_{time} = \sum_{f \in F} \sum_{t \in T} \left[t \times \sum_{i:(i,s_f) \in E^s_{S_f} \cap E_f} (a_{f,i,s_f,t} - a_{f,i,s_f,t-1}) - t \times \sum_{j:(O_f,j) \in E^o_{O_f} \cap E_f} (d_{f,O_f,j,t} - d_{f,O_f,j,t-1}) \right] \tag{2}$$

The bi-objective function can be presented as:

$$Z = \max Z_{passenger} + \min Z_{time} \tag{3}$$

Subject to:

Group 1: Train running constraints

Train timetabling is actually to determine travel routes of each train on a time-space network, so, based on it, cumulative flow variables (Meng and Zhou, (2014)) $a_{f,i,j,t}$ and $d_{f,i,j,t}$ are introduced to represent both temporal and spatial consumption of trains.

In the network, trains' start is restricted. For trains' start time, constraint (4) and (5) make sure that each train do not depart earlier than predetermined earliest starting time at their origin nodes. Within cell to cell transition, to guarantee the passing time at each cell, the time when train f departs from the forward cell (i, j) and arrives at the later cell (j, k) should be the same.

$$\sum_{j:(o_f,j)\in E_f} a_{f,o_f,j,t} = 0, \forall f \in F, t < EST_f \tag{4}$$

$$\sum_{j:(o_f,j)\in E_f} d_{f,o_f,j,t} = 0, \forall f \in F, t < EST_f \tag{5}$$

$$\sum_{i,j:(i,j)\in E_f} d_{f,i,j,t} = \sum_{j,k:(j,k)\in E_f} a_{f,j,k,t}, \forall f \in F, j \in N \setminus \{o_f, s_f\}, t = 1,...,T \tag{6}$$

In this train scheduling problem, all trains are supposed to meet the flow balance when trains run in the railway network. In this model, we separate nodes in a network into three parts (origin node, intermediate node and destination node) to explain the flow balance problem. At the origin node and destination node, there is only one routing choice for train f to go through. Constraint (7)-(9) ensure flow balance on the network at the origin node, intermediate nodes, and the destination node of train f respectively.

$$\sum_{j:(i,j)\in E^o_{of}\cap E_f} x_{f,i,j} = 1, \forall f \in F \tag{7}$$

$$\sum_{i:(i,j)\in E^s_j\cap E_f} x_{f,i,j} = \sum_{k:(j,k)\in E^o_j\cap E_f} x_{f,j,k}, \forall f \in F, j \in N \setminus \{o_f, s_f\} \tag{8}$$

$$\sum_{i,j:(i,j)\in E^s_{sf}\cap E_f} x_{f,i,j} = 1, \forall f \in F \tag{9}$$

Constraints (10) is imposed to map the variables $a_{f,i,j,t}$ in space-time network to the variables $x_{f,i,j}$ in physical network, so as to describe whether cell (i, j) is selected by train f for traversing the network from its origin to destination.

$$x_{f,i,j} = a_{f,i,j,T}, \forall f \in F, (i,j) \in E_f \tag{10}$$

Here, we use decision variables $a_{f,i,j,t}$ and $d_{f,i,j,t}$ to represent running time $TT_{f,i,j}$, which is the difference of exit time and entrance time for train f on cell (i, j), as constraint (11) shows.

$$TT_{f,i,j} = \sum_t \{t \times [d_{f,i,j,t} - d_{f,i,j,t-1}]\} - \sum_t \{t \times [a_{f,i,j,t} - a_{f,i,j,t-1}]\}, \forall f \in F, (i,j) \in E_f \tag{11}$$

In practice, due to station stops and some unexpected disturbance, such as bad weather, total travel time on cell (i, j) must be equal or larger (smaller) than its free flow travel time plus it minimum (maximum) planned dwell time at the station. Constraint (12) specifies it in an inequality. The minimum planned dwell time is larger than zero, only if there is a train stop at a station in the timetable.

$$\vartheta^{\min}_{f,i,j} + \delta_{f,i,j} \le TT_{f,i,j} \le \vartheta^{\max}_{f,i,j} + \delta_{f,i,j}, \forall f \in F, (i,j) \in E_f, p \in P \tag{12}$$

When train stops at s station, train acceleration and deceleration operations can occur in many real-world cases. In order to formalize them in train timetabling problem, the occupancy for train f on cell (i, j) is used by introducing extra running times. Constraint (13) links $u_{f,i,j,t}$ with $a_{f,i,j,t}$ and $d_{f,i,j,t}$. Hence, train f contributes a value of 1 to the

occupancy on cell (i, j) when it has arrived at cell ($a_{f,i,j,t+g}$ =1) but not departure from it by time t ($d_{f,i,j,t-h}$ =0). Furthermore, the number of trains that occupies the same cell (i, j) is limited by the capacity of cell (i, j) to avoid conflicts in railway stations. Usually, the capacity of cell (i, j) in station is set as 1.

$$u_{f,i,j,t} = a_{f,i,j,t+g} - d_{f,i,j,t-h}, \forall f \in F, (i, j) \in E_f, t = 1,...,T \tag{13}$$

$$\sum_{f:(i,j)\in E_f} u_{f,i,j,t} \le c_{i,j,t}, \ \forall (i, j) \in E_m, t = 1,...,T \tag{14}$$

To better describe the train time-dimension routing in a railway network using cumulative flow variables, we give definitional constraint (15) and constraint (16). Specifically, if train f has arrived or departed on cell (i, j) by time t , $a_{f,i,j,t}$ and $d_{f,i,j,t}$ will have a value of 1 for all later time periods.

$$a_{f,i,j,t} \ge a_{f,i,j,t-1}, \forall f \in F, (i, j) \in E_f, t = 1,...,T \tag{15}$$

$$d_{f,i,j,t} \ge d_{f,i,j,t-1}, \forall f \in F, (i, j) \in E_f, t = 1,...,T \tag{16}$$

Group 2: Passenger assignment constraints

For each passenger pair, the total number of passengers carried by planned trains should be no more than the volume of passenger pair. Besides, for each train scheduled, the total number of passengers that can be assigned to a train is limited by its maximum passenger carrying capacity. Constraint (19) is a mapping constraint between $z_{f,p,i,j}$ and $k_{p,i,j}$.

$$\sum_{f\in F} y_{f,p} \le \eta_p , \forall p \in P \tag{17}$$

$$\sum_{p:p\in P} z_{f,p,i,j} \le \alpha \times C_f, \forall f \in F, (i, j) \in E_c \tag{18}$$

$$z_{f,p,i,j} = y_{f,p} \times k_{p,i,j}, \forall f \in F, p \in P, (i, j) \in E_c \tag{19}$$

Group 3: Mapping constraints between passenger assignment and stopping pattern

The following constraint presents that if train f carry passenger pair p , the train stops at both the origin station and destination station of pair p . It is a constraint between passenger demand and stop planning.

$$y_{f,p} \le r_f^{m_p^o, m_p^s}, \forall f \in F, p \in P \tag{20}$$

Further, if train f stops at the origin station and destination station of passenger pair p ($r_f^{m_p^o, m_p^s} = 1$), the departure time and arrival time for train f departing/arriving at each station is not equal, in order to provide waiting time for trains to stop at a station. Constraint (21) and constraint (22) enforced the waiting time for train f at the origin station and destination station of pair p respectively.

$$\sum_{t\in T}\left[t\times\sum_{i:(i,S_{m_p^o})\in E^s_{S_{m_p^o}}\cap E_f}(d_{f,i,S_{m_p^o},t}-d_{f,i,S_{m_p^o},t-1})-t\times\sum_{j:(O_{m_p^o},j)\in E^o_{O_{m_p^o}}\cap E_f}(a_{f,O_{m_p^o},j,t}-a_{f,O_{m_p^o},j,t-1})\right]\succ \quad (21)$$

$$M\times(r_f^{m_p^o,m_p^s}-1),\forall f\in F$$

$$\sum_{t\in T}\left[t\times\sum_{i:(i,S_{m_p^s})\in E^s_{S_{m_p^s}}\cap E_f}(d_{f,i,S_{m_p^s},t}-d_{f,i,S_{m_p^s},t-1})-t\times\sum_{j:(O_{m_p^s},j)\in E^o_{O_{m_p^s}}\cap E_f}(a_{f,O_{m_p^s},j,t}-a_{f,O_{m_p^s},j,t-1})\right]\succ \quad (22)$$

$$M\times(r_f^{m_p^o,m_p^s}-1),\forall f\in F$$

3.3 Solution approach

Regarding the two objectives of our proposed model, one is to maximize the number of transported passengers to get on the trains from the view of passengers, and another one is to minimize total travel time from the view of train operators. When maximizing the passengers, the train has to stop to meet passenger demand and the dwell time that can increase trains travel time will occur. On the other hand, when reducing train travel time as much as possible, there will be some passengers that fail to take the train service. The problem is that these two aims are associated by different stakeholders with different cost functions (i.e. tickets) and economic interests.

The multi-objective optimization problem has been widely used in railway management. Some researches enforced ε-constraint method to solve it. (Ghoseiri et al. (2014); Yang et al. (2017); D'Ariano et al. (2017)). Meanwhile, many studies adopted weighted-sum method to handle the multi-objective model and generate the Pareto solutions. Burdett et al. (2015) used the weighted-sum method to analyse the absolute capacity in railway networks. Yang et al. (2016) optimized train scheduling problem on high-speed railway through linear weighted methods. D'Ariano et al. (2017), based on weighted-sum method, developed a formulation to integrate train scheduling and railway infrastructure maintenance.

Based on the existing literature of solving multi-objective problem, we apply a formulation that the objective functions are optimized by setting different assigned weights. It is achieved by two input parameters $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$ fixed by the decision maker. And the parameters are constrained: $\boldsymbol{\alpha}_1 \geq 0$, $\boldsymbol{\alpha}_2 \geq 0$, $\boldsymbol{\alpha}_{1+}\ \boldsymbol{\alpha}_2=1$. Therefore, the Pareto solutions can be obtained by varying $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$ to satisfy different demands. In this approach, we first get the results of f_1 and f_2, where $f_1= min\ Z_{time}$, and $f_2= max\ Z_{passsenger.}$ Then, set $m_1= \boldsymbol{\alpha}_1/\ f_1$, $m_2= \boldsymbol{\alpha}_2/\ f_2$, $Z= m_1 * Z_{time} - m_2 * Z_{passsenger.}$ Finally, replace the objective function with $N=min\ Z$, restricted by constraint (4)-(22).

4 Case study

In this section, we first describe the dataset in Section 4.1 and then we demonstrate experimental results in 4.2.

We adopt the CPLEX solver version 12.6.3 with default settings to solve the MILP models. The following experiments are all performed on a server with two Intel(R) Xeon(R)

CPU E5-2660 v4 @ 2.00GHz 2.00GHz processors and 512GB RAM.

4.1 Description of the test dataset

To evaluate the effectiveness of the model, we performed numerical experiments on a railway corridor (BJ railway line) with 6 stations of Beijing-Shanghai high-speed railway line, as shown in Fig.2. To determine the route in railway station, we illustrate BJ railway network in appendix. In BJ railway network, only the down direction is considered for simplicity. In this experiment, a total of 7 trains will be taken into consideration and a total of 15 passenger pairs among these stations is included. We assume that the start time of the first train is at 8:03 and the minimum time interval between two consecutive trains is set as 9 min. Besides, the minimum and maximum dwell time at its stop is fixed as 2 min and 5 min respectively to ensure the necessary operation time if the train needs to stop. The maximum passenger carrying coefficient of trains scheduled in this work is all set as 1.2. Detailed train information can be seen from Table 4.

Fig.2: BJ railway line

Table 4: Train origin/destination station and carrying capacity in the test

Number of trains	Train number	Origin station	Destination station	Passenger carrying capacity
7	No.1	BJS	JNW	535
	No.2	BJS	JNW	535
	No.3	BJS	JNW	450
	No.4	TJS	JNW	400
	No.5	BJS	JNW	450
	No.6	TJS	JNW	400
	No.7	BJS	JNW	463

In addition, we give the passenger demand of different origins and destinations on BJ railway line in Table 5 and the total number of passengers that are going to be transported by seven trains is 3619. The passenger data is obtained by the historical passenger flow of

one day on BJ railway line.

Table 5: Passenger volume between stations on BJ railway line

Volume	BJS	LF	TJS	CZW	DZE	JNW
BJS	-	801	211	596	241	1118
LF	-	-	141	81	10	68
TJS	-	-	-	92	44	76
CZW	-	-	-	-	13	98
DZE	-	-	-	-	-	29
JNW	-	-	-	-	-	-

4.2 Results of the experiments

In the set of experiments, we vary α_1 from 0.1 to 0.9, (by step of 0.1) to observe the set of optimal solutions. Here, we analyse the solution when $\alpha_1 = \alpha_2 = 0.5$ more in detail. We show train timetable, stopping plan and passenger assignment plan of the experiment result in Fig.3, Fig.4 and Table 6 respectively. In Fig.4, the solid dot "●" means that the train has to stop at this station for passengers getting on/off the train or for train preparing for its operation at its origin station.

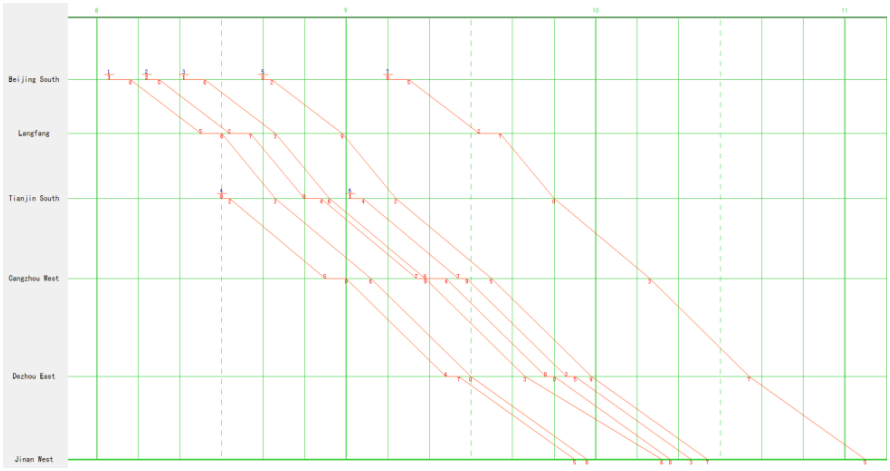


Fig.3: Train timetable for 7 trains on BJ railway line

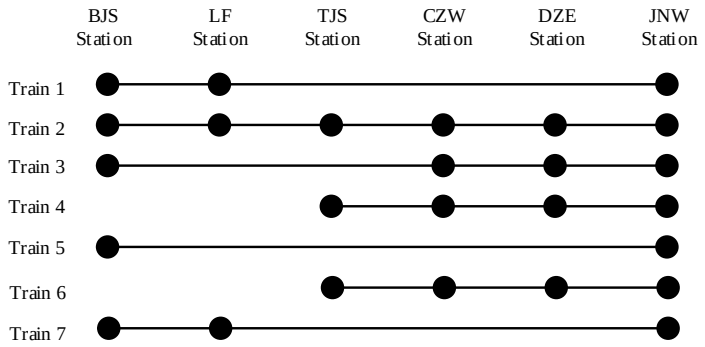


Fig.4: Train stop plan for 7 trains on BJ railway line

Table 6: Passenger assignment plan

Passenger pair	**Number of trains**							**Volume of passenger pair**
	1	2	3	4	5	6	7	
BJS- LF	642	124	0	0	0	0	45	773
BJS- TJS	0	211	0	0	0	0	0	211
BJS- CZW	0	61	289	0	0	0	0	350
BJS- DZE	0	0	241	0	0	0	0	241
BJS- JNW	0	0	0	0	540	0	510	1050
LF- TJS	0	87	0	0	0	0	0	87
LF- CZW	0	81	0	0	0	0	0	81
LF- DZE	0	10	0	0	0	0	0	10
LF- JNW	0	68	0	0	0	0	0	68
TJS- CZW	0	0	0	92	0	0	0	92
TJS- DZE	0	0	0	44	0	0	0	44
TJS- JNW	0	0	0	0	0	76	0	76
CZW-DZE	0	0	0	0	0	13	0	13
CZW-JNW	0	0	0	0	0	98	0	98
DZE - JNW	0	0	0	0	0	29	0	29

Fig.3 illustrates a train timetable of these 7 trains, in which we can obtain the information of train stop and dwell time at each station. Fig.4 details train stop plan for the tested trains on BJ railway line. Then, Table 6 represents the plan that transported passengers are assigned to the seven trains. It shows that when there are 7 trains in the railway network, total 3223 passengers have been delivered already, with 396 passengers not transported yet. To deliver these transported passengers, total travel time is 745 minutes, considering passenger demand.

5 Conclusion and future research

In this paper, we have put passenger demand into consideration when design a train timetable, and tackled the integration problem of train timetabling and train stop planning

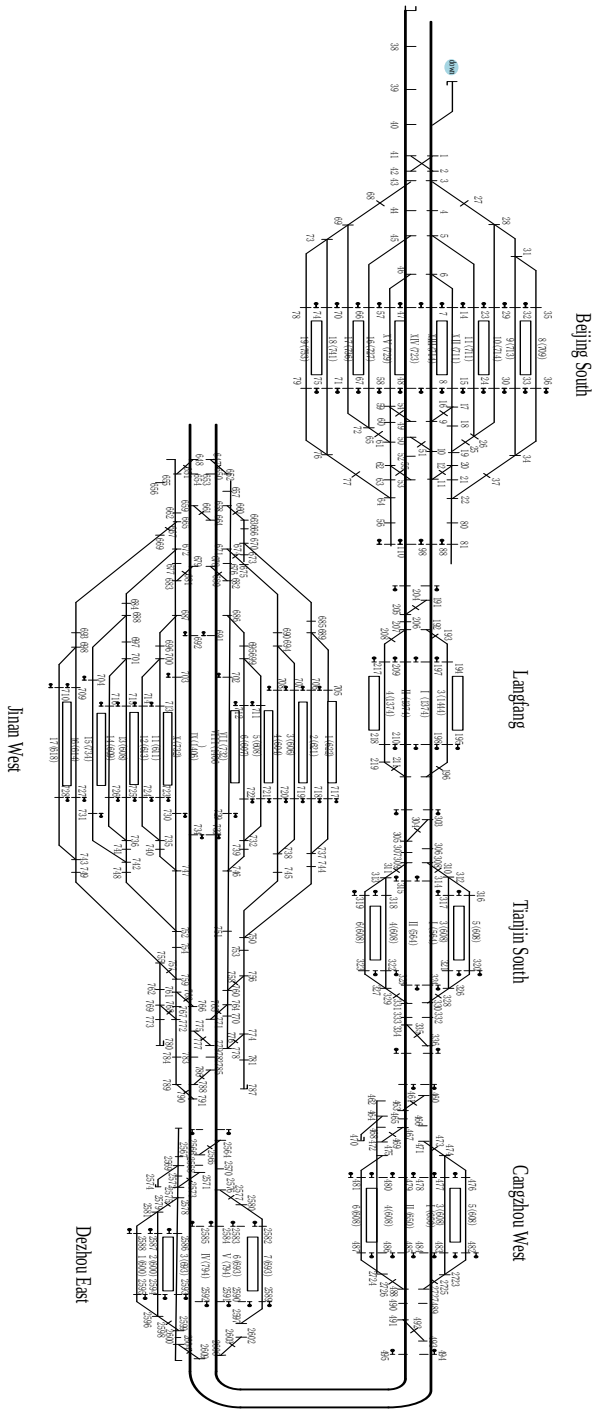
by using a bi-objective mixed integer linear programming model. Our aim is to compute train timetables (i.e. departure times and arrival times of all train at their stations), stop plan (including the choice of station that train stop and the dwell time at the station) and passenger assignment plan (including the resulting train that passenger get on it and the number of passengers that are carried). In this model, based on the origin station and destination station, we divide them into different passenger pairs in order to link the passenger pair with train stopping plans, and then generate train stop plans and timetable simultaneously. Furthermore, the weighted-sum method is used to find optimal solutions for the proposed bi-objective model. The validity of our model on solving this integration problem is shown by testing it on a part of Beijing-Shanghai high-speed railway line.

For future research, we will focus on the following main extension. Firstly, we will formulate the response of passenger behaviour to existing train service into our mathematical model to maximize the satisfaction of passengers. Train service operation is actually a mutual process. Next, a challenging extension is demand variation, as we know that passenger demand is elastic rather than fixed. Therefore, robust timetable is increasingly needed to adapt to the changing passenger demand (i.e. flexible dwell time in accordance to changing passenger demand). Finally, it is necessary to develop heuristic algorithm and dynamic programming method to improve the solution quality and computational efficiency for the real-time train scheduling problem, as passenger demand enhances the computational complexity.

Acknowledgements

The work of the first and second authors were jointly supported by a project from the National Key R&D Program of China (2018YFB1201500). The authors are of course responsible for all results and opinions expressed in this paper.

Appendix: BJ railway network in the test


Beijing South
Langfang
Tianjin South
Cangzhou West
Dezhou East
Jinan West

References

Szpigel.B. 1973. Optimal train scheduling on a single track railway. *Operation research*, 72, 333-351.

Higgins, A., Kozan, E., & Ferreira, L. 1996. Optimal scheduling of trains on a single line track. *Transportation research part B: Methodological*, 30(2), 147-161.

Lan, S. M. 2002. Study on the relevant issues of train running program along Beijing-Shanghai high speed line. *Railway Transport and Economy*, 24(5), 32-34.

Caprara, A., Fischetti, M., & Toth, P. 2002. Modeling and solving the train timetabling problem. *Operations research*, 50(5), 851-861.

Ghoseiri, K., Szidarovszky, F., & Asgharpour, M. J. 2004. A multi-objective train scheduling model and solution. *Transportation Research Part B, 38*(10), 927-952.

Caprara, A., Monaci, M., Toth, P., et al. 2006. A Lagrangian heuristic algorithm for a real-world train timetabling problem. *Discrete applied mathematics*, 154(5), 738-753.

Zhou, X., Zhong, M. 2007. Single-track train timetabling with guaranteed optimality: Branch-and-bound algorithms with enhanced lower bounds. *Transportation Research Part B: Methodological*, 41(3), 320-341.

Sun, L., Jin, J. G., Lee, D. H., Axhausen, K. W., & Erath, A. 2014. Demand-driven timetable design for metro services. *Transportation Research Part C: Emerging Technologies*, 46, 284-299.

Barrena, E., Canca, D., Coelho, L.C., Laporte, G., 2014a. Exact formulations and algorithm for the train timetabling problem with dynamic demand. *Computers and Operations Research*, 44: 66-74.

Barrena, E., Canca, D., Coelho, L.C., Laporte, G., 2014b. Single-line rail rapid transit timetabling under dynamic passenger demand. *Transportation Research Part B*, 70: 134-150.

Canca, D., Barrena, E., Algaba, E., Zarzo, A., 2014. Design and analysis of demand-adapted railway timetables. *Journal of Advanced Transportation*, 48: 119-137.

Meng, L., & Zhou, X. 2014. Simultaneous train rerouting and rescheduling on an N-track network: A model reformulation with network-based cumulative flow variables. *Transportation Research Part B: Methodological*, 67, 208-234.

Cheng, J., & Peng, Q. Y. 2014. Combined stop optimal schedule for urban rail transit with elastic demand. *Application Research of Computers*, 31(11), 3361-3364.

Niu, H., Zhou, X., & Gao, R. 2015. Train scheduling for minimizing passenger waiting time with time-dependent demand and skip-stop patterns: Nonlinear integer programming models with linear constraints. *Transportation Research Part B: Methodological*, 76, 117-135.

Wang, Y., Tang, T., Ning, B., van den Boom, T. J., & De Schutter, B. 2015. Passenger-demands-oriented train scheduling for an urban rail transit network. *Transportation Research Part C: Emerging Technologies*, 60, 1-23.

Fu, H., Nie, L., Meng, L., Sperry, B. R., & He, Z. 2015. A hierarchical line planning approach for a large-scale high speed rail network: The China case. *Transportation Research Part A: Policy and Practice*, 75, 61-83.

Burdett, R. L. 2015. Multi-objective models and techniques for analysing the absolute capacity of railway networks. *European Journal of Operational Research*, 245(2), 489-505.

Yang, L., Qi, J., Li, S., & Gao, Y. 2016. Collaborative optimization for train scheduling and train stop planning on high-speed railways. *Omega*, 64, 57-76.

Yue, Y., Wang, S., Zhou, L., Tong, L., & Saat, M. R. 2016. Optimizing train stopping patterns and schedules for high-speed passenger rail corridors. *Transportation Research Part C: Emerging Technologies*, 63, 126-146.

D'Ariano,A., Meng, L., Centulio, G., & Corman, F.2017. Integrated stochastic optimization approaches for tactical scheduling of trains and railway infrastructure maintenance. *Computers & Industrial Engineering* https://doi.org/10.1016/j.cie.2017.12.010.

Yang, X., Chen, A., Ning, B., & Tang, T. 2017b. Bi-objective programming approach for solving the metro timetable optimization problem with dwell time uncertainty. *Transportation Research Part E: Logistics and Transportation Review* 97:22-37

Robenek, T., Azadeh, S. S., Maknoon, Y., de Lapparent, M., & Bierlaire, M. 2018. Train timetable design under elastic passenger demand. *Transportation Research Part B: Methodological*, 111, 19-38.

Wang, Y., D'Ariano, A., Yin, J., Meng, L., Tang, T., & Ning, B. 2018. Passenger demand oriented train scheduling and rolling stock circulation planning for an urban rail transit line. *Transportation Research Part B: Methodological*, 118, 193-227.

Qi, J., Cacchiani, V., & Yang, L. 2018. Robust Train Timetabling and Stop Planning with Uncertain Passenger Demand. *Electronic Notes in Discrete Mathematics*, 69, 213-220.

Machine Learning based integrated pedestrian facilities planning and staff assignment problem in transfer stations

Bisheng He [a,b], Hongxiang Zhang [a], Keyu Wen [c,d], Gongyuan Lu [a,b,1]
[a] School of Transportation and Logistics, Southwest Jiaotong University P.O. Box 610031, Chengdu, China
[b] National United Engineering Laboratory of Integrated and Intelligent Transportation P.O. Box 610031, Chengdu, China
[c] China Railway Economic and Planning Research Institute P.O. Box 100038, Beijing, China
[d] School of Economics and Management, Southwest Jiaotong University P.O. Box 610031, Chengdu, China
[1] lugongyuan@swjtu.edu.cn, Phone: +86 (0) 138 8060 9100

Abstract
Optimizing the pedestrian facilities plan in transfer stations is the problem of adjusting the facilities on the layout of pedestrian flow route and the number of machines in service to service to meet the level of services requirements. In the practice, the operation of pedestrian facilities plan is always associated with the staff assignment. Hence, we develop a machine learning based integrated pedestrian facilities planning and staff assignment optimization model in transfer stations to schedule the pedestrian facilities plan and the staff assignment together. It aims to minimize the staff assignment cost and the deviation of working time of each employee of the station. The minimizing of the deviation gains the fairness of the assignment plan. The facilities plan is enforced by the level-of-services requirement in three performance indicators including transfer capacity, transfer average time and level-of-service. The performance indicators of facilities plans are evaluated by a simulation-based machine learning method. Based on simulation results, the random forest method fits a quantitative relationship among performance indicators of the facilities plans with operation scenario attributes and facilities plan attributes. The experiments on the case study of Xipu station show the integrated model can return pedestrian facilities plans which meet the level of service requirements and assign employees fairly of each period in a day and minimize the labor cost. The solutions of pedestrian facilities plan and staff assignment plan for possible operation scenarios in future are also suggested to station manager by our integrated method.

Keywords
Transfer stations, Facilities plan, Staff assignment, Simulation, Random forest

1 Introduction

Pedestrian Facilities Planning (PFP) is about adjusting the layout of pedestrian flow route and the number of machines in service to meet the passenger movement demand in the station. Given the increasing passenger demand in transfer station in China, the station

management face high pressure. Especially, the passenger demand varies from hour to hour in a day. A fixed pedestrian facilities plan usually fails to satisfy passengers' transfer demand, resulting in the bottleneck of the rail transit network. The adjustment of pedestrian facilities in different time periods in one day are applied by station managers' experience currently. Hence, the optimization of PFP is an urgent concern(Hu et al. (2015)).

Staff assignment (SA) of the employees in the station is the most important part of station management. Considering the situation in China, lots of passengers are unfamiliar with the automatic ticket machine and automatic gate for ticket checking, and they are unable to use them correctly and quickly. Some passengers may waste time on walking through passages without conductor's guidance in the station because of low sensitivity to the guide signs. And it is necessary to maintain order by staff if congest happens. Therefore, the staff are assigned to the facilities to assist passengers in passing correctly and quickly, guaranteeing the efficient operation of station. In practice, the employees are associated with pedestrian facilities plan. For example, one employee can handle four gates at the same time, if the pedestrian facilities plan with five gates in service is chosen with the increasement of passenger demand, two employees would be assigned. Hence, the staff assignment plan should be modified to correspond to the adjustment pedestrian facilities plans in time periods.

In all, PFP and SA must be managed simultaneously in the daily operation to meet the high-density transfer stations. There are three challenges in the management of station when applying the integration of PFP and SA into practice.

(1) The performance the pedestrian facilities plan is hard to quantify. A lot of researches are focused on the evaluation of pedestrian flow performance and passenger assignment in the station. The pedestrian route choice model which is developed from the route choice model in road transportation is a widely used to evaluate the pedestrian behavior(Lam et al. (1999) and Hänseler (2016)). However, the pedestrian routes in transfer station are different from the routes in road transportation, they are always overlapped. It is hard for pedestrians to find the optimal paths in most time. Then, the accuracy of pedestrian route choice model need to be improved by other measures. Berbey et al. (2012) and Xu et al. (2013) addressed a probabilistic model and a fuzzy logic approach to modeling passenger behavior on the platform. In other researches, the route choice data collected by Bluetooth and WiFi technologies are investigated. Shlayan et al. (2016) used Bluetooth and WiFi technologies to obtain origin-destination(OD) demands and pedestrian movement path in public transportation terminals. Based on the route choice and waiting time data collected by Bluetooth on two platforms, Heuvel et al. (2015) estimated the impact of vertical infrastructure like escalators and stairs on choosing pedestrian flow routes in train station. Through Bluetooth, WiFi and infrared technology, Heuvel et al. (2016) expanded the research to include station hall and non-train passengers. However, collecting data of pedestrian flow route is a difficult and time-consuming process, and the exact pedestrian flow route are hardly acquired by these technologies due to the date collecting devices only located in specific points. Hence, these technologies need to be improved to acquire more reliable data in a efficient way. Pedestrian simulation has been recognized as a powerful tool under limited data collection and has been implemented to return the performance indicators in the station for the strong computation capability (Hoy et al. (2016), Pu (2017)). However, the pedestrian facilities are fixed as the given conditions for the analysis. Some research focuses on some specific facilities in the station. Hu et al. (2015) addressed a width design of the urban rail

transit stations circulation facilities problem by a simulation-based optimization. Only the width of circulation facilities was analysand. The relationship between the whole pedestrian facilities plan and the performance of the pedestrian is not analyzed. The machine learning method has been used in railway operation with the capacity assessment (Lai et al. (2014)), train delay estimation(Kecman and Goverde (2015)),and track maintenance(Ghofrani et al. (2018)). The machine learning could be a promising way to provide the nonlinear quantitative relationship(Ghofrani et al. (2018)).

(2) The optimization operation of pedestrian facilities is interrelated with the train timetable. The development of comprehensive transportation brings the multi-mode transfer into one station, and leads to more transfer connections, designed by the transportation company. A lot of researchers focus on the optimization and the train service in the station(Wong et al. (2008), Ibarra-Rojas and Rios-Solis (2012), Dollevoet et al. (2014) and Corman et al. (2017)), but the performance indicators about level of services are ignored in their researches. D'Acierno et at. (2017) combined an analytical dwell time model and dthe railway simulation under the crowding level at platforms to support timetabling development of metro, which aims to guarantee an appropriate robustness of rail operation. Minimizing the transfer time or travelling time are always the objective function in their models. Given the tightened transfer time, the station would be into a congested situation with high density of passengers in different time periods, which may lead to negative response and interactions among pedestrians(Bandini et al. (2014) and Pel et al. (2014)). Hence, the cost and difficulty of station management may increase because the passengers are more likely to miss their trains or services and need a longer waiting time for the following services(Tirachini et al. (2013)).

(3) The staff associated with the pedestrian facilities plan should assigned with the practical constraints. Although the staff assignment is a classic assignment problem(Ernst et al. (2004)), similar to the crew scheduling problem(Huisman (2007)), the practical requirements may make the problem more the number of employees is limited, and the labor constraints, like the maximum working time, the fairness of the work plan should be considered.

We develop a machine learning based integrated pedestrian facilities plan and staff assignment method in transfer stations. The purpose of this paper is to use the facilities plan and to assign the staff of the station dynamically to ensure the adaptability of the station to different passenger demands and to minimize the labor cost in one-day operation. We need to state that our research focuses on the transfer between the metro system and railway system in China. This paper contributes to face the three above methodological issues on the integration of PFP and SA.

For point 1, a machine learning method by fitting the train data returned by an agent-based simulation model is developed to evaluate the performance of the transfer management. We use AnyLogic software which is based on the social force model to simulate the passengers behavior in station. Instead of delimiting the pedestrian routes by researchers, origin-destination(OD) is the motive force for pedestrian and the routes are formed automatically. The interaction among pedestrian also be considered in simulation model because of social force model. Then, the simulation model output the performance indicators, including the transfer capacity, the average transfer time and level-of-service of the space in the station by input the train operation attributes and passenger attributes. Then, as the simula-

tion based on the social force model, the simulation output data easily result in a randomness. To acquire a reliable performance indicators, a machine learning, random forest in the paper, fits the relationship between the pedestrian facilities plan and transfer management performance indicators.

For the point 2 and point 3, we develop a mathematical model with the considering of the level of services requirement in each time period. Based on the quantitative performance indicators acquired by the random forest, these requirements which is proposed by the station management could select the the facilities plans to face the challenge of the tightened transfer time. Then the mathematical model can allocate the selected pedestrian facilities plan and assign the station staff, which aims to minimize the staff assignment cost which is the main operation cost of station. The minimizing of the deviation gains the fairness of the assignment plan. Then, the constraints of limitation on the total working time per day, the working time range and fairness of workload are considered.

The next sections of the paper are organized as follows. Section 2 describes the pedestrian facilities planning problem, staff assignment problem and the integration problem. Section 3 presents the mathematical model proposed in this paper for the integration of PFPP and SAP problems, and explains the facilities plan performance quantization method by the random forest and simulation. Section 4 shows the experiments results by the simulation with random forest, and provides computational results on the proposed methods. Section 5 summarizes our contributions to the literature and outlines directions for further research on the integrated problem.

2 Problem Description

This section introduces the pedestrian facilities planning problem and staff assignment problem and how to integrate them.

2.1 Pedestrian facilities planning problem

Pedestrian facilities plan contains pedestrian flow route layout and number of machines in service. It can be adjusted to satisfy different passenger demand.

Two main elements of pedestrian facilities plan
We introduce the two main elements of pedestrian facilities plan.

1)Layout of pedestrian flow route

Pedestrian flow route is the path for passengers walking in the transfer station. By arranging the equipment like railings and barriers, the layout of pedestrian flow route is designed properly to make sure that there is no cross interference or convection among each other. Otherwise, a poor layout of pedestrian flow route usually leads to collisions, conflicts and dwell on the platform occur with passenger movement, which will increase the dwell time in the stations. Hence, an appropriate layout of pedestrian flow route should be selected to meet the operational requirements.

2)Number of machines in service

Except for the layout of pedestrian flow route, the number of machines in service in pedestrian facilities plans plays an essential role in the movement of transfer passenger. Concerning the queuing theory, the number of machines in service could result in the service and pass time in the machine and the neighbor area. As the transfer between the metro

system and railway system in China is very complicated, the machines in the transfer station include automatic ticket machines, automatic gates for ticket checking, etc. The passing of each machine will influence each other. Therefore, the station manager can change the number of machines in service by turning off idle machines or turning on more machines to guarantee the level of services requirement.

Evaluation of pedestrian facilities plan

The pedestrian facilities plan is selected by the station manager by its performance under the certain operation scenario. Based on the analysis in (Hänseler (2016)), we propose three indicators to evaluate the performance of pedestrian facilities plan for the transfer station: 1)Transfer capacity. It is the number of passengers transferring from the one system to another in the given time period. These indicators could be used in the oversaturate situation; 2)The average transfer time. It is the average of the time from arrival train to departure train belong to another system for each transfer passenger. It could obtain the operation of a station to meet the tighten timetable requirement; 3) level-of-service. It refers to the pedestrian density of space for each passenger, and usually measure in six levels from A to F. When level-of-service of space factor is low, the train stop time may not be sufficient for passengers to get on or get off the train, and safety accidents may also happen. The station manager can determine to choose which pedestrian facilities plan to improve the level of station management by the performance of these indicators.

Adjustment of pedestrian facilities plan

The adjustment of pedestrian facilities plan can improve the level of station management and save operation cost. Different pedestrian facilities plans are applicable for different operation scenarios.

The operation scenario contains two parts. The first part is the attributes of the passenger. Passenger demand, which affects the number and density of passengers directly, is the main factor in the operation scenario. Passenger characteristics, including whether to carry luggage, whether to buy tickets, etc., may affect the time of passengers receiving all kinds of service. The second part is train operation headway. Train operation headway, which means the change of the train frequency for multi-mode station in the operation scenario. It can affect the arrival and departure density of passengers in the station.

As the operation scenario changes in every time period in the daily operation, the performance of pedestrian facilities plan varies with the change of the three evaluation indicators. On the other side, different operation scenario will set the delicated operation requirements for operation. Corresponding to the performance indicators, the station operation provides three requirements, the minimum capacity, the maximum average transfer time and minimum level-of-service. If the performance of the three evaluation indicators shows that the current pedestrian facilities plan cannot meet the level of services requirement in current operation scenario, the pedestrian facilities plan must be adjusted. Considering the using cost and service life of facilities, under the premise of guaranteeing the level of services requirement, the pedestrian facilities plan with fewer machines in service can be chosen in low demand period. What's more, fewer machines in service means that fewer conductors are required, which can reduce the cost for the station. And it must be stated that changing pedestrian flow route between different pedestrian facilities plans, which can be completed in one minute, cost so little that it can be ignored because of using movable and retractable railings.

2.2 Staff assignments problem

Staff assignment plan is about how many employees should be assigned to and where to assign as well as how long will they work. According to the performance of evaluation indicators in different operation scenarios, the pedestrian facilities plan can be determined. Then, it is necessary to develop an appropriate staff assignment plan for pedestrian facilities plan and to minimize the labor cost at the same time, because labor cost is the main expense in the operation of rail transit stations in China. The pedestrian facilities plan determines the number of employees required. The number of machines in service is different when pedestrian facilities plan changes along with a different period, which leads to a different requirement of employees in one day because the number of machines that one employee can handle is limited.

There are two kinds of employees in the staff of rail transit station in China: regular employees and secondment employees. The beginning and the end of regular employees' working time are fixed. The secondment employees can be assigned to anywhere at any time if the requirement of employees is more than regular employees. However, the cost of assigning one secondment employee is three times more than the regular employees. Besides, the limit of employees' fatigue, including the total working time of each employee, working time range and fairness of workload should be considered. Normally, the total working time of each employee should be less than 8 hours, and at least 1-hour rest is required.

2.3 Integrated pedestrian facilities planning and staff assignment problem

As mentioned above, the simulation model and random forest are aim to obtain performance indicators for different operation scenarios, but the categoric input and output of them are not determined. Therefore, an integrated model is developed to combine the pedestrian facilities plan and staff assignment plan, and it can define the input and output clearly of simulation model and random forest. The purpose of this model is to reduce the operation cost of the station, especially the staff cost, and to improve the transfer passenger satisfaction by according to the evaluation indicators mentioned above.

We firstly develop an agent-based simulation model in Anylogic software to obtain all the evaluation indicators performances of all possible combinations between operation scenarios and pedestrian facilities plans. The simulation model can provide plenty of indicators performances results of each combination to prevent extreme situation of results that affect the final result. However, due to the randomness and large quantities of the simulation results and the nonlinear relationship between operation scenario and pedestrian facilities plan, it is very difficult for station manager to judge which simulation results can be used and choose appropriate pedestrian facilities plan for certain operation scenario. Therefore, we fit the nonlinear relationship between them by using Random Forest, and we obtain the correspondence between the performance of evaluation indicators and the combinations of operation scenarios and pedestrian facilities plans. Then, we use given operation scenario for the operation day to select out all the pedestrian facilities plans which satisfy the performance requirement provided by the station manager. Finally, the number of employees required for each pedestrian facilities plan and the cost of hiring an employee are provided. We use a mathematical model which can assign staff to minimize the operation cost and determine the integral pedestrian facilities plan from all selected out plans to satisfy the per-

formance requirement simultaneously for the station. The whole solution process is shown in Figure 1.

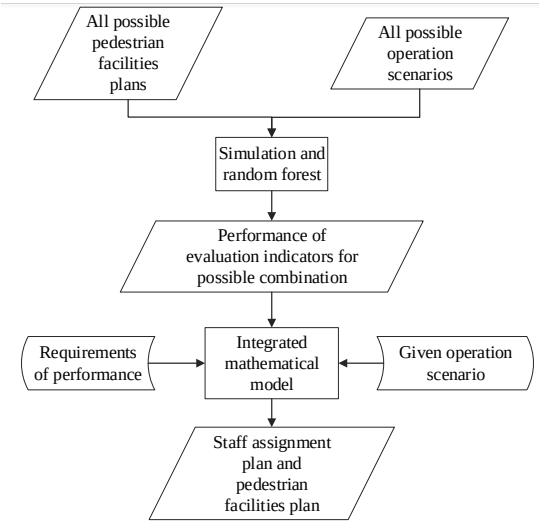


Figure 1: Flow chart of the our method.

3 Model Formulation

In this section, we address our Integrated Pedestrian Facilities Planning and Staff Assignment Problem (IPFPSA) in the transfer station. Firstly, the notation is presented. Then, we will give the mathematic model in detail. Next, the random forest connected the pedestrian facilities plan with the performance indicators in the station in IPFPSA will be introduced.

3.1 Notation

The sets, parameters and decision variables used in this paper are described in Tables 1 and 2 , respectively.

3.2 Modeling assumptions

The Modeling assumptions in our research is list as follows:

(1)Since the operation and adjustment of facilities plan cost are not considered in the daily operation, we do not minimize them in our model.

(2) A facilities plan utilized in station must be associated with a number of employees.

(3)A number of secondment employees could apply to the facilities plan when necessary. The labor cost of secondment employees is much higher than the regular employees.

(4)The number of transfer passenger is given. The station manager could acquire the passenger demand by other prediction methods.

Table 1: Definition of sets and parameters

Symbol	Definition
S	Set of staff, index by s , i.e., $s \in S$
S_{se}	Set of secondment staff in some time periods, index by s , i.e., $s \in S_{se}$
S_{re}	Set of regular staff, index by s , i.e., $s \in S_{re}$
F	Set of pedestrian facilities plan, index by f , i.e., $f \in F$
T	Set of time periods, index by t , i.e., $t \in T$
O_s	The limitation of working time for staff s
$t_{s,b}, t_{s,e}$	The begin time period and end period for staff s
$L_{f,t}$	The level-of-service indicator when facility plan f performance in time period t
$\tau_{f,t}$	The maximum average transfer time indicator when facility plan f performance in time period t
$Cap_{f,t}$	The minimum transfer capacity indicator when facility plan facility plan f performance in time period t
$L_{min,t}$	The minimum level-of-service requirement in time period t
$\tau_{avg,t}$	The transfer time requirement in time period t
$Cap_{min,t}$	The transfer capacity requirement in time period t
N_f	The needed number of staff when using facility plan f

Table 2: Decision variables

Symbol	Definition
$x_{s,f,t}$	0-1 staff assigning variable, equal to 1 if employee s assigned to facilities plan f in time period t, 0 otherwise
$y_{f,t}$	0-1 facilities planning variable, equal to 1 if facilities plan f selected in time period t to meet the station operation requirements, 0 otherwise
z_s	The actual working time for employee s
w	The average working time for all the staff $s \in S$
D_s^+, D_s^-	Deviate time to the average working time for all the staff

3.3 Mathematical Model

The aim of pedestrian facilities planning is to meet the station operation requirements, while the staff assigning focuses on operating the facilities plan in a low cost. Moreover, fairness is included to improve the satisfaction of station staff. Moreover, our constraints include four parts: 1) the station operation requirements for the facilities plan; 2) the staff assignment rules; 3) the computing of the deviation of the working time for each employee to the total staff average; 4) the relationship of facilities plan and staff assignment. IPFPSA is formulated as follows:

$$min \sum_{t \in T} \sum_{f \in f} \sum_{s \in S} c(x_{s,f,t}) x_{s,f,t} + \sum_{s \in S} D_s^+ + \sum_{s \in S} D_h^-. \tag{1}$$

$$c(x_{s,f,t}) = \begin{cases} c_{s,t,re}, s \in S_{re} \\ c_{s,t,se}, s \in S_{se} \end{cases} \tag{2}$$

$$\sum_{f \in F} y_{f,t} = 1 \, \forall t \in T. \tag{3}$$

$$\sum_{f \in F} L_{f,t} y_{f,t} \geq L_{min,t} \quad \forall t \in T \tag{4}$$

$$\sum_{f \in F} Cap_{f,t} y_{f,t} \geq Cap_{min,t} \quad \forall t \in T. \tag{5}$$

$$\sum_{f \in F} \tau_{f,t} y_{f,t} \leq \tau_{avg,t} \quad \forall t \in T. \tag{6}$$

$$\sum_{f \in F} x_{s,f,t} \leq 1 \quad \forall s \in S, t \in \{T | t_{s,b} \leq t \leq t_{s,e}\}. \tag{7}$$

$$\sum_{f \in F} x_{s,f,t} = 0 \quad \forall s \in S, t \in \{T | t \leq t_{s,b}, t \geq t_{s,e}\}. \tag{8}$$

$$\sum_{t \in T} \sum_{f \in F} x_{s,f,t} \leq O_s \quad \forall f \in F, t \in T. \tag{9}$$

$$z_s = \sum_{s \in S} x_{s,f,t} \quad \forall s \in S. \tag{10}$$

$$W = \sum_{s \in S} z_s \setminus n_s. \tag{11}$$

$$z_s = W + D_s^+ - D_s^- \quad \forall s \in S. \tag{12}$$

$$x_{s,t} \leq M y_{f,t} \quad \forall t \in T \tag{13}$$

$$\sum_{s \in S} x_{s,f,t} = N_f y_{f,t} \quad \forall f \in H, t \in T. \tag{14}$$

$$x_{s,f,t} \in 0, 1 \quad \forall s \in S, f \in F, t \in T. \tag{15}$$

$$y_{f,t} \in 0, 1 \quad \forall f \in F, t \in T. \tag{16}$$

$$z_s \in \mathbb{Z} \quad \forall s \in S. \tag{17}$$

$$w \in \mathbb{R} \tag{18}$$

$$D_s^+, D_s^- \in \mathbb{R} \quad \forall s \in S. \tag{19}$$

The objective function (1) minimizes the total staff assigning cost. To obtain the fairness of the staff, the deviations of the average working time are computed. Constraints (2) deal with the situation that the secondment employees working cost is different from the regular employees. Indeed, it is much higher, which is computed by the practical operation

experience. Constraints (2) enforce that only one facility plan s is selected in each time period t. Constraints (4), (5), and (6) ensure that in each time period t , the selected facilities plan s must meet the service level requirement $L_{min,t}$, the transfer capacity requirement $Cap_{min,t}$. and the transfer average time requirement $\tau_{avg,t}$, respectively. Constraints (7) and (8) address that each employee s only assigned to one facility plan s in their working time period t, while employee s could not work outside the working time period t. For an employee s, Constraint (9) is defined to ensure that the total working time in one day is less than O_s. The big-M in constraints (13) are used to couple the usage of an employee s with facilities plan f. Constraints (13) imply that the usage of an employee s in time period t will be enforced to be 0 if $y_{f,t}$ is equal to 0, i.e. employee s could not assign to a facilities plan s which does not implement; otherwise, $x_{s,f,t}$ is less than or equal to the value of big-M, i.e. it could be used. Constraints (14) specify that if $y_{r,t}$ is equal to 1, N_f employees are applied to the facilities plan r to operate it. The domain of variables in the model is defined by expressions (16)- (19). The staff assigning and facilities planning are defined as binary variables. The actual working time, the average working time and the deviation time to the average working time are defined as integer variables. As we only focus on one-day's planning, the IPFPSA could be solved by some commercial software. The essential part for the model is how to get the performance of the facilities plan in various scenarios. We will introduce quantization method next.

3.4 Facilities plan performance quantization method

As we stated before, the process of transfer between two modes leads to a nonlinear relationship between the facilities plan and the performance. Regression methods based on machine learning are the common ways to model that. This kind of method has been applied in railway system to predict the railway capacity, train delay, etc. For the machine learning, how to get the learning data is the most challenge task. We develop a simulation system to provide a number of results which connect the performance indicators and input data.

Simulation for transfer station
We choose Anylogic to build the simulation model of passenger transfer progress. Anylogic combines professional discrete event, system dynamics, and agent-based modeling in one platform(Anylogic (2019)). In the software, Rail Library is used to build the train operation simulation model, in which each train are agents with their own states and properties. Pedestrian Library, based on the social force model, is used to build a pedestrian moving simulation model. Pedestrians can be preassigned with individual characteristics in models. In order to integrate train and pedestrians, this paper builds a transfer simulation model, which can simulate the train operation in the station and the complete progress of transfer in one model. Therefore, the change of train operation headway and its influence on operation scenario can be presented in the simulation model. The operation scenario attributes, pedestrian facilities plan attributes and performance indicators are listed in Table 3. Facilities plan attributes have been introduced in Section 2, we will introduce simulation attributes and performance indicators in detail.

Operation scenario attributes
In multi-mode transfer station connecting railway and metro, both of the train operation headway of railway and metro affect the operation scenarios. Whether to buy tickets or not

Table 3: The input and output of the simulation model

Operation scenario attributes	Facilities plan attributes	Facilities Performance indicators
Headway of railway(HR) $H_{r,t}$,	Layout of pedestrian flow route(LPFR)R_y,	
Headway of metro(HM) $H_{m,t}$,	Automatic gate for railway tickets checking(AGRTC)$N_{rtc,y}$,	Transfer capacity $Cap_{y,t}$,
Metro inbound demand(MID) $D_{in,t}$,	Automatic ticket machines for metro(ATMM)$N_{m,y}$,	Average transfer time $\tau_{y,t}$,
Transfer demand(TD) $D_{tr,t}$,	Automatic gate for metro tickets checking(AGMTC)$N_{mtc,y}$	Service level $L_{y,t}$
Ratio of buying tickets(RBT) RA_t		

is the main passenger characteristic in this transfer progress. Passenger demands include transfer demand and inbound demand. Therefore, inbound demand from the metro, transfer demand of each arrival train, the ratio of buying tickets in transfer passengers, train operation headway of railway and train operation headway of metro are the five variables of operation scenarios.

Facilities plan attributes
As stated before, the layout of pedestrian flow route and number of machines in service are the facilities plan attributes. In multi-mode transfer station, the machines are specified in automatic gate for railway tickets checking, automatic ticket machines for metro and automatic gate for metro tickets checking.

Facilities performance indicators
The performance of evaluation indicators is returned as the results of the simulation model under the given operation scenario and pedestrian facilities plan. The method to obtain the performance of evaluation indicators shown as follows.

1 Transfer capacity

Transfer capacity is the number of passengers given by a simulation method within the time period. In the simulation model, the unit time is set to 10 minutes, so this indicator can calculate the number of passengers transferring successfully in 10 minutes after the train arrived.

2 Transfer time

The average transfer time is chosen to be the performance indicator to evaluate all transfer passengers' transfer time. In multi-mode transfer station, the progress of transfer is divided into 4 parts as shown in expression (20)and the average transfer time τ_{avg} can also be computed.

$$\tau_{avg} = \sum_{i \in N} (\tau_{i,walk1}, \tau_{i,serv} + \tau_{i,walk2} + \tau_{i,wait}) \setminus |N|. \tag{20}$$

Where $\tau_{i,walk1}$ is the time of getting off and walking of ith transfer passenger, $\tau_{i,serv}$ is the time of buying tickets and passing automatic gate for ticket checking of of ith transfer passenger, $\tau_{i,walk2}$ is the walking time after checking ticket of ith transfer passenger and $\tau_{i,wait}$ is the time of waiting to get on the train to transfer of ith transfer passenger.

3 Level-of-service

Considering China's rail transit transfer station, we choose the A-F level-of-service grading standard in (HCM,Highway Capacity Manual (2000)). This paper counts the percentage of time in the A-F level-of-service of the station in per hour, and normalizes the six values to obtain a value of (0,1), which means the level-of-service is better if the value is larger.

Each operation scenario and each pedestrian facilities plan can be combined with a possible input of simulation model, which leads to plenty of results of evaluation indicators. Due to the randomness of the simulation and the nonlinear relationship between parameters and results, the results need to be processed further by machine learning.

Random forest for quantization method

Random forest ensembles a set of decision tree prediction results to gain better predictive results for both classification and regression problems. It is an efficient machine learning model which was used widely for many real-world applications(Shafique and Hato (2017), Kecman and Goverde (2015)). Moreover, the random forest could provide the importance scores of input attributes which is useful for the analysis of the integration model.

The random forest needs lots of learning scenarios before fitting it. Hence, the simulation method is applied to provide a number of scenarios. We choose the random forest to connect the simulation attributes, pedestrian facilities plan attributes and performance indicators in Table 3 with expression (21).

$$\begin{cases} Cap_{y,t} = f_{cap}(H_{r,t}, H_{m,t}, D_{in,t}, D_{tr,t}, RA_t, R_y, N_{rtc,y}, N_{m,y}, N_{mtc,y}) \\ \tau_{y,t} = f_{\tau}(H_{r,t}, H_{m,t}, D_{in,t}, D_{tr,t}, RA_t, R_y, N_{rtc,y}, N_{m,y}, N_{mtc,y}) \\ L_{y,t} = f_{L}(H_{r,t}, H_{m,t}, D_{in,t}, D_{tr,t}, RA_t, R_y, N_{rtc,y}, N_{m,y}, N_{mtc,y}) \end{cases} \tag{21}$$

After the evolution of performance indicators, the commercial software, like the gurobi or IBM Cplex, could be used to solve IPFPSA, to get the final solution.

4 Numerical experiments

This section present the numerical experiment and solution analysis. The case study of Xipu station is used to model the transfer management. The experiments on the models proposed in this paper have been performed on a laptop computer with i7-6700HQ @ 2.6 GHz CPU and 8.0 GB RAM. The IPFPSA is solved with Gurobi 7.5, the simulation is developed with Anylogic 8.3 , the random forest is trained on Python 3.6 with scikitlearn package on Windows 10.

4.1 The case study of Xipu transfer station

The Xipu station is an appropriate station for the case study as shown in Figure 2. It is a multi-mode station as the intermediate station of the intercity railway system and terminal station of the metro system. Passengers can transfer between railway and metro on the same platform, where the platform 1 provides the transfer from the railway to metro and platform 2 provides the transfer system from metro to railway. Moreover, different ticket systems of railway and metro means more complicated procedure and longer walking distance for transfer passengers. When passengers transferring from metro to railway, they have to wait for check-in to board, which can relieve the pressure of transfer management. Therefore, we

focus on the transfer from the railway to the metro on platform 1 with complicated facilities requiring more conductors of the station staff, and more continuous transfer progress.

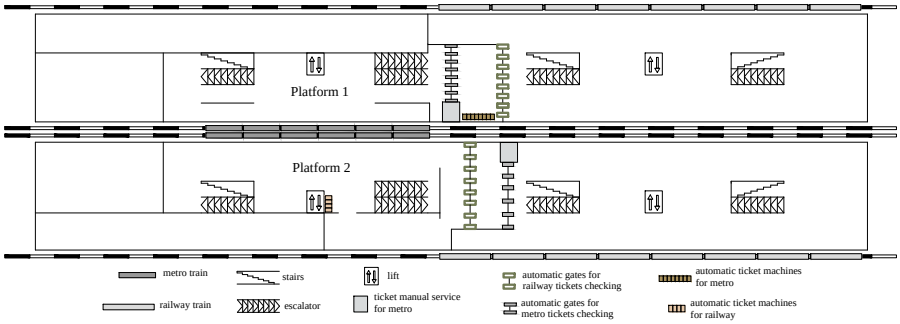


Figure 2: The facilities layout on the platforms in Xipu station

In a limited area on platform 1 as shown in Figure 3, after getting off the arrival train, transfer passengers must complete the transit through checking-out of railway system, ticket purchasing (if not have) and checking-in of metro system by 7 automatic gates for railway tickets checking, 7 automatic gates for metro tickets checking, 8 automatic ticket machines and 1 manual ticket service for metro. With so many machines available for service, the number of chosen machines for service varies in a wide range, increasing the complicate for developing pedestrian facilities plan and staff assignment plan. possible layout of pedestrian flow routes aiming to satisfy different passenger demands can be chosen after checking in the metro system as shown in Figure 3. Since most facilities belong to the metro system, the pedestrian facilities plan and staff assignment plan are mainly designed for metro station management.

4.2 Simulation experiments

simulation model

The transfer simulation model shown in Figure 4 is built in Anylogic and an display of simulation results is shown in Figure 5. The simulation model consists of train simulation and pedestrian simulation:

(1)Train operation: The three blocks of *Delay* control the stop of train and passengers getting on and off, which integrate the train operation and pedestrian.

(2)Passengers getting off and walking: The block *judge1* controls whether the passenger transfer or not.

(3)Passengers checking tickets: The block *judge2* controls whether the passenger needs to buy metro tickets or not. The number of machines in service is the parameter of block *CheckOutCRH*, *BuyTicket* and *CheckInMetro*.

(4)Passengers walking and waiting to get on: The block *Wait*, *GetOn* and *judge3* control the part of waiting to get on the metro and receiving the guidance of conductors to simulate the reality in China.

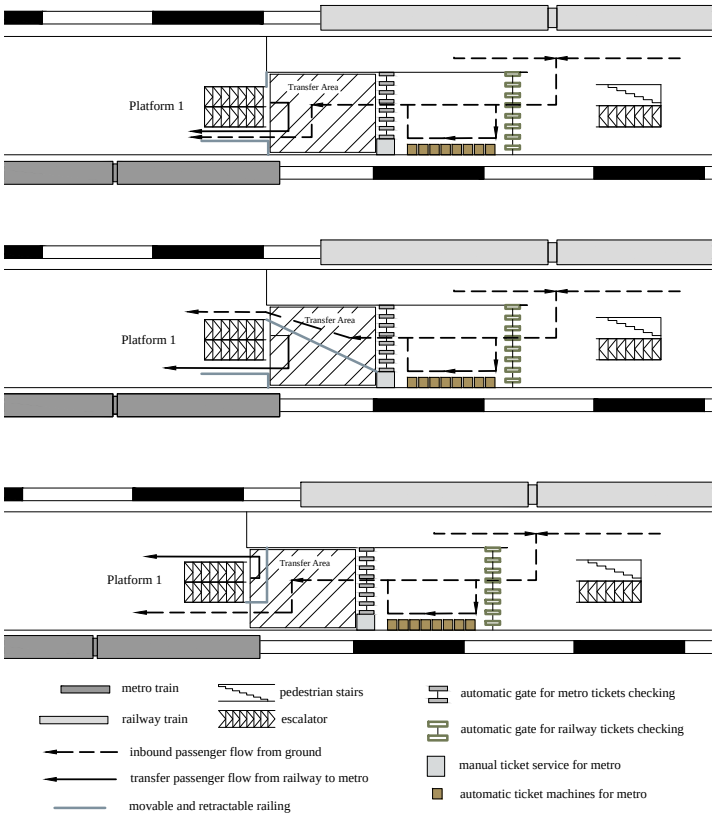


Figure 3: The facilities, three possible layout of pedestrian flow route and progress of transfer on platform 1

(5)Inbound passengers: The demand for inbound passengers also affects operation scenarios.

Parameter settings and experiment

A large number of simulation experiments are performed to obtain the performance of evaluation indicators for all possible operation scenarios with all possible pedestrian facilities plans. As shown in Figure 3, three types of layout of pedestrian flow route are designed. Then, 13 different numbers of machines in service and with its requirement of employees are designed as shown in Table 4. Since the manual ticket service for metro is fixed at 1, it will not be considered in requirement of employees for the facilities plans .

Due to 3 types of layout of pedestrian flow route and 13 different numbers of machines in service, there are 39 pedestrian facilities plans enumerated.

For operation scenarios, we developed over 400 possible operation scenarios in total by changing the attributes for diverse operation scenarios. Each operation scenario will con-

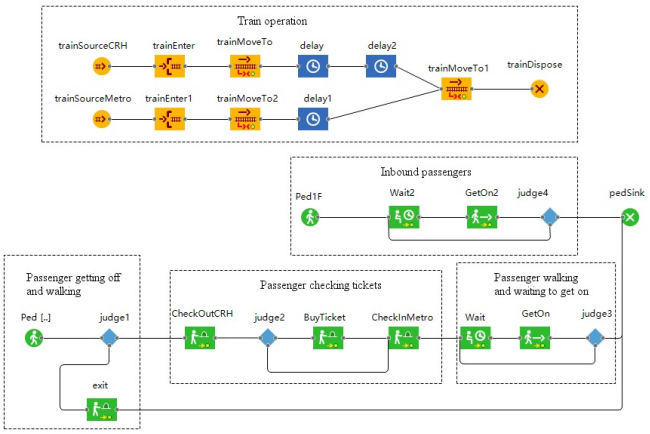


Figure 4: The transfer simulation model in Anylogic.

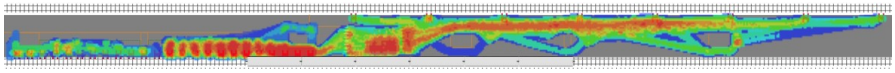


Figure 5: An display of simulation results in Anylogic.

Table 4: Different facilities plans of number of machines in service and requirement of staff

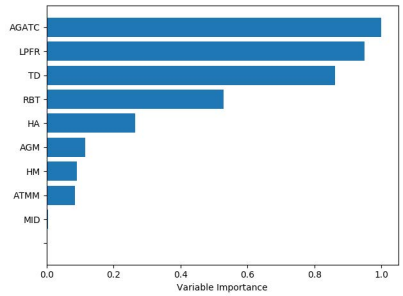
Facilities plan No.	AGRTC	ATMM	AGMTC	Requirement of employees
1	7	8	7	6
2	7	8	6	6
3	7	7	6	6
4	6	7	6	6
5	6	7	5	5
6	6	6	5	5
7	7	6	5	5
8	8	6	4	4
9	9	5	4	4
10	4	5	4	4
11	4	5	3	3
12	4	5	3	3
13	3	6	3	3

tinue at least 1 hour in one day. Considering 39 pedestrian facilities plans to be chosen for 400 operation scenarios, there are more than 15,000 combinations of operation scenarios and pedestrian facilities plans required to be simulated to obtain the performance of evaluation indicators. Considering the randomness of simulation, each combination was simulated for five times. In total, more than 75,000 results of performance of evaluation indicators are provided to random forest.

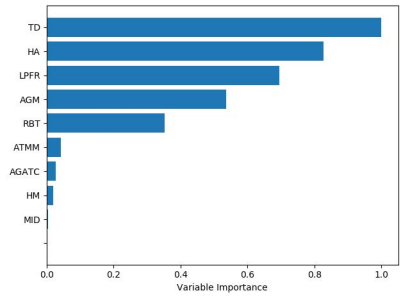
4.3 Random forest results

To improve the accuracy of machine learning, *GridSearchCV* in scikitlearn package is applied for hyperparameter tuning on *max_depth*, *max_features*, *max_leaf_nodes*, etc. Cross validation implement to avoid overfitting. Finally, the scores of three performance indicators are larger than 9.5.The score of the fitting on transfer capacity is 0.955606873045, on average transfer time is 0.97330181704 and on level-of-service is 0.972644965479. It shows the accuracy of random forest is acceptable for the fit.

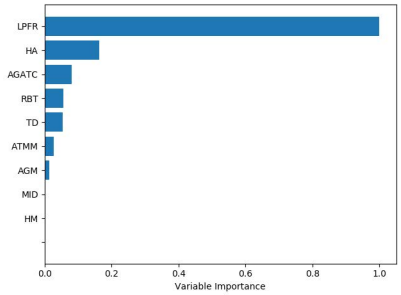
Based on the fitting results, the importance scores are shown in Figure 6. It shows that besides the facilities plan attributes, transfer demand and the ratio of buying tickets get the higher scores. Hence, we analyze the influence of them in the next.



(a) importance scores on fitting transfer capacity



(b) importance scores on fitting transfer time



(c) importance scores on fitting level-of-service

Figure 6: importance scores of the random forest

4.4 The integrated solution

The choosing of employment plan and different diagrams of pedestrian facilities plan and staff assignment plan for each day are results of the integrated mathematical model. With the potential operation scenarios in the future, suggestions for staff adjustment are provided to the station manager.

The cost and work time limit of staff

As stated before, the staff consist of regular employees and secondment employees, whose costs to assign and working time limits are different from each other. 2 teams of regular employees and 1 team of secondment employees can be assigned in Xipu station. The cost, working time range, and upper bound of working time of staff are shown in Table 5. We design 3 employment plans as shown in Table 6 for Xipu station. Each day is divided into 16 periods because the operation time of one day is from 6:00 to 22:00.

Table 5: Cost and work time limit of employees

Staff	Beginning of working time	End working time	Cost	Upper bound
Regular team 1	6:00	1400	250 per day	8
Regular team 2	14:00	22:00	250 per day	8
Secondment team	6:00	22:00	100 per hour	12

Table 6: Number of employees for each employment plan

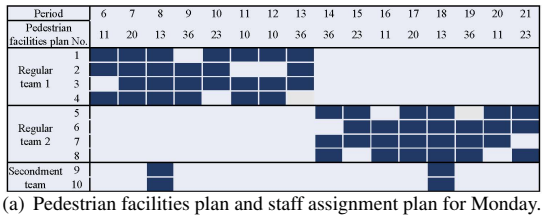
Plan No	Number of employees in Regular team 1	Number of employees in Regular team 2	Number of employees in secondment team
1	3	3	3
2	4	4	2
3	5	5	1

The different solution for four days in a week

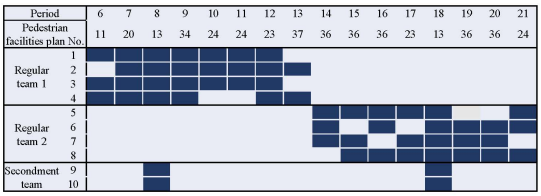
With the given requirement of performance of evaluation indicators, pedestrian facilities plans and staff assignment plans can be obtained by integrated model. Different days in a week need different pedestrian facilities plans and staff assignment plans because of different operation scenarios. A optimal solutions for Monday, Wednesday, Friday and Sunday under employment plan 2 are shown in Figure 7.

Fairness of working time of regular employees is guaranteed in all days according to Figure 7. No difference of working time is more than 1 hour between any two regular employees. We rarely use secondment employees in order to save cost since the secondment employees cost relatively higher. What's more, the integrated mathematical model provides the certain pedestrian facilities plan which can satisfy the given operation scenario and requirement of the performance of evaluation indicators for each period.

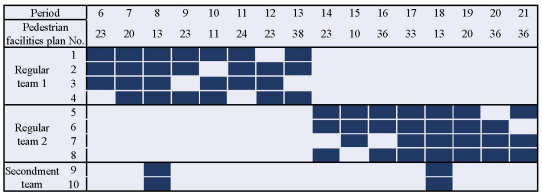
The evaluation of each plan is studied. With the given condition of operation scenarios and requirements of performance, the solution of average working time of regular employees, the total working time of secondment employees and average labor cost of Monday, Wednesday, Friday and Sunday in a week as shown in Table 7.In the table, A is short for Average working time of regular employees per day, B is short for total working time of secondment employees per day, and C is short for Labor cost per day. Average working time of regular employees aims to evaluate the work intensity of the regular. Total working time of secondment employees is aim to evaluate the rationality of the employment plan. Average labor cost aims to help station manager choose employment plan.



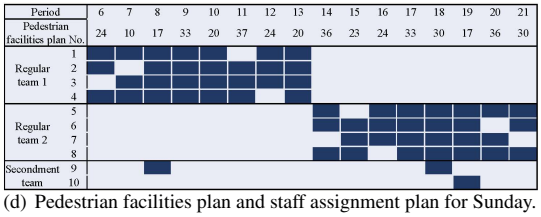
(a) Pedestrian facilities plan and staff assignment plan for Monday.



(b) Pedestrian facilities plan and staff assignment plan for Wednesday.



(c) Pedestrian facilities plan and staff assignment plan for Friday.



(d) Pedestrian facilities plan and staff assignment plan for Sunday.

Figure 7: Solutions of four days in week under employment plan 2

Table 7: The results of different employment plan

	Plan 1			Plan 2			Plan 3		
	A/hour	B/hour	C/money	A/hour	B/hour	C/money	A/hour	B/hour	C/money
Monday	8	8	2300	6.5	4	2400	5.4	2	2700
Wednesday	8	8	2300	6.5	4	2400	5.4	2	2700
Friday	8	9	2400	6.625	4	2400	5.5	2	2700
Sunday	8	10	2500	6.875	3	2300	5.8	0	2500
Average	8	8.75	2375	6.625	3.75	2375	5.525	1.5	2650

By observing Table 7, if the employment plan 1 is chosen, it is impossible for each regular employee to work all the time from beginning to ending continuously. Besides,

the average of the total working time of secondment employees is over than 8 means that the cost of assigning secondment staff is high and the station manager can add a regular employee. If the employment plan 3 is chosen, the labor cost is too high for station manager and the average working time of regular employees per day is low which means the regular employees have too more rest time and the cost of employing a regular employee is not fully utilized. Therefore, employment plan 2 is the most appropriate plan. Because the regular employees have a reasonable rest time, the working time of secondment employees is not too long, and the labor cost is the lowest.

Different ratios of buying tickets and transfer demand

With the promotion of e-tickets and QR-code, transfer passenger may no longer need to buy tickets, and due to the development of rail transit, the number of transfer passenger will increase. These will lead new operation scenarios in the station. Therefore, different ratios including 35%, 45% and 55% of buying metro tickets and increasing transfer demand are performed. The current transfer demand is considered low and the other types of operation scenarios with middle and high transfer demands are designed. The results of three different employment plans in Wednesday are reported in Table 8. And A,B and C have the same meaning in Table 7 As reported in Table 8, with the increasing of transfer demand, the

Table 8: The results of different employment plans in the potential operation scenarios

		Low transfer demand			Middle transfer demand			High transfer demand		
	Ratio	A/hour	B/hour	C/money	A/hour	B/hour	C/money	A/hour	B/hour	C/money
Plan 1	0.55	8	8	2300	8	11	2600	8	13	2800
	0.45	8	13	2800	8	16	3100	8	21	3600
	0.35	8	12	2700	8	14	2900	8	18	3300
Plan 2	0.55	6.5	4	2400	6.625	6	2600	6.75	7	2700
	0.45	7.125	4	2400	7.25	6	2600	7.5	9	2900
	0.35	7.25	2	2200	7.25	4	2400	7.5	6	2600
Plan 3	0.55	5.4	2	2700	5.7	2	2700	5.9	2	2700
	0.45	5.9	2	2700	6.2	2	2700	6.6	3	2800
	0.35	6.0	0	2500	6.2	0	2500	6.6	0	2500

working time of employees and labor cost of station increase regardless of whichever employment plan is chosen. However, the operation scenarios whose ratio of buying tickets is 0.45 requiring the highest working time of employees and labor cost. The other 27 results of Monday, Friday and Sunday are similar to this. The reason for it is that when the ratio of buying tickets is 0.55, more employees are required for automatic tickets machines for metro; when the ratio of buying tickets is 0.35, fewer employees are required for automatic gates; while the ratio of buying tickets is 0.45, both automatic tickets machines and automatic gates require more employees. With the high transfer demand and the ratio of buying tickets as 0.45 and 0.35, if employment plan 2 is chosen, the average working time of regular employees is 7.5, which means that some regular employees have to work continuously for 8 hours. The employment plan is not appropriate for the same reason as stated before. The results show that if employment plan 3 is chosen, the working time of employees and labor cost can be both reduced. Each regular employee can have rest time and the secondment employees even can be idle when the ratio of buying tickets is 0.35. Therefore, if transfer

demand is high and the ratio of buying tickets is reduced below 0.45, the employment plan 3 is the better choice.

5 Conclusions

This paper proposes a novel mix-integer linear problem formulation to deal with the integrated optimization of pedestrian facilities planning and staff assignment. The staff assigning and facilities planning are defined as binary variables to get the actual operation plan. To obtain the fairness, the actual working time, the average working time and the deviation time to the average working time are computed. The station operation requirements for the facilities plan select the qualified pedestrian facilities plan. The staff assignment rule addressed to obtain the meet the practical constraints. To acquire the performance indicators, an agent-based simulation model on Anylogic is developed to provide a huge of train data for the machine learning. Moreover, the random forest, a machine learning method, performs well to fit the non-linear relationships among the operation attributes, facilities plan attributes and the performance indicators on transfer capacity, average transfer time and level-of-service. The experiment results show the integrated model can return pedestrian facilities plans which meet the level of service requirements and assign employees fairly of each period in a day and minimize the labor cost for Xipu station. Moreover, the solutions of pedestrian facilities plan and staff assignment plan for possible operation scenarios in future show the labor cost for different employment plans. It could help the station manager to select the reasonable employment plan.

Future research efforts can be dedicated to investigating other working rules in practical operation on the staff assignment, and the staff rostering scheduling in a week. What's more, more machine learning methods, like the SVM (Support Vector Machine), ANN(Artificial Neural Network) and RNN(Recurrent Neural Network), could be tested to improve the accuracy.

Acknowledgments
This work is supported by National Natural Science Foundation of China (No. 61603317), National Key Research and Development Program of China (No. 2017YFB1200700-1), and Special Fund for Basic Scientific Research of Central Universities (No. 2682017CX023). The generous support of our colleagues Jun Zhao is greatly appreciated.

References

The AnyLogic Company, 2019. *AnyLogic Help*, Retrieved from https://help.anylogic.com/index.jsp.

Bandini, S., Mondini, M., Vizzari, G., 2014. "Modelling negative interactions among pedestrians in high density situations", *Transportation Research Part C: Emerging Technologies*, vol.40, pp. 251-270.

Berbey, A., Galan, R., Sanz, J.D., Caballero, R., 2012. "A fuzzy logic approach to modelling the passengers' flow and dwelling time", *WIT Transactions on The Built Environment*, vol.128, pp. 359-369.

Corman, F., D'Ariano, A.,Marra, A.D.,Pacciarelli, D.,Samà, M., 2017. "Integrating train scheduling and delay management in real-time railway traffic control", *Transportation Research Part E: Logistics and Transportation Review*, vol.105, pp. 213–239.

D'Acierno, L., Botte, M., Placido, A., Caropreso, C., Montella, B., 2017. "Methodology for Determining Dwell Times Consistent with Passenger Flows in the Case of Metro Service.", *Urban Rail Transit*, vol.3(2), pp. 73–89

Dollevoet, T., Corman, F., D'Ariano, A., Huisman, D., 2014. "A mathematical model for periodic scheduling problems", *Flexible Services and Manufacturing Journal*, vol.26, pp. 490–515.

Ernst, A.T., He, Q.,Jiang, H.,Krishnamoorthy, M.,Sier, D., 2004. "Staff scheduling and rostering: A review of applications, methods and models", *European journal of operational research*, vol.153, pp. 3–27.

Ghofrani, F., He, Q.,Goverde, R.M.,Liu, X., 2018. "Recent applications of big data analytics in railway transportation systems: A survey", *Transportation Research Part C: Emerging Technologies*, vol.90, pp. 226–246.

Hänseler, F.S., 2016. *Modeling and estimation of pedestrian flows in train stations*, INTER,EPFL,Lausanne.

Transportation Research Board, 2000. *Highway Capacity Manual*, Transportation Research Board,Washington, DC.

Heuvel, J.V.D., Voskamp, A., Daamen, W., Hoogendoorn, S.P., 2015. "Using Bluetooth to estimate the impact of congestion on pedestrian route choice at train stations", *Traffic and Granular Flow '13*, pp. 73–82.

Heuvel, J.V.D., Ton, D., Hermansen, K., 2016. "Advances in Measuring Pedestrians at Dutch Train Stations Using Bluetooth, WiFi and Infrared Technology", *Traffic and Granular Flow '15*, pp. 11–18.

Hoy, G., Morrow, E., Shalaby, A., 2016. "Use of Agent-Based Crowd Simulation to Investigate the Performance of Large-Scale Intermodal Facilities", *Transportation Research Record: Journal of the Transportation Research Board*, vol.2540, pp. 20–29.

Hu, L., Jiang, Y., Zhu, J.,Chen, Y., 2017. "A PH/PH (n)/C/C state-dependent queuing model for metro station corridor width design", *European Journal of Operational Research*, vol.240, pp. 109–126.

Huisman, D., 2007. "A column generation approach for the rail crew re-scheduling problem", *European Journal of Operational Research*, vol.180, pp. 163–173.

Ibarra-Rojas, O.J., Rios-Solis, Y.A., 2012. "Synchronization of bus timetabling", *Transportation Research Part B: Methodological*, vol. 46, pp. 599-614.

Kecman, P., Goverde, R.M.P., 2015. "Predictive modelling of running and dwell times in railway traffic", *Public Transport*, vol.7, pp. 295–319.

King, D., Srikukenthiran, S., Shalaby, A, 2014. "Using Simulation to Analyze Crowd Congestion and Mitigation at Canadian Subway Interchanges", *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2417, pp. 27–36.

Lai, Y.-C., Huang, Y.-A., Chu, H.-Y., 2014. "Estimation of rail capacity using regression and neural network", *Neural Computing and Applications*, vol.25, pp. 2067–2077.

Lam, W. H. K., Cheung, C. Y., Lam, C. F., 1999. "A study of crowding effects at the Hong Kong light rail transit stations", *Transportation Research Part A: Policy and Practice*, vol.33, pp.401–415.

Pel, J. A., Bel, N. H., Pieters, M., 2014 "Including passengers' response to crowding in the Dutch national train passenger assignment model", *Transportation Research Part A: Policy and Practice*, vol.66, pp. 111-126.

Pu, Y., 2017. *Capacity Analysis of the Union Station Rail Corridor using Integrated Rail and Pedestrian Simulation*, Department of Civil Engineering,University of Toronto,Toronto.

Shafique, M. A., Hato, E, 2017. “Classification of travel data with multiple sensor information using random forest”, *Transportation research procedia*, vol.22, pp. 144–153.

Shlayan, N., Kurkcu, A., Ozbay, K., 2016. “Exploring Pedestrian Bluetooth and WiFi Detection at Public Transportation Terminals”, *2016 IEEE 19th International Conference on Intelligent Transportation System*, pp. 229-234.

Tirachini, A., Hensher, D. A., Rose, J. M., 2013. “Crowding in public transport systems: Effects on users, operation and implications for the estimation of demand”, *Transportation Research Part A: Policy and Practice*, vol.53, pp. 36–52.

Wong, R. C., Yuen, T. W., Fung, K. W.,Leung, J. M., 2008. “Optimizing timetable synchronization for rail mass transit”, *Transportation Science*, vol.42, pp. 57–69.

Xu, X., Liu, J., Li, H., Zhou, Y., 2013. “Probabilistic model for remain passenger queues at subway station platform”, *J.Cent.South Univ*, vol.20, pp. 837-844.

Studies on the validity of the fixed-speed approximation for the real time Railway Traffic Management Problem

Pierre Hosteins [a,1], Paola Pellegrini [b], Joaquin Rodriguez [a]
[a] Univ Lille Nord de France, IFSTTAR, COSYS, ESTAS
F-59650 Villeneuve dAscq, France [b] Univ Lille Nord de France, IFSTTAR, COSYS, LEOST
F-59650 Villeneuve dAscq, France
[1] E-mail: pierre.hosteins@ifsttar.fr, Phone: +33 (0) 320438358

Abstract
In railway traffic management problems, a frequent approximation is the one of fixed-speed, *i.e.* the trains either run at their cruise speed or are stopped immediately without considering the acceleration and deceleration phases due to arising conflicts on the infrastructure. We assess the validity of the fixed-speed approximation for train speed dynamics in the real time Railway Traffic Management Problem. This is done through a statistical analysis on a number of perturbed scenarios on different railway infrastructures, for different objective functions commonly used in the literature. For each scenario, we analyze the ranking of the generated solutions both in the fixed-speed approximation, obtained by solving the optimization model, and with the variable-speed dynamics, obtained through micro simulation with the OpenTrack software. Our results indicate that some objective functions can be considered reliable when used in conjunction with the fixed-speed approximation, while others require more detailed studies. We also propose a modified fixed-speed approximation to better reflect the behaviour of trains speed dynamics and study its efficiency.

Keywords
Traffic Management, speed profile, fixed-speed approximation, Mathematical Programming, optimization

1 Introduction

At peak times, in critical parts of the railway network of many European countries, traffic is planned to occupy the infrastructure almost without interruption. When this happens, a delay of one train, even of a few seconds, may propagate to several other trains in a snowball effect: if one or more trains, running at the planned speed, would require the same piece of infrastructure concurrently, all but one of them must slow down or even stop to ensure safety. In this case, a *conflict* is said to emerge. Conflicts are particularly critical at *junctions*, where multiple lines cross. Here, the precedence between the involved trains must be specified and may have a strong impact on delay propagation. Moreover, it is often possible to route trains in different ways to go through a junction, also impacting delay propagation. This translates into a difficult combinatorial optimization problem.

Today, conflict management is performed mostly manually by dispatchers. Several algorithms have been proposed to solve the described routing and scheduling problem (Cac-

chiani et al., 2014), which is often named real-time Railway Traffic Management Problem (rtRTMP). Many variants exist to tackle such a problem, be it on the modelling side or on the methodology for recovering efficient solutions. For example, such models can advocate a macroscopic infrastructure representation, ignoring some details such as the locking and releasing of track sections. Conversely they can advocate a microscopic modelling of the infrastructure, taking into account all the necessary operational details. Different models focus on different aspects of the problem, such as, e.g., trains or passengers, details of the train speed variation dynamics when brakes and accelerations are necessary due to conflicts, etc... The choice of these aspects results in different objective functions to be optimized given the operational constraints of the problem. There also exists a wide range of algorithms to solve the rtRTMP. Exact resolutions usually make use of commercial branch-and-cut solvers to solve a Mixed Integer Linear Program which models the rtRTMP, as in, e.g., Caimi et al. (2011); Corman et al. (2012); D'Ariano et al. (2007a); Lamorgese and Mannino (2015); Meng and Zhou (2014); Törnquist and Persson (2007). Heuristic and meta-heuristic approaches have also been devised to tackle the problem. Prominent examples are the works of Khosravi et al. (2012); Dündar and Şahin (2013); Sama et al. (2017).

The impact of the modelling choices on the actual performance of the algorithms due to the validity of the underlying assumptions has not been deeply studied yet. In this paper, we try to assess the validity of the assumption underlying one of these choices, specifically the so-called fixed-speed model for the unplanned brakes and accelerations. According to this model, the speed profile of trains traveling according to the planned timetable is precisely computed, when the trains are free to reach their desired speed on each track section without encountering any conflict. However, if a train needs to slow down or stop due to traffic perturbations, the fixed-speed approximation considers that it passes from its planned speed to a halt in no space and time. When the track is free for the train to go, it reaches its planned speed in, again, no space and time. This means that there is inifinite acceleration and braking rate, which of course is not realistic. We illustrate in Figure 1 the difference in the speed profiles of a given train when it crosses an infrastructure without conflict, as well as when a conflict arises, using the exact speed dynamic and the fixed-speed approximation. The red part in the speed profile of train A shows where the train speed diverts from its basic conflictless profile due to the conflict with train B. While this part displays a smooth change in the speed value due to acceleration and deceleration, we can see that the speed jumps directly to 0 in the fixed-speed approximation and remains null until the train is free to move again, after which it jumps back to its maximum value over the track section.

Many optimization models regarding railway traffic management have been proposed, based on the fixed-speed approximation, as e.g. Corman et al. (2010); Pellegrini et al. (2014). The assumption behind the validity of this modelling choice is the preservation of the relative quality of solutions. In particular, if one takes two solutions A and B, then if the routing and scheduling decisions in A are better than the one in B according to the fixed-speed model, they will be better also in reality. Although the intuition suggests that this is likely to be true most often, no deep analysis has been performed so far to support this intuition. To the best of our knowledge, the only attempt to study quantitatively the fixed-speed approximation is the work of Sobieraj et al. (2011), which proposes experiments to characterize relations between some specific traffic conditions and its quality. In the following, we will try to assess on general grounds the quality of the fixed-speed approximation on a few different railways infrastructures and for different objective functions. We stress that our study aims at quantifying the quality of the fixed-speed approximation and therefore its

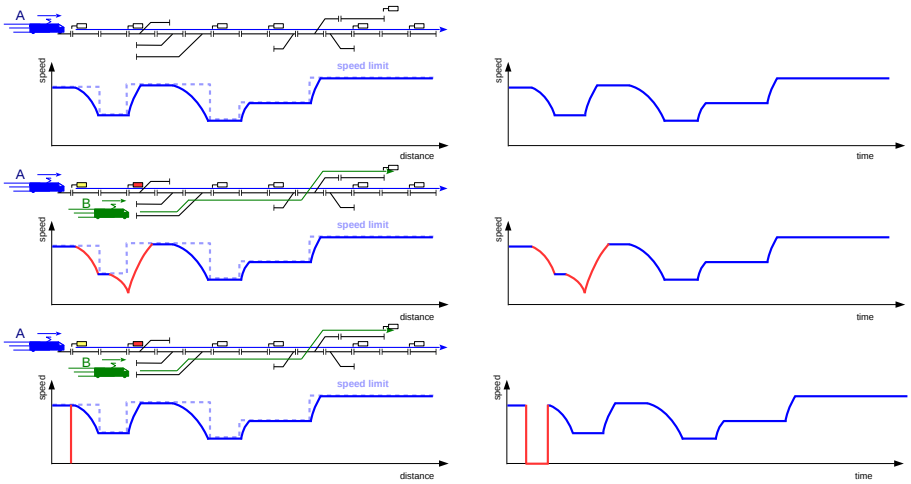


Figure 1: Speed profiles of train A (blue) in three cases (from top to bottom): no conflict, conflict with train B (green) with accurate speed dynamics (simulation) and conflict with the same train using the fixed-speed approximation. The left curves represent the value of speed as a function of distance while the right curves represente the same speed as a function of time. Dashed lines represent the speed limit on a given track section.

validity in terms of the ranking of the best solution returned by an optimization model when using such an approximation. Indeed, all solutions found are feasible from the point of view of the model constraints but the approximation used for the trains dynamics will change the value of the trains delay and therefore the value of the objective function. Let us note, however, that not all the proposed models in the literature make use of the fixed-speed approximations. Some examples exist which propose an iterative loop which solves the model with approximated speed profiles and then improve such speed profiles when they are found to be infeasible or suboptimal in the current solution D'Ariano et al. (2007b); Mazzarello et al. (2007); Lüthi (2009). This trend of work is sometimes preoccupied with the reduction of power consumption on the railway network, which is directly linked to the trains speed and accelerations. Examples of such energy consumption oriented works can be found in D'Ariano and Albrecht (2006); Albrecht (2009).

In the next section we will recall the MILP formulation already used in previous works to solve the rtRTMP and also propose a refined model that tries to better model the effects of unanticipated braking on the running time of trains. In section 2, we recall the formulation of the rtRTMP and introduce a refined model to try to take into account the effect of deceleration on the trains running time. In section 3, we will detail our methodology to assess the validity of the fixed-speed approximation by comparing its results to the variable-speed dynamic obtained by a simulation tool. In Section 4 we will present the different infrastructures that we use for our numerical experiments and then provide a comparison between the approximated and simulated speed dynamics. Finally we conclude in Section 5.

2 Integer Linear formulation for the real-time Railway Trafic Management Problem

2.1 Formulation for the classic rtRTMP with fixed-speed approximation

We will use a MILP to solve the rtRTMP, similar to the previous works of Pellegrini et al. (2014, 2015), called RECIFE-MILP. It models the infrastructure at the microscopic level and implements the route-lock sectional-release interlocking system (Pachl, 2008). The tracks are divided into *track-circuits*, i.e., track segments on which the presence of a train is automatically detected. *Block sections* represent groups of track-circuits whose access is controlled by a signal. Moreover, before a train can *occupy* a sequence of block sections, all their track-circuits must be reserved for the train itself. RECIFE-MILP uses the following sets:

- T: the set of trains;
- Θ: set of train types;
- R_t: the set of routes available to train $t \in T$, with $R = \cup_{t \in T} R_t$ the total set of routes;
- TC_t: the set of track-circuits which can be used by train $t \in T$;
- TC^r: the set of track-circuits belonging to route $r \in R$;
- $OTC_{ty,r,tc}$: set of track-circuits occupied by a train $t \in T$ of type $ty \in \Theta$ along $r \in R_t$ if t's head is at the end of $tc \in TC^r$ ($\emptyset$ if t shorter than tc);
- $TC(tc, tc', r)$: set of track-circuits between tc and $tc' \in RT^r$ along $r \in R$;
- S_t, $TCS_{t,s}$: set of stations where $t \in T$ has a scheduled stop and set of track-circuits that can be used by t for stopping at $s \in S_t$;

and parameters:

- tc_0 and tc_∞: entry and the exit locations of the infrastructure considered;
- $sched_t$: scheduled arrival time of traint $t \in T$ at destination;
- ty_t: type corresponding to train t (train characteristics);
- $init_t$, $exit_t$: earliest time at which train $t \in T$ can be operated and earliest time at which it can reach its destination given $init_t$, the route assigned in the timetable and the intermediate stops;
- $i(t', t)$: indicator function equal to 1 if t' and t use the same rolling stock and t results from the turnaround, join or split of t', 0 otherwise; $ms \equiv$ minimum separation between the arrival and the departure of two trains using the same rolling stock;
- $rt_{ty,r,tc}$, $ct_{ty,r,tc}$: running and clearing time of $tc \in RT^r$ along $r \in R$ for a train of type $ty \in \Theta$;
- $ref_{r,tc}$: reference track-circuit for the reservation of $tc \in TC^r$ along $r \in R$, depending on block-sections structure;

- $e(tc, r)$: indicator function equal to 1 if track-circuit $tc \in TC^r$ belongs to either the first or the last block section of $r \in R$, 0 otherwise;
- $bs_{r,tc}$: block section including track-circuit $tc \in TC^r$ along route $r \in R$;
- for_{bs}, rel_{bs}: formation and release time for block section bs;
- S_t, $TCS_{t,s}$: set of stations where $t \in T$ has a scheduled stop and set of track-circuits that can be used by t for stopping at $s \in S_t$;
- $dw_{t,s}, a_{t,s}, d_{t,s}$: minimum dwell time, scheduled arrival and scheduled departure times for train $t \in T$ at station $s \in S_t$;
- $p_{r,tc}$, $s_{r,tc}$: set of track-circuits preceding and following $tc \in RT^r$ along $r \in R$;
- M: a large constant.

We also make use of the following variables:

- $sU_{t,tc}$, $eU_{t,tc}$: continuous positive variable representing the time at which $tc \in TC_t$ starts and ends being utilized by $t \in T$;
- $x_{t,r}$: binary variable equal to 1 if train $t \in T$ uses route $r \in R_t$, 0 otherwise;
- $y_{t,t',tc}$: binary variable equal to 1 if train $t \in T$ utilizes track-circuit tc before train t', such that the index t is smaller than the index t' ($t \prec t'$), with $tc \in TC_t \cap TC_{t'}$, and 0 otherwise;
- $o_{t,r,tc}$: time in which $t \in T$ starts the occupation of $tc \in TC^r$ along $r \in R_t$;
- $l_{t,r,tc}$: longer stay of $t \in T$'s head on $tc \in TC^r$ along $r \in R_t$, due to dwell time and scheduling decisions (delay).

In addition to the existing track-circuits, we introduce the dummy ones: tc_0 and tc_∞ which represent the entry and the exit locations of the infrastructure considered. Depending on the objective function used, we also have to define the following variables:

- D_t: delay suffered by train t when exiting the infrastructure;
- Δ: maximum secondary delay among all trains;
- δ_t: binary variable equal to 1 if train $t \in T$ suffers some delay compared to its original timetable.

The model has to respect the following sets of constraints:

$$o_{t,r,tc} \geq init_t \, x_{t,r} \qquad \forall t \in T, r \in R_t, tc \in TC^r, \tag{1}$$

$$o_{t,r,tc} \leq M x_{t,r} \qquad \forall t \in T, r \in R_t, tc \in TC^r, \tag{2}$$

$$o_{t,r,tc} = o_{t,r,p_{r,tc}} + l_{t,r,p_{r,tc}} + rt_{r,ty_t,p_{r,tc}} x_{t,r} \qquad \forall t \in T, r \in R_t, tc \in TC^r, \tag{3}$$

$$o_{t,r,s_{r,tc}} \geq \sum_{\substack{s \in S_t: \\ tc \in TCS_{t,s} \cap TC^r}} d_{t,s} x_{t,r} \qquad \forall t \in T, r \in R_t, tc \in \bigcup_{s \in S_t} TCS_{t,s}, \tag{4}$$

$$l_{t,r,s_{r,tc}} \geq \sum_{\substack{s \in S_t: \\ tc \in TCS_{t,s} \cap TC^r}} dw_{t,s} x_{t,r} \qquad \forall t \in T, r \in R_t, tc \in \bigcup_{s \in S_t} TCS_{t,s}, \tag{5}$$

$$\sum_{r \in R_t} x_{t,r} = 1 \qquad \forall t \in T, \tag{6}$$

$$\sum_{\substack{r \in R_t, tc \in TC^r: \\ p_{r,tc} = tc_0}} o_{t,r,tc} \geq \sum_{\substack{r \in R_{t'}, tc \in TC^r: \\ s_{r,tc} = tc_\infty}} o_{t',r,tc} + (ms + rt_{r,ty_{t'},tc}) x_{t',r}$$

$$\forall t, t' \in T : i(t',t) = 1, \tag{7}$$

$$\sum_{\substack{tc \in TC_t: \\ \exists r \in R_t, p_{r,tc} = tc_0}} sU_{t,tc} \leq \sum_{\substack{tc \in TC_{t'}: \\ \exists r \in R_{t'}, s_{r,tc} = tc_\infty}} eU_{t',tc} \qquad \forall t, t' \in T : i(t',t) = 1, \tag{8}$$

$$sU_{t,tc} = \sum_{\substack{r \in R_t: \\ tc \in TC^r}} \left(o_{t,r,ref_{r,tc}} - for_{bs_{r,tc}} x_{t,r} \right) \qquad \forall t \in T,\ tc \in TC_t :$$

$$(\nexists t' \in T : i(t',t) = 1) \vee (\forall r \in R_t : ref_{r,tc} \neq s_{r,tc_0}), \tag{9}$$

$$eU_{t,tc} = \sum_{\substack{r \in R_t: \\ tc \in TC^r}} o_{t,r,ref_{r,tc}} + (for_{bs_{r,tc}} + rel_{bs_{r,tc}})\, x_{t,r} + ul_{t,r,tc} \qquad \forall t \in T, tc \in TC_t, \tag{10}$$

$$eU_{t,tc} - M(1 - y_{t,t',tc}) \leq sU_{t',tc} \qquad \forall t, t' \in T, \text{index } t < \text{index } t',\ tc \in TC_t \cap TC_{t'} :$$

$$i(t,t') \sum_{r \in R_t} e(tc,r) = 0 \ \wedge \ i(t',t) \sum_{r \in R_{t'}} e(tc,r) = 0, \tag{11}$$

$$eU_{t',tc} - My_{t,t',tc} \leq sU_{t,tc} \qquad \forall t, t' \in T, \text{index } t < \text{index } t',\ tc \in TC_t \cap TC_{t'} :$$

$$i(t,t') \sum_{r \in R_t} e(tc,r) = 0 \ \wedge \ i(t',t) \sum_{r \in R_{t'}} e(tc,r) = 0. \tag{12}$$

Eqs. (1) and (2) force train t to be operated no earlier than $init_t$ on its chosen route. In Eq. (3), a train starts occupying a given track-circuit after spending its effective running time in the preceding one (if the route is used). Eq. (5) ensures that train t which stops at station s along route r does not leave track-circuit $tc \in TCS_{t,s}$ before the scheduled departure time from s. In (6), a single route is chosen for train t. Eqs. (7) and (8) are used to guarantee consistency for trains using the same rolling stock, i.e. minimum time between arrival and departure of such trains and consistent *utilization* time of the track-circuits. The utilization time is defined as the sum of the *reservation* and the *occupation* time. In Eq. (9), a train's utilization of a track-circuit starts as soon as the train starts occupying the track-circuit $ref_{r,tc}$ along one of the routes including it, minus the formation time. Furthermore, in Eq. (10), the utilization of a track-circuit lasts till the train utilizes it along any route, plus the formation and the release time. Here $ul_{t,r,tc}$ is the total utilization time which includes: the running time of all track-circuits between $ref_{r,tc}$ and tc, the longer stay of the train's head on each of these track-circuits and the clearing time of tc. Finally, Eqs. (11) and (12) ensure that the track-circuit utilizations by two trains must not overlap. We refer the reader to Pellegrini et al. (2015) for additional details about the above formulation.

In our subsequent analysis, we consider the four following objective functions commonly used in the literature:

- the total delays (i.e. the sum of delays for each train with respect to its timetable):

$$\min \quad \sum_{t \in T} w_t D_t, \tag{13}$$

$$D_t \geq \sum_{r \in R_t} o_{t,r,tc_\infty} - sched_t \qquad \forall t \in T; \tag{14}$$

- the maximum secondary delay (i.e. the maximum propagation delay between all trains):

$$\min \quad \Delta, \tag{15}$$
$$\Delta \geq D_t \qquad \forall t \in T, \tag{16}$$
$$D_t \geq \sum_{r \in R_t} o_{t,r,tc_\infty} - exit_t \qquad \forall t \in T; \tag{17}$$

- the number of delayed trains:

$$\min \quad \sum_{t \in T} \delta_t, \tag{18}$$
$$M\delta_t \geq \sum_{r \in R_t} o_{t,r,tc_\infty} - exit_t \qquad \forall t \in T; \tag{19}$$

- the total travel times of all trains in the infrastructure:

$$\min \quad \sum_{t \in T} \sum_{r \in R_t} (o_{t,r,tc_\infty} - o_{t,r,tc_0}). \tag{20}$$

2.2 Modified formulation for the Min-fixed-speed approximation

In order to try to include the effects of deceleration on the trains dynamics more accurately, we introduce a new approximation based on a slight modification of the fixed-speed based model described above. The basic idea is that when a train is forced to decelerate because of a conflict, it loses a minimum amount of time, even if in the fixed-speed approximation it would only have to stop for a couple of seconds. We therefore introduce a minimum *forfait* f for the stopping time of a train: when a train has to stop due to a conflict, we impose that its running time increases of at least f seconds. The value of f is a parameter that might depend on the infrastructure considered and the type of train considered.

In order to generalize the fixed-speed model we introduce the new binary variables:

- $\sigma_{t,r,tc}$: variable equal to 1 if train $t \in T$ needs to stop on track circuit $tc \in TC_t$ on route $r \in R_t$,

and we add the following constraints to the fixed-speed model described above:

$$l_{t,r,tc} \leq M\sigma_{t,r,tc}, \qquad t \in T, r \in R_t, tc \in TC_t, \tag{21}$$
$$l_{t,r,tc} \geq f\sigma_{t,r,tc}, \qquad t \in T, r \in R_t, tc \in TC_t. \tag{22}$$

Hence, when variable l must be greater than 0 on a certain track-circuit due to some conflict, the corresponding binary variable σ is equal to 1, thus forcing l to be at least equal to the forfait value f.

3 Methodology

We now present the statistical methodology chosen for our numerical experiments and analysis. In order to perform a relevant statistical analysis, we need to generate a large set of

different solutions for a given perturbation scenario. For a given scenario on a given infrastructure, we generate five hundred solutions[1]. For each solution, we randomly fix the route of the trains among all available routes. We also randomly set 5% of scheduling decisions (y variables in the MILP described in Section 2). The remaining scheduling decisions are decided by solving RECIFE-MILP with CPLEX 12.8 with three minutes of running time and extracting the best found solution[2]. These solutions are then executed using the OpenTrack microscopic railway simulator for obtaining the exact speed profile results and we use CPLEX 12.8 to compute the objective functions in the (min-)fixed-speed approximation using RECIFE-MILP with fixed binary variables. Then, for each scenario, we identify the ranking of the solutions in terms of objective function value according to either the (min-)fixed-speed model or the simulation results.

We focus on the relative ranking of the solutions since our aim is to identify the optimal (or best possible) solution regarding trains routing and scheduling. Therefore, it is not crucial that the approximate objective is the same as the objective obtained in the microsimulation, as long as the optimization algorithm provides the best possible solution for the network operator. In other words, the relative ranking of a set of solutions according to the fixed-speed approximation and the simulated speed dynamics (also called *variable-speed dynamics* in contrast to the fixed-speed approximation) should be the same. Therefore, when we choose to represent the solutions in a plane where the x and y-axis are the relative ranking of each solution in both (min-)fixed-speed approximation and with variable-speed dynamics, the points should be located on the diagonal. We will quantify the departure from the diagonal by performing a linear regression and extract the linear coefficient and the correlation factor. We perform the same statistical analysis for the four different objective functions described in Section 2, as well as for the min-fixed-speed approximation with different forfait values.

We also decided to perform an aggregation of the solutions which have very close objective values, in order to avoid the artificial discrimination of somewhat equivalent solutions. For each objective function we choose a threshold value θ which fixes the precision with which we intend to round the value of the objective function for each solution. In practice, we divide the value of the objective by θ, round the obtained value to the nearest integer and multiply the result by θ. This operation creates more solution with the same objective value, which won't be discriminated by the ranking procedure. We fix the value of θ to 100 seconds for the total delay and total travel time objective functions, 10 seconds for the maximum consecutive delay and 1 for the number of late trains. In practice, this means that we do not perform any aggregation for the number of late trains as we feel that the difference is already clear enough between two different solutions.

4 Statistical analysis

4.1 Railway infrastructures and scenarios

We propose an experimental analysis based on two French control areas: the Pierrefitte Gonesse junction and the Parisian St. Lazare station. These two control areas have very different characteristics. The former is a complex junction about 18 km long where freight,

[1] Note that equivalent solutions are discarded, so that some scenarios have less than 500 solutions in practice.

[2] This method aims at generating sufficiently different solutions while avoiding very bad ones which are not likely to be returned by optimization algorithms.

conventional and high speed lines cross. A weekday timetable includes 336 trains. The latter is a terminal station area of slightly more than 7 km, with 27 platforms. A weekday timetable includes 459 trains, most of which linked by rolling-stock re-utilization constraints. A simplified map of the Pierrefitte Gonesse junction is shown in Figure 2 while a map of the St. Lazare station is shown in Figure 3.

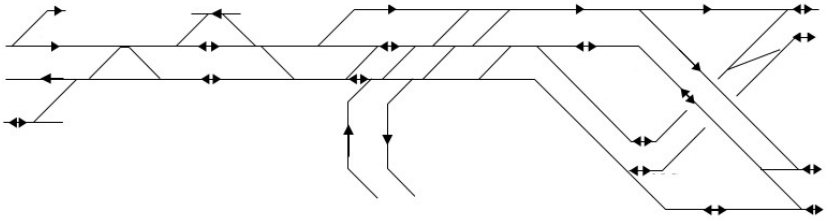


Figure 2: Simplified map of the Pierrefitte Gonesse junction.

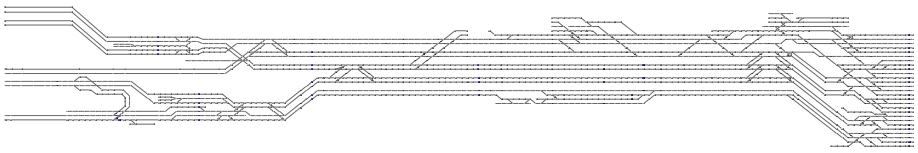


Figure 3: Simplified map of the St. Lazare station.

For each control area, we consider four different daily perturbation scenarios. These scenarios are obtained by randomly assigning an entrance delay between 5 and 15 minutes to 20% of the trains. From each of these daily scenarios we extract three peak-hour scenarios, starting at 6, 7 and 8 a.m. and lasting one hour. Therefore we have a total of twelve different perturbation scenarios for each railway infrastructure.

4.2 Numerical Results

We first present the results for the Pierrefitte Gonesse junction. We start by providing an example of the objective function values for both fixed-speed approximation and variable-speed dynamics for each solution in Figure 4 (left panels) on a representative perturbation scenario. We also plot on the same figure (right panels) the variable-speed ranking of the same solutions as a function of the fixed-speed rankings. As advocated in the Section 3, we perform a linear regression on the obtained cloud of points. We display on Figure 4 the straight line obtained, together with the diagonal. As can be seen, even though the linear regression and the diagonal are very close, there is a substantial dispersion of the solutions around it. In order to quantify the dispersion, we compute the average correlation

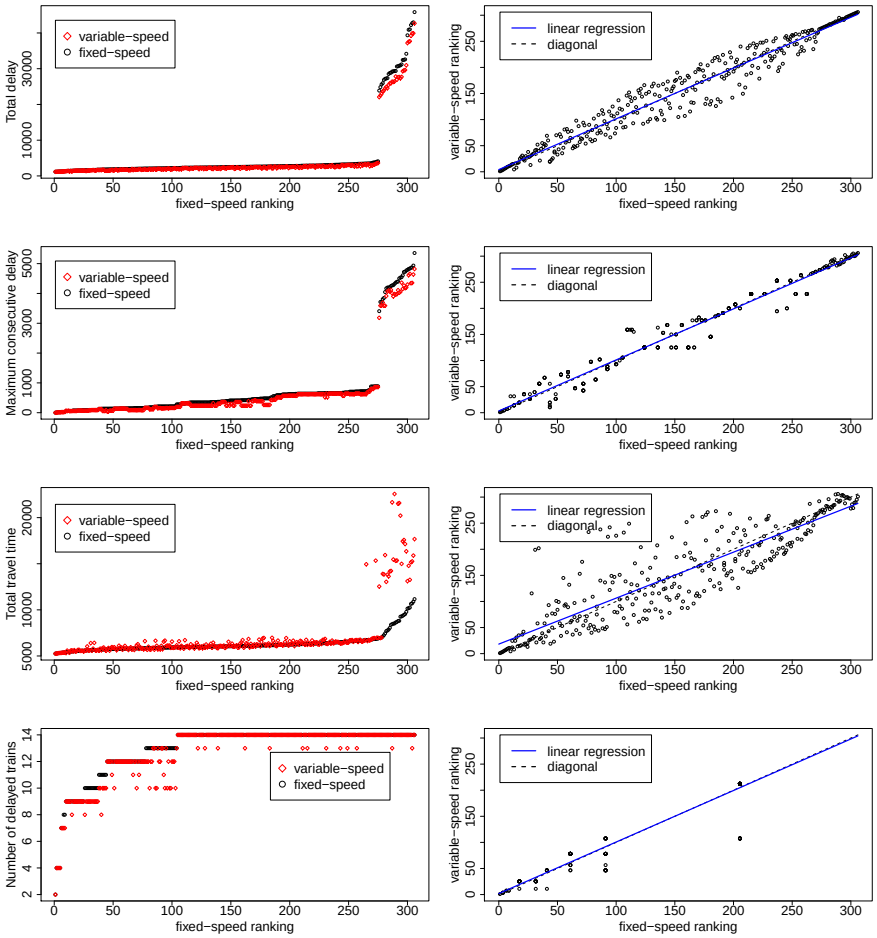


Figure 4: Value of the objective function (left panels) as a function of the ranking in the fixed-speed approximation (black circle) and variable-speed model (red diamond) and ranking with the variable-speed dynamics (right panels) as a function of the ranking in the fixed-speed approximation for a perturbation scenario on the Pierrefitte Gonesse junction. The different objective functions from top to bottom are: total delay, maximum consecutive delay, total travel time and number of delayed trains. The first three objective functions are expressed in seconds.

coefficient over the 12 perturbation scenarios and display it in Table 1, for the four objective functions, both for the fixed-speed and min-fixed-speed approximation with a forfait $f = 30$ seconds. Since our aim is to minimise a given objective, it is reasonable to suppose that an optimization algorithm will tend to provide good quality solutions. Hence, in a sense, one may consider the correlation of the set of best solutions as more important than the one of

just any solution. Therefore, we also indicate the same correlation coefficient computed on a fraction of the best solutions, i.e., on the 50% and 25% best solutions.

Table 1: Average correlation factor for the fixed-speed approximation on the Pierrefitte Gonesse junction, over the whole sample, the best 50% and the best 25% solutions, for the four objective functions. The best correlation between all four objective functions is displayed in bold font. Results are displayed both for the fixed-speed (left) and min-fixed-speed (right) approximations with a forfait of $f = 30$ seconds.

	fixed-speed			min-fixed-speed		
	whole sample	50% best	25% best	whole sample	50% best	25% best
totD	0.92	0.81	**0.74**	0.92	0.81	**0.74**
maxConsD	**0.93**	0.77	0.59	**0.93**	0.77	0.58
num	0.91	**0.89**	0.73	0.91	**0.88**	0.72
totT	0.63	0.42	0.38	0.62	0.40	0.35

What we can infer from this table is that on this particular infrastructure, the total travel time does not seem to be a very reliable objective function with respect to the fixed-speed approximation. The other three objectives provide good correlation factors on the whole sample, however the correlation decreases when we restrict ourselves to a fraction of the best solutions. This might be due to the lower statistics but also to the stabilizing effect of the worst solutions, who tend to have the same ranking in both models. In this case, the total delay and number of delayed trains seem to hold better. We observe that the results in the min-fixed-speed approximation are very similar to those of the fixed-speed approximation with a forfait of 30 seconds. We also tested the min-fixed-speed model with a larger forfait of 60 seconds and obtained no better results than the more basic fixed-speed approximation. This result seems to imply that differences between the exact speed profiles and the fix-speed approximation do not come from very short decelerations (akin to a short stop in the fixed-speed model) but from more complicated dynamics between the train conflicts. In general, the correlation coefficient shows that both models with all objective functions but the total travel time are able to properly distinguish very bad from very good solutions. However, the discrimination ability decreases when only good solutions are concerned. This says that an optimization algorithm implementing these models may possibly not return the very best solution, evaluating it slightly worse than other ones. However, these other solutions will not be much worse than the best one.

In order to provide a better idea of the statistical distribution of our results, we also display in Figure 5 histograms indicating the difference for each solution between the ranking in the fixed-speed approximation and the variable-speed model. We report one histogram per objective function. Ideally, one would expect a bell-shaped, narrow distribution centered on value 0 and with a short and thin tail. This would mean that the ranking of a solution in each model is very close, which would guarantee the validity of the approximation. What we can see in the figure is that the distribution for the total delay and maximum consecutive delay are roughly bell-shaped and centered around 0, though their tails are fatter than one would expect in an ideal situation. The distribution for the total number of delayed trains has a thinner width, though it is not centered around value 0. This may explain the good correlation factors for that particular objective function. Finally, the shape of the distribution for the total travel time sheds some light on the weak correlation factors displayed in Table 1, since it has a shape almost opposite to what one could hope for. Specifically, we

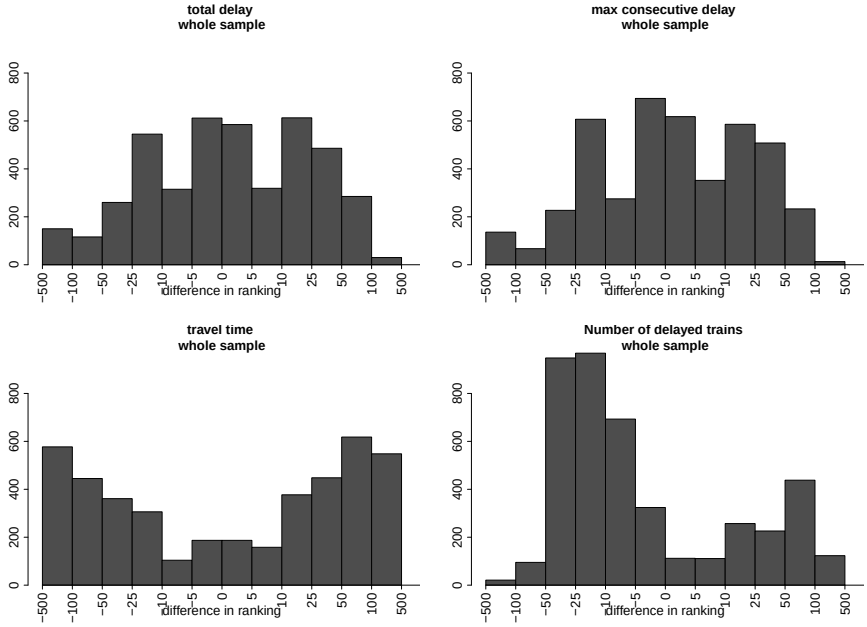


Figure 5: Histograms of the difference between the fixed-speed and variable-speed rankings for the four objective functions, on the Pierrefitte Gonesse junction (total over the twelve perturbation scenarios). Each rectangle represents the total number of solutions with a difference in ranking located in the range displayed on the x axis.

see that many solutions have an absolute rank difference larger than 50 or even 100.

Table 2: Same results as Table 1 but with aggregated solutions.

	fixed-speed			min-fixed-speed		
	whole sample	50% best	25% best	whole sample	50% best	25% best
totD	**0.99**	0.77	0.78	**0.99**	0.77	0.78
maxConsD	0.98	0.71	0.61	0.98	0.71	0.61
num	0.93	**0.91**	**0.88**	0.93	**0.91**	**0.88**
totT	0.58	0.27	0.29	0.57	0.24	0.26

In accordance with the methodology described in Section 3, we perform the same statistical analysis after rounding the objective values with respect to some precision threshold. The effet of this rounding is to aggregate some solutions together, which means that they will not be differentiated by the ranking procedure. The results displayed in Table 2 are interesting, since the correlation factors evolve significantly for different objectives. Specifically, the correlation factor for the number of late trains generally improves, especially for the 25% best solutions, while the correlation for the total travel time is even further degraded.

We can now perform the same statistical analysis for the St.Lazare station, where trains

Table 3: Average correlation factor for the fixed-speed approximation on the St.Lazare station, over the whole sample, the best 50% and the best 25% solutions, for the four objective functions. The best correlation between all four objective functions is displayed in bold font. Results are displayed both for the fixed-speed (left) and min-fixed-speed (right) approximations with a forfait of $f = 30$ seconds.

	whole sample	50% best	25% best	whole sample	50% best	25% best
totD	0.94	0.83	**0.80**	0.94	0.83	**0.80**
maxConsD	**0.96**	**0.85**	0.77	**0.96**	**0.85**	0.77
num	0.82	0.67	0.57	0.82	0.67	0.57
totT	0.87	0.68	0.60	0.87	0.68	0.60

circulate at lower speed and can encounter more conflicts at junctions. The results, as displayed in Table 3, lead to slightly different conclusions than those of the Pierrefitte Gonesse junction. In particular, it is now the number of delayed trains whose correlation factor decreases more when restricted to the best 25% solutions while the maximum delay remains more reliable in the same conditions. The total delay, however, seems to be reliable on both infrastructures and imposes itself as the most reliable objective function so far. As before, the min-fixed-speed results do not improve on the simple fixed-speed model.

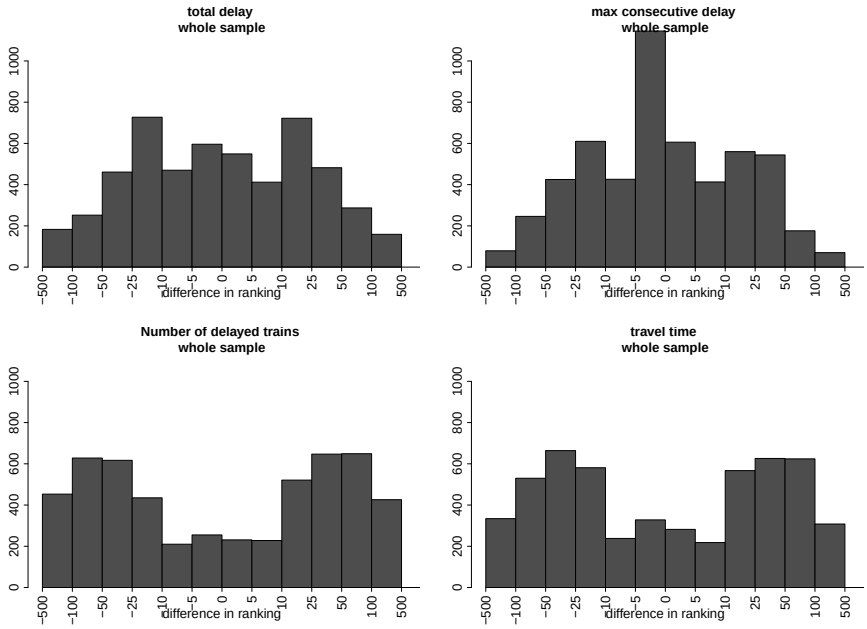


Figure 6: Same as Figure 5 with the St-Lazare station.

We complete the analysis of the second infrastructure with the histograms of the distribution of the difference in fixed and variable-speed rankings in Figure 6 and the average correlation factors with aggregated solutions in Table 4. The histograms show that the distri-

Table 4: Same results as Table 3 but with aggregated solutions.

	fixed-speed			min-fixed-speed		
	whole sample	50% best	25% best	whole sample	50% best	25% best
totD	0.95	0.88	**0.87**	0.95	0.88	**0.87**
maxConsD	**0.98**	**0.92**	**0.87**	**0.98**	**0.92**	**0.87**
num	0.85	0.75	0.72	0.85	0.75	0.72
totT	0.88	0.76	0.72	0.88	0.76	0.72

butions for the total delay and maximum consecutive delay are still roughly bell-shaped but now both the total travel time and the number of delayed trains look like double bell-shaped symmetric distributions, which once again explains the degredation of the correlation coefficient for the latter objective function. This time the correlation factors for aggregated solutions improve for all objective functions, in particular the correlations for the 25% best solutions improve between 0.07 and 0.15. This is a hint that the fixed-speed approximation cannot pretend to compute the objective functions reliably with a precision of the order of a second, or at least to discriminate between solutions which differ by a relatively small delay, in our case a handful of seconds.

5 Conclusion

We have provided a study of the validity of the so-called fixed-speed approximation for train dynamics in optimization models of the real time Railway Traffic Management Problem. We used two different railway infrastructures with various train behaviours, generated a dozen perturbation scenarios for each and hundreds of solutions for said scenarios. A statistical analysis was performed over the solutions for each scenario and four different objective functions used in the literature, to assess whether the ranking of solutions was the same with both the fixed-speed approximation and variable-speed dynamics. We also considered a slightly more refied model, in which the fixed-speed approximation is slomehow brought a step closer to the variable-speed dynamics. However, this model did not prove to behave differently from the fixed-speed one.

The average results on all perturbation scenarios show that the fixed-speed approximation seems to be reliable for most objective functions on the whole set of solutions. However, when we look specifically at the best solutions, the correlation factors between fixed and variable-speed rankings tend to decrease. Overall, the total delay objective function seems the most robust in terms of fixed-speed approximation while the total travel time provides questionable correlation factors with respect to the variable speed dynamics computed through micro-simulation. The results generally tend to improve when we aggregate the solutions with a larger granularity, which hints at the fact that the fixed-speed model cannot claim to reliably differentiate between solutions that differ by a handful of seconds. Instead, when solutions are significantly different, the fixed-speed model correctly capture the quality of the different routing and scheduling decisions. We now plan to extend the analysis to a third infrastructure in order to confirm our results.

References

Albrecth, T., 2009. "The influence of anticipating train driving on the dispatching process in railway conflict situations". *Networks Spatial Economics*, vol. 9:1, pp. 85–101.

Cacchiani, V., Huisman, D., Kidd, M., Kroon, L., Toth, P., Veelenturf, L., and Wagenaar, J., 2014. "An overview of recovery models and algorithms for real-time railway rescheduling". *Transportation Research Part B: Methodological*, vol. 63, pp. 15–37.

Caimi, G., Chudak, F., Fuchsberger, M., Laumanns, M., and Zenklusen, R., 2011. "A new resource-constrained multicommodity flow model for conflict-free train routing and scheduling". *Transportation Science*, vol. 45:2, pp. 212–227.

Corman, F., D'Ariano, A., Pacciarelli, D., and Pranzo, M., 2010. "A tabu search algorithm for rerouting trains during rail operations". *Transportation Research Part B*, vol. 44:1, pp. 175–192.

Corman, F., D'Ariano, A., Pacciarelli, D., and Pranzo, M., 2012. "Optimal inter-area coordination of train rescheduling decisions". *Transportation Research Part E*, vol. 48, pp. 71–88.

D'Ariano, A., and Albrecht, T., 2006. "Running time re-optimization during real-time timetable perturbations". In: *WIT Transactions on the Built Environment*, vol. 88. WIT Press, pp. 532–540.

D'Ariano, A., Pacciarelli, D., and Pranzo, M., 2007. " A branch and bound algorithm for scheduling trains in a railway network". *European Journal of Operational Research*, vol. 183, pp. 643–657.

D'Ariano, A., Pranzo, M., and Hansen, I.A., 2007. "Conflict Resolution and Train Speed Coordination for Solving Real-Time Timetable Perturbations". *IEEE Transactions on Intelligent Transportation Systems*, vol. 8:2, pp. 208–222.

Dündar, S., and Şahin, I., 2013. "Train re-scheduling with genetic algorithms and artificial neural networks". *Transportation Research Part C: Emerging Technologies*, vol. 27:0, pp. 1–15.

Khosravi, B., Bennell, J.A., and Potts, C.N., 2012. "Train scheduling and rescheduling in the UK with a modified shifting bottleneck procedure". In: *Delling, D., Liberti L. (Eds), Proceedings of the 12th Workshop on Algorithmic Approaches for Transportation Modelling, Optimization and Systems*, vol. 25 of OpenAccess Series in Informatics (OASIs), Dagstuhl, Germany, Schloss Dagstuhl-Leibniz-Zentrum für Informatik, pp. 120–131.

Lamorgese, L. and Mannino, C., 2015. "An exact decomposition approach for the real-time train dispatching problem". *Operations Research*, vol. 63:1, pp. 48–64.

Lüthi, M., 2009. "Improving the efficiency of heavily used railway networks through integrated real-time rescheduling". *Diss., Egenössische Technische Hochschule Zürich*.

Mazzarello, M., and Ottaviani, E., 2007. "A traffic management system for real-time traffic optimization in railways". *Transportation Research Part B*, vol. 41:2, pp. 246–274.

Meng, L., and Zhou, X., 2014. "Simultaneous train rerouting and rescheduling on an n-track network: A model reformulation with network-based cumulative flow variables". *Transportation Research Part B: Methodological*, vol. 67, pp. 208–234.

Pachl, J. (2008). 'Timetable design principles". In Hansen, I. and Pachl, J., editors, *Railway Timetable & Traffic*, chapter 2, pages 9–42. Eurailpress — DVV Rail Media, Hambourg, Germany.

Pellegrini, P., Marlière, G., and Rodriguez, J., 2014. "Optimal train routing and scheduling for managing traffic perturbations in complex junctions". *Transportation Research Part*

B: Methodological, vol. 59, pp. 58–80.

Pellegrini, P., Marliere, G., Pesenti, R., and Rodriguez, J., 2015. "RECIFE-MILP: An Effective MILP-Based Heuristic for the Real-Time Railway Traffic Management Problem". *IEEE Transactions on Intelligent Transportation Systems*, vol. 16:5, pp. 2609–2619.

Samà, M., D'Ariano, A., Corman, F., and Pacciarelli, D., "A variable neighbourhood search for fast train scheduling and routing during disturbed railway traffic situations". *Computers & Operations Research*, vol. 78, pp. 480–499.

Sobieraj, S., Marlière, G., and Rodriguez, J., 2011. "Simulation of solutions of a fixed-speed model for the real-time railway traffic optimization problem". In: *Proceedings of The 4th International Seminar on Railway Operations Modelling and Analysis (RailRome2011)*, Rome, Italy.

Törnquist, J., and Persson, J., 2007. "N-tracked railway traffic re-scheduling during disturbances". *Transportation Research Part B*, vol. 41, pp. 342–362.

Statistical Modeling of the Distribution Characteristics of High-Speed Railway Disruptions

Ping Huang [a,b,c,1]

[a] National Engineering Laboratory of Integrated Transportation Big Data Application Technology, Southwest Jiaotong University, Chengdu Sichuan 610031, China;

[b] National United Engineering Laboratory of Integrated and Intelligent Transportation, Southwest Jiaotong University, Chengdu Sichuan 610031, China;

[c] Railway Research Centre, University of Waterloo, Waterloo N2L3G1, Canada

[1] E-mail: huangping129@my.swjtu.edu.cn, +86 18200298902

Abstract

Studies on the spatiotemporal distribution and duration characteristics of railway disruptions are very significant for the advanced prediction of disruption and development of real-time dispatch strategies. In this study, historical disruption records of some Chinese High-Speed Railways (HSRs) lines from 2014–2016 were used to investigate the distribution characteristics of railway disruptions. The spatiotemporal probability distribution of four railway lines were calculated and their hotspots (coordinates with high probabilities) and coldspots (coordinates with low probabilities) were revealed using heatmaps. Furthermore, all the disruptions were classified into seven clusters based on their causes, and statistical analysis was carried out on each cluster. In addition, three right-skewed distribution models, namely Log-normal, Weibull, and Gamma distributions, were used to fit the duration of each cluster to uncover its duration regularities. Finally, goodness-of-fit test was performed on the models using the Kolmogorov-Smirnov method, indicating that the duration of each classified disruption can be estimated using a Log-normal distribution function. The obtained spatiotemporal probabilities and duration time distribution models thus can be further applied into estimating the occurrence and duration of railway disruption in real-time dispatching to help dispatchers make advanced decisions.

Keywords

High-speed-railway; disruption; spatial-temporal distribution; duration; Log-normal distribution

1 Introduction

The disruptions encountered in railway systems are caused by humans, equipment, and the environment, which can lead to considerable losses for managers and travelers. For example, the statistics from a Dutch railway network show that infrastructure-related disruptions occur approximately 22 times per day and each disruption lasts for an average of 1.7 h (Jespersen-Groth et al, 2009). Furthermore, the Austrian Federal Railways suffer huge financial losses of more than EUR 100 million every year due to flooding (Kellermann, Schönberger and Thieken, 2016). Meanwhile, the average departure punctuality in China at various origin stations was 98.8% in 2016. However, the average punctuality at the final destination stations was less than 90% due to various disruptions during operation, although delays smaller than 5 min are considered punctual (Lessan et al, 2018). Hence, train dispatchers are faced with the challenge of reducing the influence of disruptions by developing effective strategies in advance. In other words, the dispatchers can make effective decisions during or before disruptions for efficient timetable re-scheduling if they

can predict when and where the disruptions would occur and how long the disruptions would last. Therefore, studies on the rules and distribution characteristics of railway disruptions are significant for the real-time dispatch of trains.

However, there are several challenges in the accurate prediction of the occurrence of train disruption and duration which are as follows: 1) the disruption is unexpected; and 2) the maintenance duration is highly dependent on the experience and skill of the maintenance staff. Functional models are not sufficient to explain the complex relationship between the disruptions and their potential influence factors. However, skilled dispatchers usually predict the disruption duration empirically, which tends to cause ineffective dispatching when disruptions happen. However, data-mining approaches have recently gained more attention because they can efficiently model train operations and can support robust timetables and real-time dispatching (Wallander and Mäkitalo, 2012). Historical disruption records are considered as interactive consequences of all potential influence factors such that the disruption rules can be determined from the historical performances rather than influence factors. Thus, advanced data-mining techniques, as well as big data, enable us to address these problems using data analysis.

This paper aims to discover the spatiotemporal distribution and duration characteristics of railway disruptions based on data obtained from Guangzhou Railway Group in China. Thus, the spatiotemporal probability distribution of disruptions on four railway lines (Wuhan-Guangzhou HSR line, Shanghai-Shenzhen HSR line, Guangzhou-Shenzhen HSR line, and Guangzhou-Shenzhen intercity line) were analyzed. The disruptions were then classified into seven categories based on their source, and statistical analyses were conducted on the duration of each category. Furthermore, three right-skewed distribution models were used to fit the duration of disruption of each category. The histograms indicated that the duration has a right-skewed and heavy-tailed distribution. Finally, Kolmogorov-Smirnov method was used to perform the goodness-of-fit test in order to select the optimal models for each category.

2 Literature review

Generally, railway disruptions can be caused by exogenous factors, such as natural disasters, and bad weather conditions and endogenous factors, such as operation interference resulted from equipment failure, man-made faults, railway construction, temporary speed limitations, defective braking systems, signal and interlocking failures, and excessive passenger demand (Olsson and Haugland, 2004; Hartrumpf et al, 2009; Higgins, Kozan and Ferreira 1995). Many methods and models have been suggested to manage these disruptions. Traditionally, train operation simulated systems such as LUKS (Janecek and Weymann, 2010), RailSys (Wiklund, 2003), and OpenTrack (Nash and Huerlimann, 2004) have been used by railway researchers and managers worldwide. However, the disruption or delay parameters in these systems mainly depend on hypothetical and theoretical models. (Corman, D'Ariano and Hansen 2014) examined the resisting disturbance abilities of normal traffic and robust timetables using a simulation method. (Huisman and Boucherie, 2001) established a delay propagation model considering the routes occupation relations to predict the knock-on delays, under the condition that train delays follow an Exponential distribution.

Data-driven approaches are also widely used in railway disruption/delay management. These approaches aim to discover the delay and disruption patterns from historical train operation data or disruption records. (Murali et al, 2010) introduced a delay regression-based estimation technique that models delay as a function of train mix and network

topology. (Kecman and Goverde, 2015) developed separate predictive models for the estimation of running and dwell times by collecting data on the respective process types from a training set. (Lessan et al, 2018) examined different distribution models for running times of individual sections in an HSR system and showed that the Log-logistic probability density function is the best distributional form to approximate the empirical distribution of running times on the specified line. It was shown that the distributional form of primary delays, and the affected number of trains could be well-approximated by Log-normal distribution and linear regression models (Wen et al, 2017). A *q*-exponential function is used to demonstrate the distribution of train delays on the British railway network (Takimoto, 2000). Using spatial and temporal resolution transport data from the UK road and rail networks, and the intense storms of 28 June 2012 as a case study, a novel exploration of the impacts of an extreme event has been carried out in (Hartrumpf et al, 2009). Regression trees were trained using Hong Kong subway incident data to estimate the affected delay trains in (Weng et al, 2015). However, the environment of HSR trains is more complex than subway systems. Copula Bayesian networks were developed to predict the duration of turnout faults (Zilko, Kurowicka and Goverde, 2016). A hybrid Bayesian network model is also established to predict arrival and departure delays for Wuhan-Guangzhou HSR (Lessan, Fu and Wen, 2018).

3 Data description

The data used in this study were obtained from the disruption records of Guangzhou Railway Group from 2014-2016, for Wuhan-Guangzhou, Shanghai-Shenzhen, and Guangzhou-Shenzhen HSR lines, as well as Guangzhou-Shenzhen intercity line, as shown in Figure 1. The trains have a maximum speed of 350 km/h when operated on Wuhan-Guangzhou and Guangzhou-Shenzhen HSR lines and 250 km/h when operated on Shanghai-Shenzhen HSR line. In addition, the trains have a maximum speed of 200 km/h when operated on Guangzhou-Shenzhen intercity line. Thus, 2,256 disruptions attributed to nine causes were recorded which are Automatic Train Protection (ATP) system faults, turnout faults, track faults, pantograph faults, rolling stock faults, catenary faults, signal system faults, foreign body invasions, and severe weather. Table 1 shows four cases of the disruptions in the database.

Table 1: Records of HSR disruptions

Line	Date	Train	Time	Duration(min)	Cause
Wuhan-Guangzhou HSR	2014.05.19	G275	19:10	19	Catenary faults
Wuhan-Guangzhou HSR	2014.05.20	G6313	14:30	63	Severe weather
Wuhan-Guangzhou HSR	2015.09.27	G1133	17:06	15	Severe weather
Shanghai-Shenzhen HSR	2015.10.24	G530	16:42	19	Pantograph faults

Figure 1: Sketch map of HSR lines in the jurisdiction of Guangzhou Railway Group.

4 Spatiotemporal probability distribution of disruptions

Railway disruptions are unexpected. However, they tend to appear as regularities that can be investigated from large-scale historical records due to the influence of external factors such as weather and climate, and internal factors such as the characteristics and coordination of equipment, and train interval. Figures 1–4 show the spatiotemporal probability distributions of HSR disruptions, where darker colors represent higher probabilities. Owing to the low probabilities and frequencies of disruptions, each HSR line was divided into several segments to improve the statistical effects. For example, Wuhan-Guangzhou HSR line which has 17 stations was divided into four segments from south-north, such as GZS-SG, SG-HYE, HYE-CSS, and CSS-WH. Figures 1–4 indicate that different segments have different probabilities in the time domain. The peak hours occurred between 12:00 and 20:00. However, GZS-SG segment has the highest probabilities for Wuhan-Guangzhou HSR line, while SZN-SW and CS-ZA segments have higher probabilities for Shanghai-Shenzhen HSR line. Similarly, GZS-HM has the highest probabilities for Guangzhou-Shenzhen HSR line, while GZ-DG and ZMT-SZ have higher probabilities for Guangzhou-Shenzhen intercity line. The spatiotemporal characteristics of disruptions indicate that the probabilities of the disruptions depend on the number of train operations in time domain. However, its influence factors are complex in space domain owing to weak regularities. The probabilities in the space domain tend to be influenced by the status of the equipment, skill

and experience of dispatchers, weather, and climate. However, these factors are different for different locations.

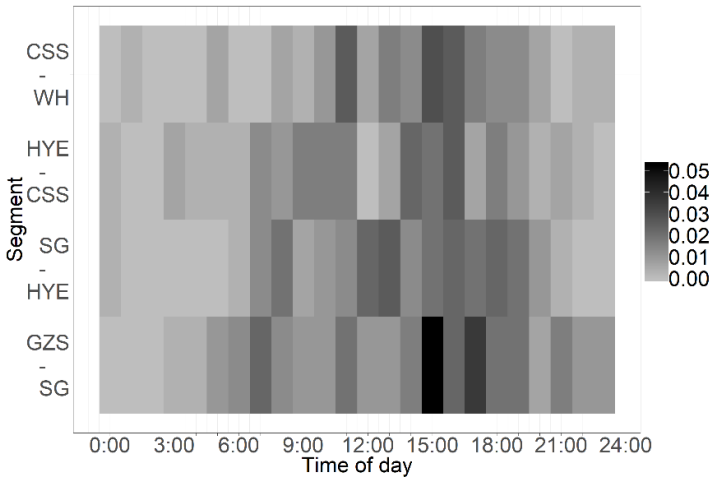


Figure 2: Spatial-temporal distribution of Wuhan-Guangzhou HSR disruptions.

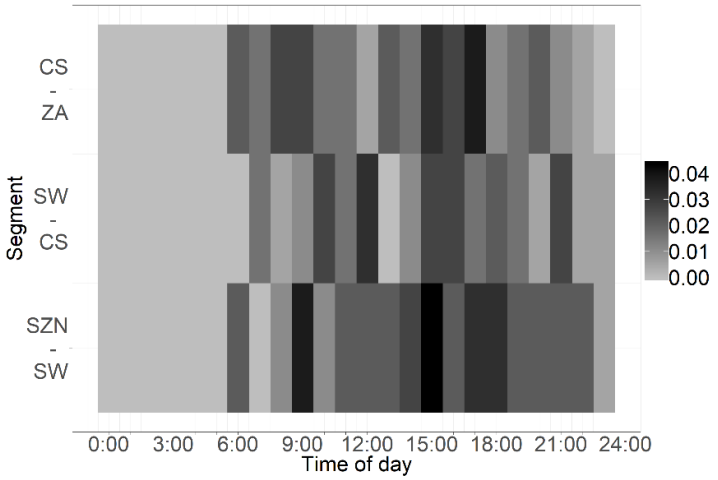


Figure 3: Spatial-temporal distribution of Shanghai-Shenzhen HSR disruptions.

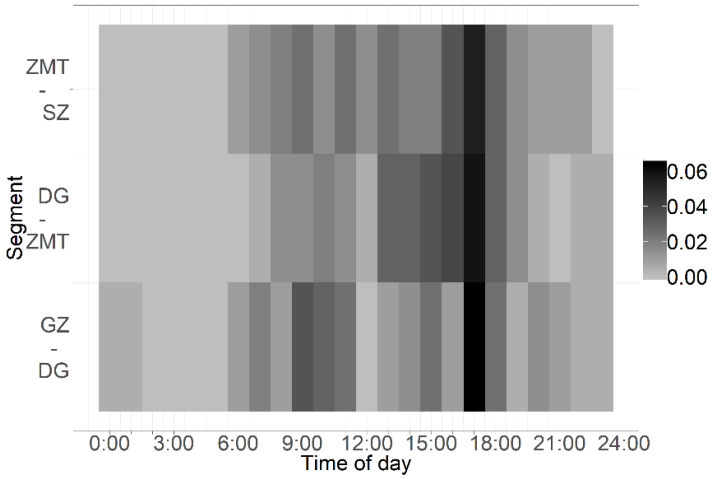


Figure 4: Spatial-temporal distribution of Guangzhou-Shenzhen Intercity Railway disruptions.

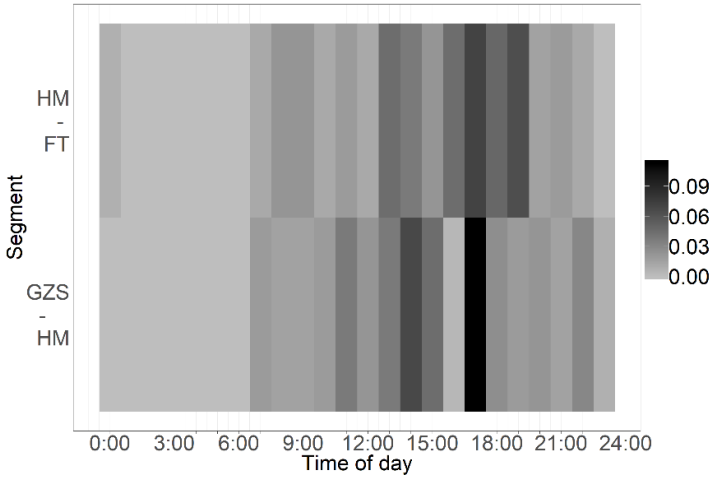


Figure 5: Spatial-temporal distribution of Guangzhou-Shenzhen HSR disruptions.

5 Investigation of disruption duration characteristics

The spatiotemporal distribution probabilities can help dispatchers predict the occurrence of disruptions. However, in practice, it is also necessary to know the duration of disruptions to better understand the characteristics of disruptions, and estimate their influences on train operation, as the durations of disruptions can have different influence on railway systems. Therefore, in this section, we examine the rules of disruption durations using statistical

method.

5.1 Statistics analyses

Based on the coordinated relationship between each equipment in HSR systems, pantograph faults and catenary faults can be regarded as a single category called power supply faults as they have the same effect on HSR systems. Likewise, track faults and turnout faults can be regarded as a single category called turnout-track faults. Thus, the disruptions were classified into seven clusters, namely ATP faults (ATPFs), rolling stock faults (RSFs), turnout-track faults (TTFs), power supply faults (PSFs), signal faults (SFs), severe weather (SW), and foreign body intrusions (FBIs). Statistical analyses were conducted to examine the differences in duration between each category, as shown in Table 2. The results show that the mean values of TTF and SW durations are higher than other values and are longer than 40 min, which indicates that these two categories have stronger influence on the HSR system. In addition, the variances of these two categories are larger than the other values, indicating that a larger uncertainty exists. Meanwhile, the mean and variance of ATPF duration have the least values, indicating that ATPF has the least influence on the HSR system. Its duration has a more centralized distribution.

Table 2: Statistics on duration time of disruptions with different causes(min).

Cause	Min	Mean	Max	Variance	Sample size
RSF	13	31.69	506	1148.22	472
ATPF	10	20.68	154	327.16	328
TTF	8	42.71	579	3185.21	149
PSF	9	33.21	295	1340.07	543
SW	4	41.97	286	2612.29	263
FBI	11	34.35	372	1336.38	289
SF	6	30.19	376	977.90	212
Total	4	27.39	577	1568.39	2256

5.2 Distributional models for disruption duration

The duration of disruption is the difference between its starting and ending time. Figure 6 shows a real disturbance in YDW-SG section on W-G HSR line. This figure defines the disruption length, which is from the time when the station/section is blocked to the time when the first train is allowed to pass. Longer durations can lead to stronger influence on the HSR system, causing more damage and significant losses to railway managers and travelers. Hence, the duration distribution models of the disruptions were investigated to discover the characteristics of disruptions so that dispatchers can predict and control the disruptions effectively. The database just recorded the disruptions whose length are longer than 4 minutes, because the delays longer than 4 minutes are labelled as delayed trains by the China Railway corporation. In addition, samples with durations longer than 120 min were regarded as outliers because they had extremely low frequencies. In Figure 7, the histograms show the duration distribution of each category and all samples, which indicate that both each cluster and all samples have a long-tailed and right-skewed distribution. To quantitatively examine their duration, three right-skewed probability models were selected to fit the data:

1) Log-normal distribution.
If the logarithm of a random variable follows a normal distribution, the random variable also follows a Log-normal distribution. The probability density of a Log-normal model is

$$f\left(x;\mu,\sigma\right)=\frac{1}{x\sigma\sqrt{2\pi}}\exp\left(-\frac{\left(\ln x-\mu\right)^2}{2\sigma^2}\right) \tag{1}$$

where x is a random variable, σ is the standard deviation, and μ is mean.

2) Weibull distribution.

$$f\left(x;\lambda,k\right)=\begin{cases}\frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1}\exp\left(-\left(x/\lambda\right)^k\right) & x\geq 0\\ 0 & x<0\end{cases} \tag{2}$$

where x is a random variable, λ>0 is the scale parameter, and k>0 is the shape parameter.

3) Gamma distribution.

$$f\left(x;\alpha,\beta\right)=\frac{1}{\beta^{\alpha}\Gamma\left(\alpha\right)}x^{\alpha-1}\exp\left(-\frac{x}{\beta}\right)\qquad x>0 \tag{3}$$

$$\Gamma\left(x\right)=\int_0^{\infty}t^{x-1}\exp\left(-t\right)dt \tag{4}$$

where x is a random variable, α is the shape parameter, and β is the scale parameter.

The models above were used to fit the duration of the disruptions as shown in Figure 7. Meanwhile, the fitted parameters of each category are shown in Table 3.

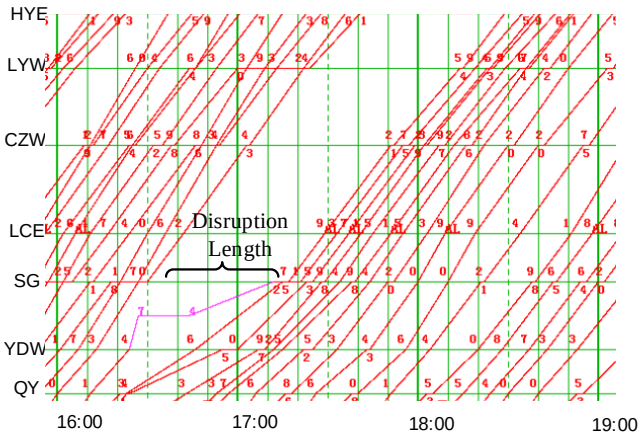


Figure 6: A real disturbance happened on W-G HSR line shown in time-space diagram (horizontal axis is time, and vertical axis is space).

Table 3: Fitted parameters of each category.

Cause	Log-normal		Weibull		Gamma	
	μ	σ	k	λ	α	β
RSF	3.100	0.772	1.416	32.418	1.954	0.066
ATPF	2.760	0.685	1.480	22.186	2.328	0.117
TTF	3.301	0.707	1.553	38.491	2.272	0.066
PSF	3.082	0.727	1.407	31.290	2.091	0.074
SW	3.091	0.780	1.304	32.774	1.761	0.058
FBI	3.138	0.733	1.434	33.230	2.067	0.069
SF	2.768	0.845	1.323	23.836	1.739	0.079
Total	3.029	0.766	1.376	30.211	1.933	0.070

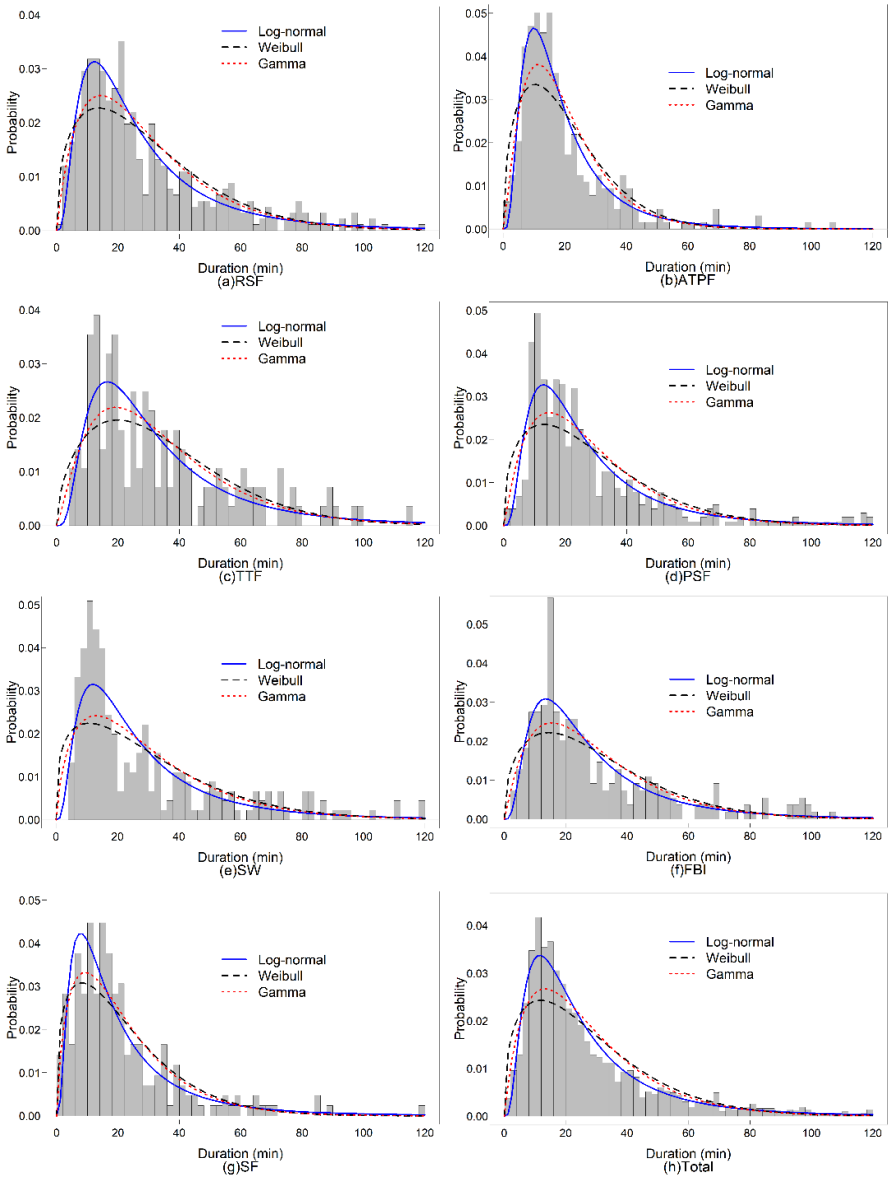


Figure 7: Fitting results of duration time of each disruption category.

5.3 Goodness-of-fit testing

To select the model that has the best performance for each category, a Kolmogorov-Smirnov (K-S) (Massey Jr, 1951) method was used to test the goodness-of-fit of the models. K-S method tests if one random variable follows a theoretical distribution, or if two random variables have the same distribution. Its null hypothesis is as follows:

H0：a random variable follows a theoretical distribution, or two random variables have the same distribution.

Its test statistic (*T*) is the largest difference between the cumulative distribution function (CDF) of the data and the theoretical distribution, as described by (5). However, some random numbers, which follow an uniform distribution were added to the data in order to satisfy the continuity requirement of K-S because the historical train operation data were recorded in the minute timescale

$$T = max\left|F'(x) - F(x)\right| \tag{5}$$

where $F'(x)$ is the CDF of the observed data, which consists of the duration of each category, $F(x)$ is the CDF of the theoretical distribution models, which consists of three alternative distribution models. A significance level of $\alpha = 0.05$ was chosen for the test. As *T* becomes smaller, the sample distribution tends to follow the theoretical distribution. The K-S test results of all the models are summarized in Table 4.

The results indicate that Log-normal models fitted using RSF, ATPF, FBI, all samples, as well as all alternative models fitted using TTF, and SF samples, passed K-S test. However, the Log-normal models had the least *T* value. Meanwhile, no model based on PSF and SW samples passed K-S test. However, Log-normal model had the least *T* value, and the p-values were very close to α. Therefore, the CDF of Log-normal model had the smallest distance with that of PSF, and SW, and Log-normal model thus was chosen as the distribution model of all HSR disruption clusters. The parameters of each category are shown in Table 5. The fitted probability models can be used to estimate the duration of any disruption, once its causes are ascertained.

Table 4: KS testing result of each cluster.

Cause	Log-normal		Weibull		Gamma	
	T	p-value	T	p-value	T	p-value
RSF	0.028	0.863	0.078	0.007	0.067	0.031
ATPF	0.041	0.635	0.099	0.003	0.086	0.016
TTF	0.052	0.848	0.072	0.450	0.073	0.436
PSF	0.066	0.021	0.109	0.000	0.083	0.001
SW	0.094	0.034	0.119	0.003	0.129	0.000
FBI	0.037	0.843	0.093	0.017	0.081	0.052
SF	0.065	0.324	0.083	0.104	0.071	0.224
Total	0.031	0.729	0.077	0.001	0.073	0.016

Note: underline fonts mean passing K-S test

Table 5: Fitted Log-normal distribution parameters for each category.

Cause	Model	μ	σ	Cause	Model	μ	σ
RSF	Log-normal	3.100	0.772	SW	Log-normal	3.091	0.780
ATPF	Log-normal	2.760	0.685	FBI	Log-normal	3.138	0.733
TTF	Log-normal	3.301	0.707	SF	Log-normal	2.768	0.845
PSF	Log-normal	3.082	0.727	Total	Log-normal	3.029	0.766

6 Conclusion

In this paper, we investigated the spatiotemporal distribution and duration distribution characteristics of railway disruptions based on the historical disruption records of four HSR lines in China. The conclusions made are as follows:

1) The probabilities of railway disruptions are spatiotemporally different.
2) Railway disruptions can be classified into seven categories based on their causes and influence on the HSR system.
3) The statistical analyses of each category revealed that the average duration of TTF and SW is the highest and longer than 40 min, whereas ATPF has the least value.
4) The duration of each category can be well fitted using Log-normal distribution model.

The results can assist dispatchers in understanding the distribution characteristics of disruptions, thereby improving the quality of their decisions. In particular, they can obtain the real-time and future probabilities of disruption at any coordinates of the timetable to enable them develop strategies that can prevent the disruptions. Furthermore, they can estimate the duration of disruptions using fitted Log-normal distribution models in order to make better decisions. The probability models can also improve train operations and disruption management in simulated systems as they are more accurate than hypothetical models. Hypothetical models introduce certain gaps into the simulations and usually overestimate or ignore some situations and constraints of train operations, which are needed by dispatchers in rescheduling the timetable.

Acknowledgments

This work was supported by the China Scholarship Council [grant number 201707000038]; National Nature Science Foundation of China [grant numbers 71871188, 61503311]; Science & Technology Department of Sichuan Province [grant number 2018JY0567]; and the Doctoral Innovation Fund Program of Southwest Jiaotong University [grant number D-CX201827]. We are grateful for the useful contributions made by our project partners, and we would like to thank the China Railway Guangzhou Group Co., Ltd for the data support.

References

Corman, Francesco, Andrea D'Ariano, and Ingo A Hansen. 2014. "Evaluating disturbance robustness of railway schedules." *Journal of Intelligent Transportation Systems* 18 (1):106-120.

Hartrumpf, M., T. Claus, M. Erb, and J. M. Albes. 2009. "Surgeon performance index: tool for assessment of individual surgical quality in total quality management."

European Journal Of Cardio-Thoracic Surgery 35 (5):751-758. doi: 10.1016/j.ejcts.2008.12.006.

Higgins, Andrew, E Kozan, and L Ferreira. 1995. "Modelling delay risks associated with train schedules." *Transportation Planning and Technology* 19 (2):89-108.

Huisman, Tijs, and Richard J. Boucherie. 2001. "Running times on railway sections with heterogeneous train traffic." *Transportation Research Part B: Methodological* 35 (3):271-292.

Janecek, David, and Frédéric Weymann. 2010. "LUKS-Analysis of lines and junctions." Proceedings of the 12th World Conference on Transport Research (WCTR).

Jespersen-Groth, Julie, Daniel Potthoff, Jens Clausen, Dennis Huisman, Leo Kroon, Gábor Maróti, and Morten Nyhave Nielsen. 2009. "Disruption management in passenger railway transportation." In *Robust and online large-scale optimization*, 399-421. Springer.

Kecman, Pavle, and Rob MP Goverde. 2015. "Predictive modelling of running and dwell times in railway traffic." *Public Transport* 7 (3):295-319.

Kellermann, Patric, Christine Schönberger, and Annegret H Thieken. 2016. "Large-scale application of the flood damage model RAilway Infrastructure Loss (RAIL)." *Natural Hazards and Earth System Sciences* 16 (11):2357-2371.

Lessan, Javad, Liping Fu, and Chao Wen. 2018. "A hybrid Bayesian network model for predicting delays in train operations." *Computers & Industrial Engineering*. In press, DOI:10.1016/j.cie.2018.03.017.

Lessan, Javad, Liping Fu, Chao Wen, Ping Huang, and Chaozhe Jiang. 2018. "Stochastic Model of Train Running Time and Arrival Delay: A Case Study of Wuhan–Guangzhou High-Speed Rail." *Transportation Research Record*. DOI:10.1177/0361198118780830.

Massey Jr, Frank J. 1951. "The Kolmogorov-Smirnov test for goodness of fit." *Journal of the American statistical Association* 46 (253):68-78.

Murali, Pavankumar, Maged Dessouky, Fernando Ordóñez, and Kurt Palmer. 2010. "A delay estimation technique for single and double-track railroads." *Transportation Research Part E: Logistics and Transportation Review* 46 (4):483-495. DOI: 10.1016/j.tre.2009.04.016.

Nash, Andrew, and Daniel Huerlimann. 2004. "Railroad simulation using OpenTrack." *WIT Transactions on The Built Environment* 74.

Olsson, Nils OE, and Hans Haugland. 2004. "Influencing factors on train punctuality—results from some Norwegian studies." *Transport policy* 11 (4):387-397.

Takimoto, T. 2000. "Development of efficient operational control using object representation." *Computers In Railways Vii* 7:837-841.

Wallander, Jouni, and Miika Mäkitalo. 2012. "Data mining in rail transport delay chain analysis." *International Journal of Shipping and Transport Logistics* 4 (3):269-285.

Wen, Chao, Zhongcan Li, Javad Lessan, Liping Fu, Ping Huang, and Chaozhe Jiang. 2017. "Statistical investigation on train primary delay based on real records: evidence from Wuhan–Guangzhou HSR." *International Journal of Rail Transportation* 5 (3):170-189. DOI:10.1080/23248378.2017.1307144.

Weng, Jinxian, Yang Zheng, Xiaobo Qu, and Xuedong Yan. 2015. "Development of a maximum likelihood regression tree-based model for predicting subway incident delay." *Transportation Research Part C: Emerging Technologies* 57:30-41.

Wiklund, Mats. 2003. "SERIOUS BREAKDOWNS IN THE TRACK INFRASTRUCTURE: CALCULATION OF EFFECTS ON RAIL TRAFFIC." *VTI MEDDELANDE* (959).

Zilko, Aurelius A, Dorota Kurowicka, and Rob MP Goverde. 2016. "Modeling railway disruption lengths with Copula Bayesian Networks." *Transportation Research Part C: Emerging Technologies* 68:350-368.

Mining Train Delay Propagation Pattern from Train Operation Records in a High-Speed System

Ping Huang [a,b,c,1], Chao Wen [a,b,c] Zhongcan Li [a,b]
[a] National Engineering Laboratory of Integrated Transportation Big Data Application Technology, Southwest Jiaotong University, Chengdu Sichuan 610031, China;
[b] National United Engineering Laboratory of Integrated and Intelligent Transportation, Southwest Jiaotong University, Chengdu Sichuan 610031, China;
[c] Railway Research Centre, University of Waterloo, Waterloo N2L3G1, Canada
[1] E-mail: huangping129@my.swjtu.edu.cn, +86 18200298902

Abstract
This study aims to investigate delays, delay increases, and delay recovery characteristics, by using statistical methods to clarify delay propagation patterns according to historical records of the Wuhan-Guangzhou high-speed railway (HSR) in China in 2014 and 2015. Specifically, we examined arrival and departure delay duration distributions and used heatmaps to demonstrate the spatiotemporal frequency distribution of delays, delay increases, and delay recovery, and the heatmaps clearly show hot spots (coordinates with high frequencies) in a timetable. Then, we separated delays as discrete intervals according to their severity, and analyzed the delay increasing frequency and the delay increasing severity within each interval, so as to clarify the relationships of delay increasing probability and delay increasing severity with delay extents. Next, we investigated the observed delay recoveries and prescheduled buffer times at (in) station (section), which demonstrate the recovery ability of each station and section. Finally, to understand the key influencing factor of delay propagation, we analyzed the relationship between capacity utilization and delays, delay increases, and delay recoveries, by examining their Pearson correlation coefficients. These indicate that delay frequencies and delay increasing frequencies with Pearson correlation coefficients as high as 0.9 are highly dependent on capacity utilization. The uncovered delay propagation patterns can enrich dispatchers' experience, and improve their decision-making ability during real-time dispatching in HSR.

Keywords
High-speed railway, train delays, delay increases, delay recoveries, capacity utilization

1 Introduction

Train operations are subject to various disturbances, such as severe weather, power outages, and facility failure, and all of these can result in train delays and lead to considerable losses for both railway operators and travellers (Khadilkar, 2016). For instance, the statistics from a Dutch railway network show that infrastructure-related disruptions occur approximately 22 times per day, and each disruption on average can last 1.7 hours (Jespersen-Groth et al. 2009). The Austrian Federal Railways had to cope with financial losses of more than EUR 100 million every year due to flooding (Kellermann, Schönberger and Thieken, 2016). In China, the average departure punctuality in origination stations was as high as 98.8% in 2016, but because of various disturbances during their operations, the average punctuality at the final destination stations was less than 90%, though delays smaller than 5 minutes are

considered punctual (Lessan et al, 2018). For the train dispatchers, the key steps to reduce loss are not only managing the unexpected delays, but also making decisions in advance. In other words, if the dispatchers can know the delay probabilities at different times and locations and how the delay would evolve and propagate, they can make decisions before train delays that can result in more efficient timetable re-scheduling. Therefore, examining patterns of delay propagation is of great significance for improving railway delay management and real-time decision-making abilities.

However, accurate train delay and propagation pattern recognition presents challenges, mainly because: 1) disturbances are totally unexpected; 2) delay propagation is spatiotemporal, and its influencing factors are complex; and 3) prescheduled time supplements and buffer times cannot be fully utilized (delay recoveries are stochastic). In practice, some skilled dispatchers usually predict delays and delay propagation empirically, leading to different decision-making standards even for the same dispatcher. Data-mining approaches have recently gained more attention, due to their better understanding of train delay concatenation and the fact they are more supportive of robust timetables and real-time dispatching (Wallander and Mäkitalo, 2012). Historical train operation records can be regarded as the interactive consequences of all potential influencing factors, and this supports us exploring the rules of delays and propagation from their historical performances, rather than fitting functional expressions. Hence, mining the delay and propagation patterns from historical operation records can provide more accurate and comprehensive results for railway operation managers.

This study aims to recognize train delay and delay propagation patterns from train operation data of the Wuhan-Guangzhou (W-G) HSR in China. To this end, we first analyzed the duration and spatiotemporal distribution of train delays. Next, we split train delays as discrete intervals with a width of 5 minutes according their length, and investigated the delay increasing probabilities and severity on different delay extents. We also examined the delay recovery abilities and prescheduled buffer times at(in) each station(section). Finally, in order to understand the influences of capacity utilization on delays, delay increases, and delay recoveries, we investigated their relationships by calculating Pearson correlation coefficients.

2 literature review

Generally, train delays are caused by exogenous factors, such as natural disasters and bad weather conditions, and endogenous factors, such as operation interference resulting from equipment failure, man-made faults, railway construction, temporary speed limitations, defective braking systems, signal and interlocking failures, and excessive passenger demand (Olsson and Haugland, 2004; Hartrumpf et al, 2009; Higgins, Kozan and Ferreira 1995). In addition, if the running and dwell times increase due to unexpected disturbances, it can result in knock-on delays and delays for other trains (Milinković et al, 2013). Serious disruptions such as switch or signal failures, if not managed effectively, can result in queuing of trains, creating a chain of delayed trains. The experience from the Taiwan HSR shows that shortening the maintenance cycle can effectively alleviate the problem of train delay caused by signal failures (Hasan, 2011). Some studies have made contributions on statistical models of delay, and the respective fitness models. The Weibull, Gamma, and Log-normal distributions have been adopted in several studies (Yuan, Goverde and Hansen, 2002; Higgins and Kozan, 1998). It was shown that the distributional form of primary delays and the affected number of trains could be well-approximated by classical methods, such as Log-normal distribution and linear regression models (Wen et al, 2017). A q-

exponential function is used to demonstrate the distribution of train delays on the British railway network (Takimoto, 2000). Using spatial and temporal resolution transport data from the UK road and rail networks, and the intense storms of 28 June 2012 as a case study, a novel exploration of the impacts of an extreme event has been carried out in (Hartrumpf et al, 2009). Given the HSR operation data, the maximum likelihood estimation method was used to determine the probability distribution of the different disturbance factors and the distributions of affected trains. However, the models of primary delay consequences have not been established in detail (Xu, Corman and Peng, 2016). Probabilistic distribution functions of both train arrival and departure delays at the individual station were derived in general, based on the data from Beijing-Shanghai HSR (Liang et al, 2009).

Data-driven research studies proposed for delay management mainly focused on using regression or distribution approaches to fit delay data. (Milinković et al, 2013) mined data from peak hours, including rolling-stock and weather data, and developed a predictive model involving the mining of track occupation data for delay estimations. A data-mining approach was used for analyzing rail transport delay chains with data from passenger train traffic on the Finnish rail network, but the data from the train running process was limited to one month (Wallander and Mäkitalo, 2012). (Murali et al, 2010) reported a delay regression-based estimation technique that models delay as a function of train mix and network topology. A statistical analysis of train delays in the Eindhoven Station in the Netherlands was used to explain systematic delay propagation, based on the use of a robust linear regression model to uncover the correlations between arrival delays (Goverde, 2005). Recently, (Kecman and Goverde, 2015) developed separate predictive models for the estimation of running and dwell times by collecting data on the respective process types from a training set. (Lessan et al, 2018) examined different distribution models for running times of individual sections in an HSR system, and showed that the log-logistic probability density function is the best distributional form to approximate the empirical distribution of running times on the specified line. A hybrid Bayesian network model is also established to predict arrival and departure delays for the Wuhan-Guangzhou HSR (Lessan, Fu and Wen, 2018).

3 Data description

The data used in this study are the real-world train operation records of the Wuhan-Guangzhou HSR in China. This line connects to the Guangzhou-Shenzhen HSR line at GZS station, to the Hengyang-Liuzhou HSR line at HYE station, and to the Shanghai-Kunming HSR line at CSS station, respectively. All the trains operating on this line are equipped with the Chinese Train Control System (CTCS), which allows a maximum speed of 350 km/h, and the Automatic Train Supervision system that records the movements of all trains. We considered data from trains in the northbound direction comprising the segment from Guangzhou South (GZS) to Changsha South (CSS), as shown in Figure 1. The collected data contain 57,796 HSR trains in the GZS-HYE section and 64,547 HSR trains in the HYE-CSS segment, comprising information about train operations from March 24, 2015 to November 10, 2016. The scheduled/actual arrival/departure records of each train and station, the number of trains, dates, occupied tracks, and section lengths were collected to construct a database with data recorded every minute. Figure 2 shows the accumulative HSR trains of each station during each hour, clearly indicating the differences of train services along with space and time axes. Therefore, in the following sections, we will investigate the spatiotemporal differences of delay and delay propagation characteristics.

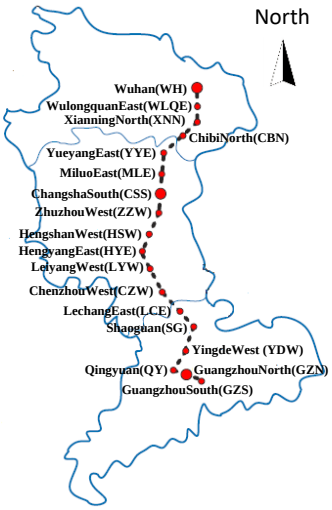


Figure 1: Map of Wuhan–Guangzhou high-speed railway line

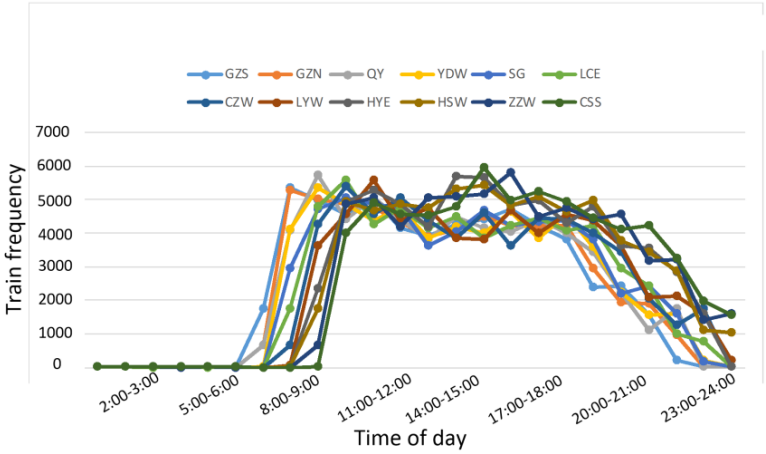


Figure 2: Cumulative HSR trains per hour

4 Delay Characteristics Investigation

Longer delays can have stronger influences on railway systems and can propagate farther, whereas shorter delays have smaller influences on railway systems, or can even be assimilated at the moment of occurrence. To understand their characteristics, we first

visualized the duration distribution of arrival and departure delays. The histograms in Figure 3 clearly show that both arrival and departure delays follow a right-skewed and heavy-tailed distribution, which indicates that the longer the delays, the lower the frequencies. Also, train delays can propagate along time and space axes, which can result in different delay frequencies in a timetable. To understand the spatiotemporal distribution pattern of train delays, we analyzed the frequency distribution of delays. We separated delays as longer than 4 minutes and as longer than 30 minutes, to better understand the spatiotemporal distribution characteristics of different delay severities. The length of 4 minutes was chosen because it is the criteria set by the Chinese Railway Company to label trains as delayed, and the length of 30 minutes was chosen to understand the spatiotemporal distribution of longer delays. Figure 4 and Figure 5 clearly show that both 4 minute and 30-minutes-or-longer delay frequencies at the original station are extremely low, but they became much higher with the operation of trains. In addition, along the time axis, their frequencies are low during off-peak hours, and high during peak hours. In short, the hot spots of both 4 minute and 30-minutes-or-longer delays appear during 14:00 to 20:00, in the LCE-CSS section.

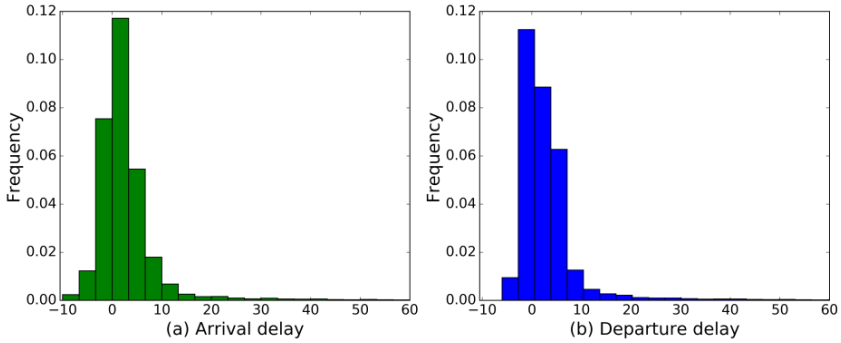


Figure 3: Delay length (min) distribution

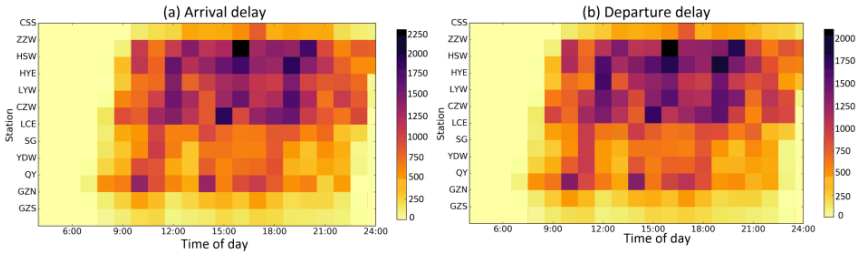


Figure 4: Spatiotemporal distribution of delays

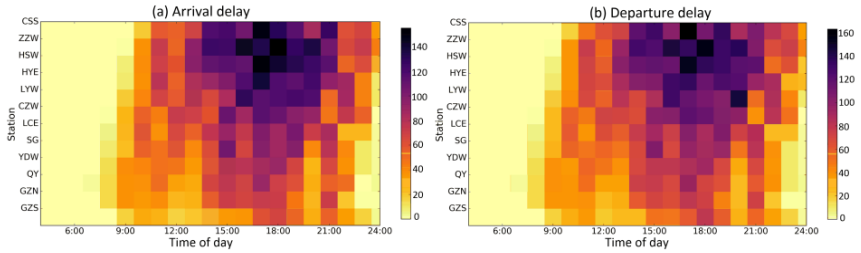


Figure 5: Spatiotemporal distribution of delays longer than 30 minutes

5 Delay propagation pattern investigation

5.1 Delay increasing characteristics

Delays can increase due to secondary disturbances. In order to understand the delay increasing pattern in the timetable, we first made statistics about the delay increasing (growth more than 4 minutes) frequency, at station and in section. In this process, a train whose departure delay was 4 minutes longer than its previous arrival delay was labeled as delay increase at station, and an arrival delay that was 4 minutes longer than its previous departure delay was labeled as delay increase in section. Figure 6 clearly shows that the delay increasing frequency at station is high at junction stations (CSS, HYE, and GZS). This conclusion is understandable, as the junction stations have more tasks (such as trains turning-over, crossing-line, and terminating) than other stations, which makes the equipment utilization more frequent, leading to the higher disturbance probability. However, without evident task volume differences, probabilities in sections do not appear with any explicable regularity, as they are mainly related to equipment status, climate, weather, and the experience and skills of dispatchers.

Then, we conducted sensitive analyses of delay increasing frequency and delay increasing severity on different delay severities. We transferred train delays as discrete intervals with a width of 5 minutes, and separated delays into the intervals that they fall in. Likewise, the sensitive analyses were also conducted on delay increases at station and in section, as shown in Figure 7 and Figure 8.

Figure 7 shows that both delay increasing frequencies at station and in section rise with the growth of delay extent, meaning that longer delays are more likely to encounter secondary disturbances. An exception happens on the early-arrival-trains interval (the first interval), where the delay increasing probability at station is abnormally high, but that is not in the case in section. This can be explained by the dispatching principle that trains are only allowed to arrive early, but cannot depart early due to the passenger boarding requirements. Therefore, early arrival trains tend to be given more dwelling times to depart on schedule. Figure 8 shows the sensitivity of delay increasing severity, where, at stations, it is shorter with the growth of delay length, but, in sections, it keeps stable with the growth of delay length. Also, an exception happens on the interval of early arrival trains at station (the first interval). This result was caused by the recovery of early arrival trains, as their early arrival times (the smallest is -10 minutes) are not as long as delays (can reach 190 minutes), thus limiting their increasing extent.

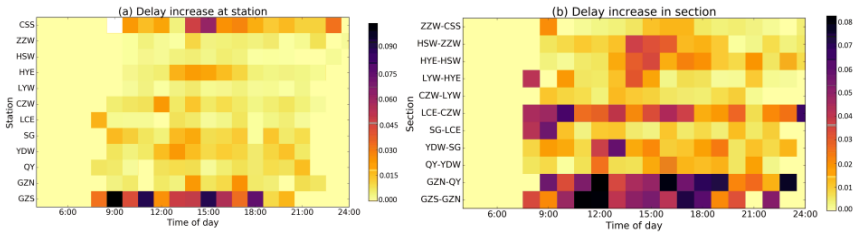


Figure 6: Spatiotemporal frequency distribution of delay increase

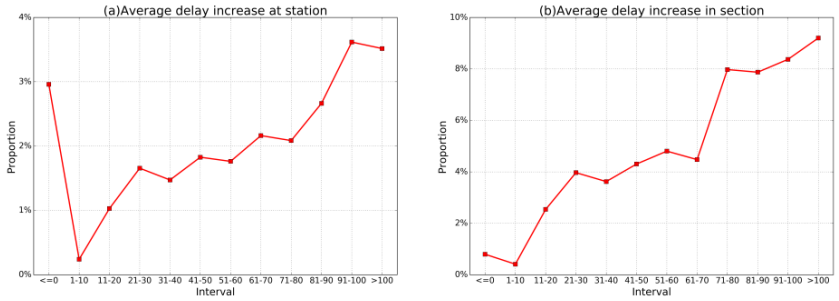


Figure 7: Sensitive analysis of delay increasing frequency on delays severity

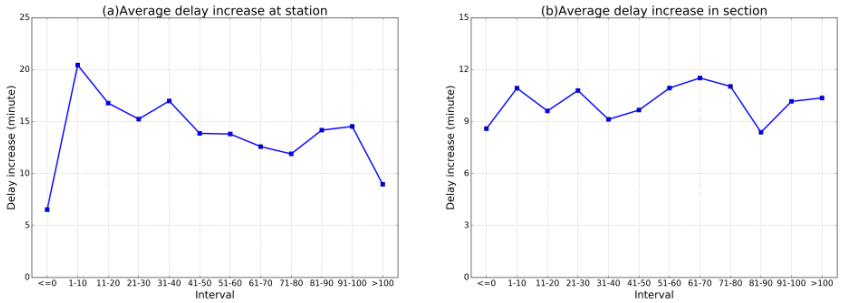


Figure 8: Sensitive analysis of delay increasing severity on delays severity

5.2 Delay recovery characteristics

Delays can be recovered using buffer times prescheduled in sections and stations. To understand the delay recovery characteristics of each station and section, we conducted statistics analyses about the spatiotemporal probability distribution of delay recoveries, which is the proportion of the trains with delay recoveries to all delayed trains, as shown in Figure 9. Like the spatiotemporal distributions of delay increases in sections, delay recoveries do not have centralized hot spots, but their probabilities in section are much higher than those at station. Besides, delay recovery probabilities at(in) one station(section) are evidently different from others, coordinating the empirical conclusion that delay

recoveries are dominantly influenced by the buffer times distribution in the timetable. In practice, the buffer time allocation methods differ from timetables/railway lines, such as allocating according to section length and travel times or according to the specific recovery requirements of sections (Huang et al, 2018). Therefore, we investigated the observed delay recoveries and pre-scheduled buffer times of each station and section as shown in Figure 10 and Figure 11, respectively. Figure 10 denotes the comparisons of scheduled running (dwell) times and practical running (dwell) times at(in) each station(section). Comparisons of the bar pairs indicate that the practical running and dwelling times are smaller than the scheduled running and dwelling times, implying that buffer times were somewhat effective in reducing delays at(in) station(section). However, different stations and sections appear to have different recovery values, as the left-hand bars were ranked from small to large (from top to bottom), but the right-hand bars were opposed to this rule, and the recovery volumes in sections are lower than those at stations. We thus calculated the available (prescheduled) buffer times at each station using prescheduled running times and minimum running times, and those in sections using prescheduled dwell times and minimum dwell times, given by (1) and (2).

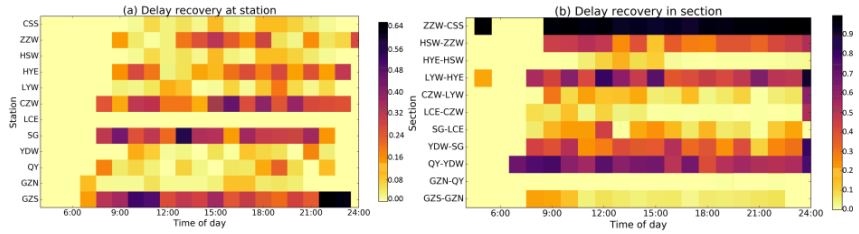


Figure 9: Spatiotemporal distribution of delay recovery probabilities

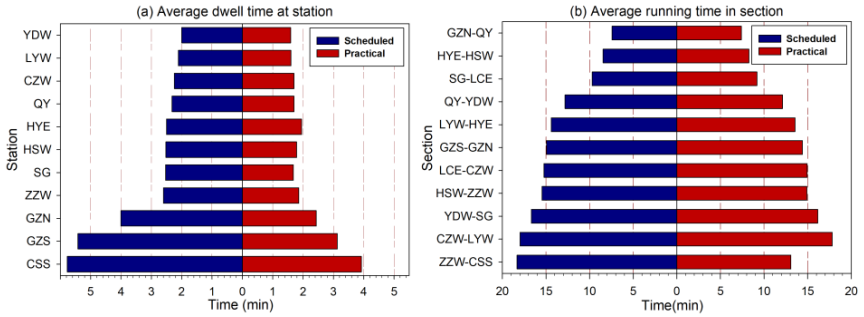


Figure 10: Comparison of scheduled and practical dwell (running) times

$$BT_{\text{station}} = T_{\text{station}} - T_{\min}, \quad (1)$$

$$BT_{\text{section}} = T_{\text{section}} - L/S_{\max}. \quad (2)$$

where BT_{station} and BT_{section} are buffer times at(in) station(section); T_{station} and T_{section} are

prescheduled dwell times and running times at(in) station(section); $T_{\min}$ is the minimum dwell times of trains at station, where, at junction stations, it is 2 minutes, and at other stations, it is 1 minute; L is the distance of every adjacent station; and S_{max} is the maximum speed of HSR trains, i.e., 310 km/h during the time span in the collected data, according to the technique documents from China Railway Company. The available buffer times distributions are shown in Figure 11, where buffer times value is extremely high at the GZS station and in the ZZW-CSS section, which can explain the high probabilities at GZS stations and in the ZZW-CSS section in Figure 9, and the large recovery ability of ZZW-CSS in Figure 10.

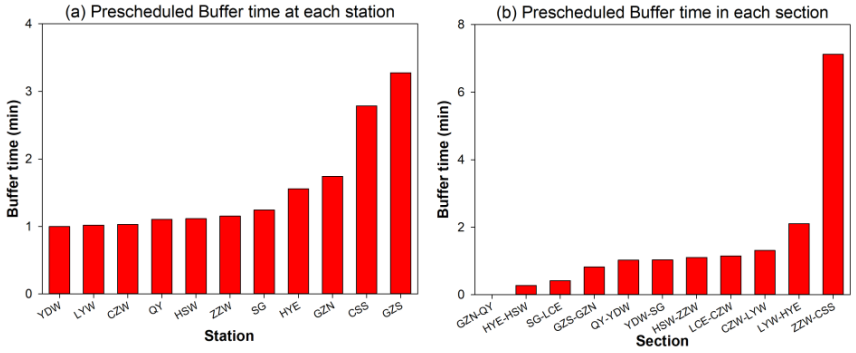


Figure 11: Prescheduled buffer times at(in) each station(section)

5.3 Correlation analyses with capacity utilization

Figure 2, Figure 4, Figure 5, and Figure 6 indicate that both train delays and delay increasing have high frequencies during peak hours. To quantitatively estimate their relationships with capacity utilization, we calculated the Pearson correlation coefficients (φ , see (4)) of delays and delay increases with the number of trains per hour (N), given by (3).

$$N = {N_{total}}/{d} \ . \tag{3}$$

where N_{total} is the total train services of each hour shown in Figure 2, and d is the number of days the data included.

$$\varphi_{X,Y} = \frac{E(XY) - E(X)E(Y)}{\sqrt{E(X^2) - E^2(Y)}\sqrt{E(Y^2) - E^2(X)}} \ . \tag{4}$$

In the above equation, X and Y are two variables, and $E(X)$ and $E(Y)$ are the expectations of X and Y, respectively.

Table 1 clearly shows that delays (including arrival and departure delays) and delay increases (including delay increases at station and in section) have strong relationships with the number of trains per hour whose Pearson correlation coefficients can reach as high as 0.9, but the Pearson correlation coefficients between delay recoveries and the number of trains per hour are as low as 0.413 and 0.598 at station and in section, respectively. Pearson correlation coefficients further clarify the quantitative relationships between the number of trains per hour with delays, delay increases, and delay recoveries. Specifically, the scatter-lines in Figure 12 and Figure 13 show the matching effects of delays and delay increases

with the number of trains per hour. Considering these figures and the Pearson correlation coefficients, the linear relationship between the probabilities of delays and delay increases with the number of trains per hour is high.

Table 1: Pearson correlation coefficients of delays, delay increases, and delay recoveries and capacity utilization

Arrival delays	Departure delays	Delay increase at station	Delay increase in section	Delay recovery at station	Delay recovery in section
0.936	0.933	0.915	0.945	0.413	0.598

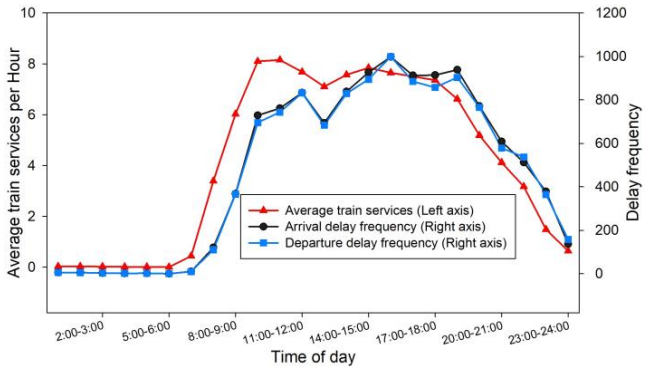


Figure 12: Correlation of delay and capacity utilization

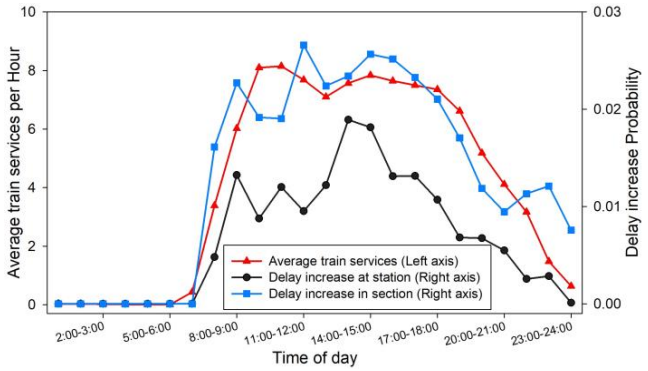


Figure 13: Correlation of delay increase and capacity utilization

6 conclusion

The paper presents how to recognize train delays and delay propagation patterns from historical train operation records. The following conclusions were obtained.

1) Train delay frequencies and delay increasing probabilities are spatiotemporally different, and are highly dependent on capacity utilization; the more the capacity utilization, the higher probabilities of delays and delay increases (with Pearson correlation coefficients over 0.9).

2) For both arrival and departure delays, the longer the delays, the higher the delay increasing probabilities.

3) Longer arrival delays could result in shorter delay increases, whereas departure delays do not influence the delay increasing extent.

4) The delay recoveries, which are mainly influenced by prescheduled buffer times, have higher probabilities but lower volumes in section as compared against those at stations.

The spatiotemporal probabilities and analyses on delay increase and recovery can help dispatchers improve their decision-making qualities. Explicitly, with the spatiotemporal distributions, the dispatchers can obtain the real-time and future probabilities of delays, delay increases, and delay recoveries. With the sensitivity analyses between delays and delay increase, the dispatchers can acquire their increase probabilities and severities under any delay length; with the relationship analyses between delay recoveries and total buffer times, the dispatchers can have a better understanding of the recovery abilities of each station and section. Additionally, the spatiotemporal probabilities can also be applied to train operation simulation systems to optimize disturbance setting and timetable rescheduling programs, as they are more practical than hypothetical models that bring certain gaps between simulations and practice, and usually over assume and ignore some situations and constraints of train operations.

Acknowledgments

This work was supported by the China Scholarship Council [grant number 201707000038]; National Nature Science Foundation of China [grant numbers 71871188, 61503311]; Science & Technology Department of Sichuan Province [grant number 2018JY0567]; and the Doctoral Innovation Fund Program of Southwest Jiaotong University [grant number D-CX201827]. We are grateful for the useful contributions made by our project partners, and we would like to thank the China Railway Guangzhou Group Co., Ltd for the data support.

References

Goverde, Rob MP. 2005. "Punctuality of railway operations and timetable stability analysis." Delft.

Hartrumpf, Martin, Thomas Claus, Michael Erb, and Johannes M Albes. 2009. "Surgeon performance index: tool for assessment of individual surgical quality in total quality management." *European Journal of Cardio-thoracic Surgery* 35 (5):751-758.

Hasan, Nazmul. 2011. "Direct Fixation Fastener (DFF) Spacing and Stiffness Design." 2011 Joint Rail Conference, *American Society of Mechanical Engineers*.

Higgins, Andrew, E Kozan, and L Ferreira. 1995. "Modelling delay risks associated with train schedules." *Transportation Planning and Technology* 19 (2):89-108.
Higgins, Andrew, and Erhan Kozan. 1998. "Modeling train delays in urban networks." *Transportation Science* 32 (4):346-357.
Huang, Ping, Chao Wen, Qiyuan Peng, Javad Lessan, Liping Fu, and Chaozhe Jiang. 2018. "A data-driven time supplements allocation model for train operations on high-speed railways." International Journal of Rail Transportation:1-18.
Jespersen-Groth, Julie, Daniel Potthoff, Jens Clausen, Dennis Huisman, Leo Kroon, Gábor Maróti, and Morten Nyhave Nielsen. 2009. "Disruption management in passenger railway transportation." In *Robust and online large-scale optimization*, 399-421. Springer.
Kecman, Pavle, and Rob MP Goverde. 2015. "Predictive modelling of running and dwell times in railway traffic." *Public Transport* 7 (3):295-319.
Kellermann, Patric, Christine Schönberger, and Annegret H Thieken. 2016. "Large-scale application of the flood damage model RAilway Infrastructure Loss (RAIL)." *Natural Hazards and Earth System Sciences* 16 (11):2357-2371.
Khadilkar, Harshad. 2016. "Data-enabled stochastic modeling for evaluating schedule robustness of railway networks." *Transportation Science* 51 (4):1161-1176.
Lessan, Javad, Liping Fu, and Chao Wen. 2018. "A hybrid Bayesian network model for predicting delays in train operations." *Computers & Industrial Engineering*. DOI:10.1016/j.cie.2018.03.017
Lessan, Javad, Liping Fu, Chao Wen, Ping Huang, and Chaozhe Jiang. 2018. "Stochastic Model of Train Running Time and Arrival Delay: A Case Study of Wuhan–Guangzhou High-Speed Rail." *Transportation Research Record*. DOI:10.1177/0361198118780830
Liang, Zhang, Liu Jianhua, Wu Ruofei, and Gong Xiaobin. 2009. "Design of Performance Testing System for Train Air Conditioning." Energy and Environment Technology, 2009. ICEET'09. International Conference on IEEE 1: 85-89.
Milinković, Sanjin, Milan Marković, Slavko Vesković, Miloš Ivić, and Norbert Pavlović. 2013. "A fuzzy Petri net model to estimate train delays." *Simulation Modelling Practice and Theory* 33:144-157. DOI: 10.1016/j.simpat.2012.12.005.
Murali, Pavankumar, Maged Dessouky, Fernando Ordóñez, and Kurt Palmer. 2010. "A delay estimation technique for single and double-track railroads." *Transportation Research Part E: Logistics and Transportation Review* 46 (4):483-495. DOI: 10.1016/j.tre.2009.04.016.
Olsson, Nils OE, and Hans Haugland. 2004. "Influencing factors on train punctuality—results from some Norwegian studies." *Transport policy* 11 (4):387-397.
Takimoto, T. 2000. "Development of efficient operational control using object representation." *WIT Transactions on The Built Environment* 50.
Wallander, Jouni, and Miika Mäkitalo. 2012. "Data mining in rail transport delay chain analysis." *International Journal of Shipping and Transport Logistics* 4 (3):269-285.
Wen, Chao, Zhongcan Li, Javad Lessan, Liping Fu, Ping Huang, and Chaozhe Jiang. 2017. "Statistical investigation on train primary delay based on real records: evidence from Wuhan–Guangzhou HSR." *International Journal of Rail Transportation* 5 (3):1-20.
Xu, Peijuan, Francesco Corman, and Qiyuan Peng. 2016. "Analyzing railway disruptions and their impact on delayed traffic in Chinese high-speed railway." *IFAC-PapersOnLine* 49 (3):84-89.

Yuan, J, RMP Goverde, and IA Hansen. 2002. "Propagation of train delays in stations." *WIT Transactions on The Built Environment* 61.

EVALUATION of TRAVEL TIME RELIABILITY USING "REVEALED PREFERENCE" DATA & BAYESIAN POSTERIOR ANALYSIS

Sida Jiang [a,1], Christer Persson [b], Karin Brundell-Freij [c]
[a] WSP Sweden 121 88
Arenavägen 7 Stockholm-Globen, Sweden
[1] E-mail: sida.jiang@wsp.com, Phone: +46 (0) 70-231 68 77
[b] Urban Planning & Environment, Royal Institute of Technology KTH
Teknikringen 10, Stockholm, Sweden
[c] K2 – National Science Centre for Public Transport, Lund University
Bruksgatan 8, Lund, Sweden

Abstract

In Swedish context, the value of delay is deemed equalling to the value of travel time reliability (VoR), which is a factor of value of travel time (VoT) and mostly derived from Stated Preference (SP) studies. According to our knowledge, there are several issues with the SP method for obtaining VoR, for example, its deficiency in harmonizing the stated choices with the actual choices. On the other hand, Revealed Preference (RP) data from ticket sales has its limit in, for example, socioeconomic information of travellers and scenario variation.
This project aimed to use a RP method to evaluate travel time reliability through reliability ration (RR) - the relation between VoR and VoT upon several selected railway corridors, with Bayesian posterior analysis to infer socioeconomic differences between passengers given on their actual choices.

The data in the study are from two sources, ticket data from a major passenger operator SJ, and data on train movements from Trafikverket's (Swedish Transportation administration) TFÖR database. Both data sources are for the whole year 2009. The data includes 60 545 individual observations on traveler's route choice for two specific trip relations. The chosen trip relations are long-distance non-commuting trips with travel distances between 200 and 250 kilometers.

The project is a "proof-of-concept" for possible use of ticket data for the evaluation of travel time reliability. We can conclude that the estimated VoR – 1,13 times value of travel time, is in compliance with results from previous international studies using SP and/or RP data. The simulated distribution of RR from posterior analysis also clearly indicates a bimodal pattern of valuing travel reliability, probably due to socio-economic characteristics or trip purposes.

Keywords

Value of reliability (VoR), Bayesian model, posterior analysis, mixed logit, revealed preference (RP).

1. Background

With the deregulation of railway operations and modern information system practiced in an increasing manner in Sweden, travelers have more accessibility and flexibility yet inevitably more complicated travel choices to make from time to time. In this context, travelers encounter reliability issues in the form of delays when using railway. There are reasons to assume that travelers evaluate reliability in form of the expected day-to-day variability differently from delays relating to unexpected or surprising events. This study is an attempt to provide estimates of the valuation of travel time reliability, in economic sense. Areas of application for these valuations are in planning of maintenance measures and informing travelers about the rescheduling.

According to the latest socio-economic evaluation guide from Swedish National Transport Administration (Trafikverket, ASEK6.1), value of reliability (VoR) is 3.5 times VoT which has been derived from stated preference (SP) studies. Nevertheless, there are issues with the SP method for obtaining VoR. For example, travel reliability that is a measure of a probability distribution is not a straight forward concept to present to the respondents; also, it remains unclear whether there is a robust consistency between choice that are made in a specific hypothetical situation and revealed behavior observed in reality. These issues would be even larger for this proposed study where we need to distinguish between expected reliability and unexpected delays, and the former will be modelled in as a perceived factor determining how traveler choose ahead of different alternatives, while the latter can't be counted before the trip is made.

Due to the above-mentioned issues with the SP method, we have chosen to use revealed preference data recorded in statistics from year 2009. The data in the study are from two sources, ticket data from a Swedish railway operator (SJ) and data on train movements from Trafikverket's TFÖR database. The main reason for using relatively old data is that it has given us the opportunity to use detailed ticket data without further concerns on commercial disclosure. The data includes 60 545 individual trips on traveler's route choice for two specific trip relations: 1) from Örebro to Stockholm and 2) from Borlänge to Stockholm. These two relations have been chosen to homogenize the data. Homogenization of data is a reasonable strategy for a new research area since it will make estimated valuations more accurate, but in the same time compromises the generalizability of these estimates to trip relations with other characteristics than the relations contained in the sample. The chosen trip relations can be described as long-distance non-commuting trips with travel distances between 200 and 250 kilometers.

2. Hypothesis and Model Design

2.1 Empirical Setup

To evaluate travel time and its reliability, we use revealed route preference for railway traffic. This is a new area of research and it may be difficult to empirically establish the proposed valuation of travel time reliability from other determining factors. Therefore, the evaluation is based on a homogenized route choice data set consisting of three choice relations with the central station in Stockholm as one end-point. All three choice relations have similar trip distances 200-250 kilometers. The choice relations are depicted in figure 1 below.

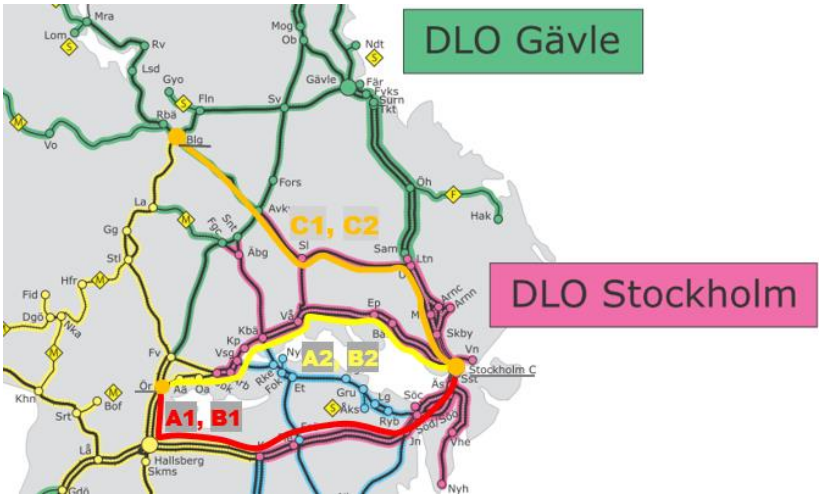


Figure 1 Illustration of studied railway routes

A. From Örebro (Ör) to Stockholm Center (Stockholm C) in the morning peak hours: alternative 1 is a transfer via Hallsberg and alternative 2 is a non-transfer train,
B. From Örebro to Stockholm C in the off-peak afternoon: alternative 1 is a transfer via Hallsberg and alternative 2 is a non-transfer train,
C. From Borlänge (Blg) to Stockholm C in the morning peak hours: alternative 1 is regional train and alternative 2 is high-speed x2000.

For instance, in choice task A, travelers can only choose either alternative 1 or 2, and both alternatives can be characterized by planned travel time PTT, departure delay (usually informed before departure from the start station) D^{inf} and the uninformed but self-estimated travel time reliability – standard deviation of travel time SD(TT), travel cost C and alternative specific constants ASC. A linear form is assumed for the specification of the utility for a traveler, which therefore attain the following form:

$$U_1 = \boldsymbol{\beta_T} * \mathrm{PTT}_1 + \boldsymbol{\beta_{SD}} * \mathrm{SD}(\mathrm{TT}_1) + \boldsymbol{\beta_{D^{inf}}} * D_1^{inf} + \boldsymbol{\beta_C} * \mathrm{C}_1 \quad (1)$$

$$U_2 = \boldsymbol{\beta_T} * \mathrm{PTT}_2 + \boldsymbol{\beta_{SD}} * \mathrm{SD}(\mathrm{TT}_2) + \boldsymbol{\beta_{D^{inf}}} * D_2^{inf} + \boldsymbol{\beta_C} * \mathrm{C}_2 + \boldsymbol{ASC} \quad (2)$$

Travel time PTT is the planned travel time or interchangeably timetable travel time. PTT is rather fixed and can be considered as only varying over alternatives and choice tasks, thus the difference between PTT over two alternatives can function as choice-task specific constants (CTSC), and due to its importance in explaining the utility and to prevent high correlation with ASC, PTT (or CTSC) is employed to replace ASC.

Travel cost C is not available at the individual purchase level therefore been assumed the same and quantified with an average over all individual tickets in the studied alternative, and in this case it is missing thus enters into the constant term ASC.

Travel time reliability or the risk of delay against time table is difficult to model and results in a massive variety of indicators for theoretical and practical appliance. Börjesson & Eliasson (2011) concluded evaluation of travel time reliability varies over the frequency of travel delay; Fosgerau and Karlström (2010) are giving an expression for VoR when standard deviation is used as the attribute for reliability. The expression for VoR, in this framework, is a function of the travel time distribution. In accordance with (Fosgerau and Karlström, 2010) this study uses standard deviation of travel time as the indicator of travel time reliability. However, in contrast to their approach, (i. e. to estimate the travel time distribution and then compute VoR from the distribution), we estimate reliability ratio RR empirically, as the ratio $\boldsymbol{\beta_{SD}/\beta_T}$ between the parameters given by the utilities in equation (1) and (2). Since standard deviation is by far the most common attribute for travel time variability, which makes the results comparable with other SP/RP studies in Sweden and worldwide. Yet it is not informed or deterministic, which invites random effects from person to person, together with the needs of posterior analysis, the project therefore employs also mixed logit model for better design and modal fit.

2.2 Mixed Logit Model

The utility equation for mixed logit (McFadden& Train, 2000) with randomized effects of travel time and travel time reliability assumed to follow a normal distribution over the travelers, so that the unconditional choice probability:

$$P_{n,\ i}(\Omega) = \int_{\beta} \left[\frac{e^{\beta x_i}}{\sum_{j=1}^{J} e^{\beta x_j}} \boldsymbol{f(\beta_T|\,\Omega)} \right] d\beta \qquad (3)$$

where $\boldsymbol{f}(\cdot\,|\boldsymbol{\Omega})$ is the density of a normal distribution with parameters $\boldsymbol{\Omega}$. The difference from ground model, a binomial logit with utitlities given by (1) and (2), is that the choice probabilities $P_{n,\ i}(\Omega)$ is the average binomial choice probabilities over the normal distribution with parameters $\boldsymbol{\Omega}$. To establish a computationally efficient estimation of this mixed model, uniform random draws are first generated using a Halton sequences[1], then the inverse function of the cumulative density function is used to derive standardized **normal draws** with mean 0 and standard deviation 1. With this standardized normal distribution, the utility functions with randomized effects of travel time and travel time reliability can be written as follows:

$$U_1 = (\boldsymbol{\beta_T} + \boldsymbol{\sigma_T} * \boldsymbol{draws_T}) * \mathrm{PTT}_1 + (\boldsymbol{\beta_{SD}} + \boldsymbol{\sigma_{SD}} * \boldsymbol{draws_{SD}}) * \mathrm{SD}(\mathrm{TT}_1) + \boldsymbol{\beta_{D^{inf}}} * D_1^{inf} \qquad (4)$$

$$U_2 = (\boldsymbol{\beta_T} + \boldsymbol{\sigma_T} * \boldsymbol{draws_T}) * \mathrm{PTT}_2 + (\boldsymbol{\beta_{SD}} + \boldsymbol{\sigma_{SD}} * \boldsymbol{draws_{SD}}) * \mathrm{SD}(\mathrm{TT}_2) + \boldsymbol{\beta_{D^{inf}}} * D_2^{inf} \qquad (5)$$

That the choice probability can therefore be reformulated with *N(0,1)* random draws as follows:

[1] Halton method divides $0-1$ space into p_k segments (with p_k giving prime used as base for parameter k), and by systematically filling the empty spaces, using cycles of length p_k.

$$P_{n,\ i}(\beta_x, \sigma_x) = \int_{\varepsilon_x} \left[\frac{e^{\beta x_i + \sigma_x \varepsilon_x x_i}}{\sum_{j=1}^{J} e^{\beta x_j + \sigma_x \varepsilon_x x_j}} \boldsymbol{\phi}(\mathbf{0}, \mathbf{1}) \right] d\varepsilon_x \quad (6)$$

Instead of estimating fixed parameter for travel time and travel time reliability, both the mean $\boldsymbol{\beta_T}$, $\boldsymbol{\beta_{SD}}$ and standard deviation $\boldsymbol{\sigma_T}, \boldsymbol{\sigma_{SD}}$ are to be estimated based upon simulated log-likelihood maximization (Bhat, 2001).

3. Model Estimation

The travel time and travel time reliability has been randomized with standard normal draws, according to the preceding sections, so that the mixed logit model requires estimation of both mean and sigma for how the sensitivity of travel time and travel time reliability vary over individuals. Also, the mixed logit model is a non-linear model which can result in numerous local optima. To handle the local optima issue, different initial values has been run for the first 100 iterations and the initial values with the best log-likelihood has been chosen for further model estimation. The initial log-likelihood is LL(0): -41966.6 and the model converged at **LL(final): -24 151.5**.

Table 1 Estimation of Mixed Logit Model (with 300 random draws)

	Est.	*s.e.*	*t-value(0)*	*robust s.e.*	*robust t-value(0)*
$\boldsymbol{\beta_{SD}}$	-0.46	0.01	*-48.98*	0.01	*-46.72*
$\boldsymbol{\beta_T}$	-0.37	0.01	*-49.59*	0.01	*-46.97*
$\boldsymbol{\sigma_{SD}}$	-0.25	0.01	*-29.10*	0.01	*-23.64*
$\boldsymbol{\sigma_T}$	-0.31	0.01	*-40.95*	0.01	*-33.44*
$\boldsymbol{\beta_{D^{inf}}}$	0.17	0.00	*48.22*	0.01	*21.68*

Rho-sq: 0.42; adj. rho-sq: 0.42; AIC: 48 312.9

Standard deviation $\boldsymbol{\sigma}$ for both travel time (T) and travel time reliability (SD) are (strongly) significant different from 0. This means that we have shown that there is variation across individuals of how those attributes as trade-off against each other. Including a random effect can hence improve the modal fit. The same conclusion can be found by comparing LL(final), adjusted rho square and AIC that mixed logit model fits better than corresponding basic logit model.

Coefficient for departure delay (D^{inf}) is positive, which is contrary with the expectations. the likely explanation is that the explanatory power of departure delay is likely to be confounded by other variables that varies over choice tasks. Because of this problem, which is inherent in data, the project conclusions focus more on the importance of planned travel time and travel time reliability. Again, this counter-intuitive finding is most probably limited to the specific study scope and in the many cases with only small departure delay. In the rarer cases where there is a long departure delay, it is most probable that departure delay will be used (righteously) as a predictor by the traveler, indicating that one can expect

a longer than usual delay at the final destination also.

4. Posterior Analysis

The mixing distributions given by the parameters in table 1 can be used directly to construct the distribution of the reliability ratio RR. However, in the specification, the mixing distributions are assumed to be independent. This assumption is likely to not be fully valid. Since the distribution of especially very high RR as well as very low RR among individuals is highly dependent on the validity of this assumption, the computed RR from the mixing distributions can be expected to deviate considerably from its true value in the population. Further the mixing distributions are assumed to be Gaussian. Under the estimated parameters given in table 1 there will, for example, be a sizeable proportion of the population for which RR will have a wrong sign. Therefore, it is a need for a more robust estimation of RR. The method used is a so called posterior analysis (Hess, 2007) where, in a Bayesian spirit, the posterior distribution of RR is obtained by applying the mixing distributions to the likelihood of the data. This method can be seen as a way to correct the mixing distributions such that they comply to observed dependencies in the data (i.e. the likelihood). In this sense, the obtained distribution for RR can be seen as more robust than the distribution obtained directly from the estimation of mixed logit model.

Knowing that different individual has significantly different evaluation of both travel time and travel time reliability, we can further divide the individuals into several groups conditional on their observed choices. In the meantime, the reliability ratio is no longer limited to the average level as earlier illustrated ratio of coefficients, by using posterior analysis upon the mixed logit model. Each individual is assumed to follow a random distribution (with simulated random draws) and each individual is assigned with the expected values of this random distribution termed as conditional mean. Notice that conditional mean is not the actual sensitivities for that individual but the expected mean, in other words, it is associated with different simulation of corresponding distribution and how many random draws one allows for each individual. The study uses 300 as number of random draws, furthermore sensitivity to both travel time and travel time reliability is assumed to be normally distributed.

For more details in posterior analysis of mixed logit model please refer to Hess (2007). The probability of observing the specific value of β given the choice of individual *n:*

$$\widehat{\beta_n} = \frac{\sum_{r=1}^{R}[L(Y_n|\beta_r)\beta_r]}{\sum_{r=1}^{R} L(Y_n|\beta_r)} \quad (7)$$

Where β_r with $r = 1, \dots, R$ are i.i.d draws, this method will relax the independent assumption of composing variables imposed by unconditional estimation, in other words, the resulting ratio of estimated coefficients is supposed to be more robust and fit into the reality revealed by the data. However, this would again lead to problems with data outliers.

The descriptive statistics of posterior analysis results for travel time, travel time reliability and reliability ratio is summarized in table 2:

Table 2 descriptive statistics of posterior analysis results for the parameters for travel time, travel time reliability and reliability ratio

Statistics	*Travel time*	*Travel time reliability*	***Reliability ratio***
1st quartile	-0,207	-0,252	0,806
median	-0,106	-0,213	1,280
mean	-0,101	-0,214	**1,132**
3rd quartile	-0,040	-0,197	3,808
Std. dev. of Sample	0,137	0,050	94,517
Std. dev. of Sample mean	0,0006	0,0002	**0,384**

As the table shows, the mean and median of RR from mixed logit model are both slightly above 1 and quite close to RR value from the recent international studies summarized by Carrion and Levinson (2012), see the figure 2 bellows.

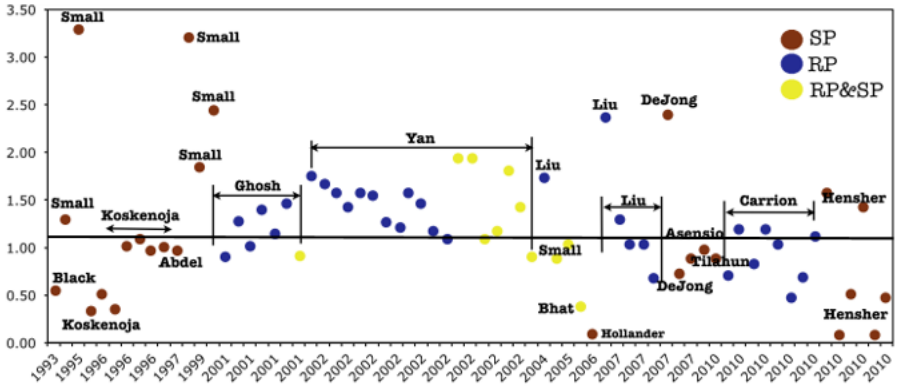


Figure 2 reliability ratio of selected studies (Carrion and Levinson, 2012)

One may argue, based on the figure, that estimated reliability ratios have declined, over the past two decades, in both SP and RP data, with RP estimates more constrained in middle of the span. In other words, travel time reliability seems to play relatively less and less role compared to travel time in the utility function. This may be explained by easier accessibility to travel information in general: travelers can forecast the coming journey and make changes accordingly, so that unreliable travel time become more and more predictable; also, substitution with digital activity can help travelers make as efficient use of travel delay as she/he can make of planned travel time. This declining trend of reliability ratio should probably be considered when conducting cost-benefit analysis (CBA) with travel time reliability involved. Results from our model (illustrated as a line in figure 2) is quite in line with other RP studies in which RR has varied from 0.5 to nearly 2.5.

But also, from posterior analysis, a significant spreading of RR between 1 and 4 has been observed, which could potentially and partly be explained by different trip purposes.

This variation needs to be further examined and put in relation to complementary information about trip. The detailed distribution of the estimated reliability ratio is indicated in the figure 3:

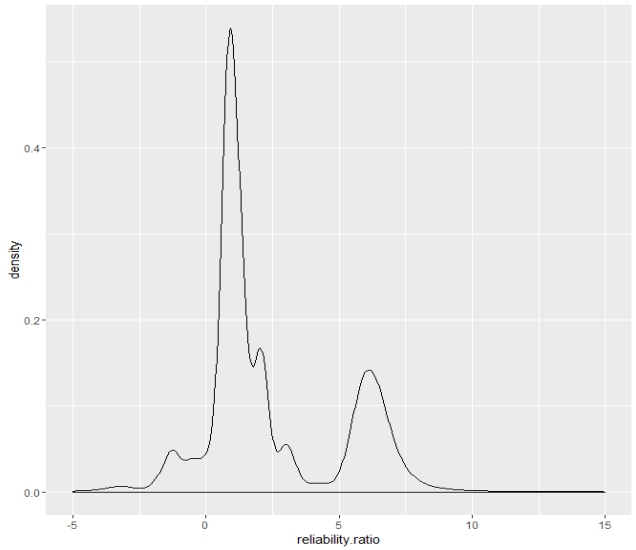


Figure 3 Probability density function of RR from posterior analysis

As expected, the major part of distribution is positive with several modes, the 1st group with around 50% of the individuals has a RR slightly less than 1. This group of travelers are the ones for which their observed choices indicate that they are relatively less sensitive towards travel time reliability. This group is also the majority in our data, but different data and scope of study can change its dominance and thus yields very different statistics. One important point of our results is therefore that multiple clusters or groups can be seen clearly in the posterior analysis, and we may need further data such as SP to understand better how socio-economic variables or trip purpose divide the sample, and then to specify reliability ratio with respect to e.g. private/business/work trip.

The 2nd largest group with around 15% of individuals has a high RR, with an average very close to 6. This group is thus about 6 times more sensitive of travel time reliability than the overall average. The division into two groups is however not absolute: the analysis also suggests that there are also individuals with RR between 1 and 6.

As with all other results, our conclusions are limited to the range of travel distances for which we have data. Obviously, the magnitude of RR can vary with among others travel distance, and the results illustrated above can only draw insights about the trips with distance between 200 and 300 km. Nonlinearity of RR with respect to travel distance can be complemented with data of other routes at different length.

5. Conclusions

In this project, ticket sales data from SJ has been transformed and treated as revealed preference data combined with travel time, travel time reliability and departure delay from

TFÖR database. Both data sources can be obtained from historical database and studies over different routes and years can be conducted for other analysis purposes, but one can also – as was done in this study – use the data as "RP" (revealed preferences) to conceptually examine how travelers react to travel time uncertainty or different forms of delay. Main results of this project can therefore be argued to be that it proved possible to use the type of data on observed behavior (RP-data), to estimate a model of how uninformed delays (travel time uncertainty) affect individual travel behavior.

As a basis for our analyses, we have used the choice task when a traveler chooses between different scheduled travel options. (One of the drawbacks of this approach is that the behavioral response not to go by train is not included). Our analysis circles around the travelers' trade-off between three central qualities of the travel options: Travel time (as planned in the time table), delay at departure, travel time uncertainty (based on the distribution of real travel times in the recent past)

Our data only comprises three distinct choice situations (Choice situation = A pair of adjacent scheduled train departures for the same destination). Although the choice tasks have been selected so that they are similar in nature, the alternatives will inevitably differ in many more aspects than the three measured qualities that is introduced in our analyses. Therefore, there is a risk that our results are confounded by other variables with which our explanatory variables co-vary over alternatives and choice tasks. Also, only travel distances in the range 200-300 km is covered in data.

For travel time (as scheduled in time table) our estimations give the intuitively correct sign for the estimated parameter. Travel time varies only with single minutes for the same alternative in a given choice situation (due to minor modifications of the timetable during the observed year). Thus, the estimated value is based almost entirely on the variation between the three choice tasks. Therefore, it is reassuring that the estimated value has the correct sign. Never the less, it is clear that a larger data set (that is many more choice tasks) would have been highly desirable. Since it was not possible to estimate parameters for the sensitivity to costs, it is not possible to check whether the estimated sensitivity to travel time is reasonable in terms of "value-of-time". However, we can conclude that the ratio between the parameters estimated for travel time and travel time reliability, was estimated to 1.13, which is very much in line with what has been estimated in previous studies. The reliability ratio can be used for socio-economic evaluation regarding investment and maintenance for a more robust railway service, from the perspectives of passengers.

6. Future Work

Before our study was conducted, the novel approach we were proposing raised concerns as to whether:

1. Data quality was good enough, given that data from multiple sources were combined and one data source (ticket statistics) had not previously been used.
2. It would be possible to estimate reliable parameter values for the two relevant variables journey time and travel time uncertainty.

We can now conclude that data quality seems to be sufficient, and that the data allow the estimation of models that are suitable for the purpose.

A particular difficulty is that we have studied how travelers choose between trains. This means that we miss in our analysis the traveler's option to adapt to uncertainties and delays by abandoning the train altogether, either by switching to another mode, or to forgo the trip.

If future studies are extended to include also such alternatives, it may help to allow the estimation of more relevant parameters.

To summarize what is probably needed to increase the possibility to estimate the effects / parameters for informed delay also:

- Allow the option "not-use-train" into the described choice situation. In practice, this can be done by using not only distribution of rail passengers between train alternatives, but also the total number of train travelers.
- Include more data (more choice situations) to provide (1) better estimates, (2) reduced risk of unmeasured attributes that vary between choice tasks being confounded into the estimated parameters and (3) possibility to study how values differ between different types of travel (there are indications of the estimates on the valuation of uncertainty is bimodal).

Mixed logit model has been tested with better fit for the data. In future work it would also be useful to develop that approach further, for example test different random distribution, number of draws as well as modified specifications of utility function for improvement of model fit. In a word, current results have shown differences between how travel time, travel time uncertainty and different types of delay is evaluated, and also significant variance cross observed individuals. Future analysis is to extend the model so that it can utilize RP data to calculate VoT, VoR and RR over different trip purposes, travel distance and for other analysis practices.

References

Bhat C.R., 2001. Quasi-random maximum simulated likelihood estimation of the mixed multinomial logit model [J]. *Transportation Research Part B: Methodological*, 35(7): 677-693.

Börjesson M., Eliasson J., 2011. On the use of "average delay" as a measure of train reliability [J]. *Transportation Research Part A: Policy and Practice*, 45(3): 171-184.

Carrion C, Levinson D, 2012. Value of travel time reliability: A review of current evidence [J]. *Transportation research part A: policy and practice*, 46(4): 720-741.

Fosgerau, M., Karlström, A, 2010. The value of reliability, *Transportation Research Part B: Methodological* 44, no. 1, 38-49.

Greene W.H., Hensher D.A., 2003. A latent class model for discrete choice analysis: contrasts with mixed logit [J]. *Transportation Research Part B: Methodological*, 37(8): 681-698.

Hess S., 2007. Posterior analysis of random taste coefficients in air travel behaviour modelling [J]. *Journal of Air Transport Management*, 13(4): 203-212.

McFadden, D., & Train, K. 2000. Mixed MNL models for discrete response. *Journal of applied Econometrics*, 15(5), 447-470.

Trafikverket, 2018. Analysis method and socioeconomic calculation values for transport sector: ASEK 6.1. Chapter 8 cost for congestion and delay. Table 8.1.

Walker J., Ben-Akiva M., 2002. Generalized random utility model [J]. *Mathematical Social Sciences*, 43(3): 303-343.

Applied Timetabling for Railways: Experiences with Several Solution Approaches

Julian Jordi [a], Ambra Toletti [a], Gabrio Caimi [a,1], Kaspar Schüpbach [b]
[a] SBB AG, Eigerstrasse 13, 300 Bern 65, Switzerland
[1] E-mail: gabrio.caimi@sbb.ch, Phone: +41 (0) 79 124 77 29
[b] ELCA Informatik AG, Switzerland

Abstract
As part of the smartrail 4.0 program, SBB is focusing with the project TMS (Traffic Management System) on algorithmic supported, optimized and integrated capacity planning. For solving this problem, we have experimented with different approaches from the literature and have compared their quality and performance. In this paper, we present the results of this comparison on a set of 8 specific instances that we also used for a crowdsourcing challenge. We also discuss how we intend to use our lessons learned for having the best possible solution for our ambitious goal.

Keywords
Timetabling, MILP, Constraint Programming, Alternative Graph

1 Introduction

Swiss Federal Railways (Schweizerische Bundesbahnen, short SBB) is pushing ahead digitization and automation of railway planning and operations. Customers are to benefit from higher capacities, less disturbances, better radio communication, improved customer information and lower overall costs. Railway infrastructure utilization is to be increased by shorter headway times and more precise planning. For this purpose, SBB has launched the *smartrail 4.0* program along the following principles.

- Algorithmically supported, optimized and integrated capacity planning
- Advanced control systems for railway operations
- New generation of digital interlocking systems
- Significant reduction in quantity and diversity of signaling systems
- Network-wide roll-out of the ETCS cab signalling
- Increased data transmission capacity
- Highly available and precise tracking of trains

The smartrail 4.0 program is organized in 4 principal streams: ETCS Interlocking (EI), Localization, Connectivity and Security (LCS), Automatic Train Operation (ATO) and Traffic Management System (TMS). In the context of smartrail 4.0, the TMS-stream strives to reach the following goals:

- Integrated and automated planning
- Automation of the operations centers. The employee develops himself from user to manager of the system.
- Enabling of efficient, real-time and precise automation and control of train movements and speeds.
- Precise coordinated remote control of departure, driving and arrival of trains.

The opportunities for the railway system behind these goals are pointed out in Weidmann et al. (2014).

The program clearly aims for an evolutionary approach towards automation of the planning and operation systems, in order to realize improvements step by step. During transition, it is crucial to take the human factor and the interaction between manual and automated processes into account. Certain roles will have to prepare the inputs and have to understand and post-process computational results. Particularly in the long-term planning process with many commercial and political aspects, the human factor will remain still important for long time, even if also at this stage algorithmic decision support will be important.

The key factor for the success of TMS is to find the right approach which enables on the one hand to have a strong algorithmic performance to solve big instances of the size of Switzerland, and on the other hand to enable a continuous integration in the current processes, with particular focus on the man-machine interaction.

2 Problem Formulation

The timetabling rules we consider were set by experienced railway planners. Planners also provided the test scenarios on which the different formulations are tested (see 4.2).

The Timetabling Problem considered in this paper is as follows: Given a list of trains to be scheduled, and for each train

- a list of commercial requirements, such as earliest departure times, latest arrival times, minimum dwell times and connections to other trains.
- a directed acyclic *route graph* that defines the routes the train could take from origin to destination. Each arc in the graph is called a *route section* and has associated to it the minimum running time for this train on this edge and a list of resources that are occupied while the train is on this section. A section may also have a non-negative *penalty* attached that is counted in the objective function if this section is used in the solution

Choose for each train exactly one path from origin to destination in the route graph and assign to each node on this path a time such that

- all commercial requirements are satisfied
- all minimum running times on the route sections are satisfied
- no resource occupation conflict results
- the objective function (see below) is minimized

The only commercial requirement that may be violated are latest arrival times. A violation of a latest arrival is a *delay*. All other rules are hard constraints. The *objective function* is the sum of all delays plus the sum of all route penalties for the chosen routes.

In this paper, we only consider *blocking* resources. To avoid *resource occupation conflicts*, it must be allocated to one train movement exclusively. The next train may only allocate it once the previous train has released it and a given release time has elapsed.

3 Approaches

3.1 MILP Model

One standard approach is to formulate this timetabling problem as a Mixed Integer Linear Programming (MILP) and to solve it by general-purpose MILP solvers. These can in principle solve instances to optimality, if problem size and computation time limits permit. For this study, we have used IBM Cplex Version 12.8.

In our model, each train run is modelled as a series of events with trip sections in-between. Attached to the events are commercial requirements such as earliest/latest departure/arrival times and connections. The sections contain information on required process durations and resource occupations.

Continuous decision variable are introduced for the event times. Binary variables define the routing alternatives taken and the precedence order of train pairs on sections with common resource occupations.

Our model is similar to the ones described in Pellegrini et al. (2012) and Fischetti and Monaci (2015). Dependencies between event times are modelled as simple time difference constraints: Trip times, stop times, connection times and release times. Big-M constraints are used to switch off all constraints which become oblivious depending on binary routing and precedence decisions.

Different to Pellegrini et al. (2012) where each origin-destination route gets it own decision variable, we introduce one variable for each local routing alternative. Continuity of the local decisions is enforced by flow conservation constraints. Other differences are that we minimize the weighted sum of delays and that we currently omit constraints on rolling stock circulation.

As in Fischetti and Monaci (2015), we limit the maximum delay allowed for each event as hard constraint and only introduce precedence decision variables and constraints for pairs of trains, which are temporally close enough to be potentially conflicting.

Further we merge precedence decisions for resource occupations of neighboring sections, whenever change of precedence ordering is not possible.

A comparison of the solver performance with and without these enhancements is available in Schupbach et al. (2018).

3.2 CP Optimizer

IBM's CP Optimizer[1] (CPO for short) is a general-purpose constraint solver with a strong focus on scheduling problems. As part of this focus it provides *interval* variables out of the box as central model element. An interval variable represents an activity that is to be

[1] https://www.ibm.com/analytics/cplex-cp-optimizer

planned. An expressive and intuitive constraint language is available to model conditions on the activities. See Laborie et al. (2018) for a general overview of CP Optimizer.

The expressiveness and flexibility of the language makes it comparatively easy to incorporate new planning rules into the model. In our experience this makes CPO especially suitable for prototyping new ideas before implementing them also in other solvers.

Also, CPO naturally allows a resource view on the problem, rather than a train-pair view as in 3.1 and 3.3. Instead of demanding that *for each train pair* occupying a common resource, one of the train has to precede the other, one constrains that *for each resource*, the intervals *of all trains* must not overlap. This leads to models that grow only linearly in the number of trains and thus remain fairly compact.

The CPO Model

For each route section of each service intention introduce an interval variable. It represents the activity of traveling through this route section for this service intention.

Time windows (such as earliest departure, latest arrival, etc.) are handled by setting the allowed domain of the associated interval variables accordingly. Similarly, minimum running times are guaranteed by setting a minimum length for the interval. For connections, introduce timing constraints such as startBeforeEnd between the appropriate intervals.

Resource occupation conflicts are avoided with appropriate no-overlap constraints. Routing alternatives are handled using the presence status of the intervals. Just as in the flow formulation for the MILP, one has to make sure that the solver chooses exactly one path through the routing graph for the solution.

3.3 ASP

Answer Set Programming (ASP) is a declarative problem solving approach. For our application we used Potassco, the library developed at Potsdam University. See Gebser et al. (2012) for a description of the language and tools. We transformed the problem described in section 2 into a set of declarations. Then, the Potassco grounder aggregated the declarations and defined the decision variables for the optimizer. These decision variables correspond to the routes and time events that are not naturally involved by the declarations. Finally, the optimizer selected the route and the time variables that return the best objective.

Similarly to the other formulations, for each route section of each service intention a binary decision must be taken. The selection of the previous section is a precondition for each route choice. Similarly to the MILP formulation (Section 3.1) time variables represent the entrance of a train into a section and the exit from the last sections.

The time variables are linked together via difference logic constraints see for example Kaminski et al. (2017), which are linear inequalities that are activated when the declared preconditions are satisfied (cfr. Big-M in Section 3.1). For instance, minimum travelling time constraints are activated if the corresponding route sections are selected. Analogously, minimum time separations are activated if different trains are routed through common resources.

3.4 Alternative Graph Library

The Alternative Graph Library (AGLib) is an academic optimization framework specialized for scheduling and routing trains in real-time. [2]

As described in Samà et al. (2015), the solver splits scheduling and routing into separate steps. Starting from an initial set of routes, the first schedule is computed applying a branch-and-bound scheme on the so-called alternative graph. Then, in the routing step, neighborhood search algorithms look for improving routings. Applying scheduling and routing steps alternatingly, AGLib iteratively improves solution quality.

The objective function built into AGLib is the minimization of the maximum delay, measured over all events. When comparing to the cost function described in the challenge formulation, this can lead to different optimal solutions and needs to be kept in mind when reading the following results section 4.

3.5 The 'IQUADRAT' Solver: A Greedy Algorithm Using Resource Usage Density

From August through November 2018 SBB conducted a crowdsourcing Challenge[3] to solicit algorithms for solving the timetabling problem described in section 2.

By design, the challenge was limited to a fixed set of nine problem instances. Due to technical limitations, it was not possible to evaluate the algorithms on an independent test set. This setup probably explains why greedy approaches were quite successful, as they could be fine-tuned on these particular instances. Nonetheless, the winning algorithms appear reasonably generic, using at most a handful of parameters.

Among the leading participants a wide variety of approaches and technologies were tried. A common theme among the very best submissions was the use of greedy algorithms, although MILP and Constraint Programming approaches as well as enhancing greedy scheduling with reinforcement learning techniques[4] (Q-Learning) were also successfully applied.

Some of the leading participants have made their code publicly available[5]. We encourage an academic review and possible extension of these ideas.

For our comparison tests in section 4, we include from this challenge the winning algorithm by participant 'iquadrat'.[6]

The Algorithm

The IQUADRAT algorithm greedily solves the most *critical* conflicts first. A *conflict* is defined as any two train sections that potentially occupy a common resource simultaneously. This also takes into account routing alternatives (so there may be many conflicts on a given resource between two trains). *Criticality* of a conflict is determined by a function that takes into account usage density of the associated resource and the planning flexibility of the associated trains (if the trains have a large time window, the conflict is not critical).

For each conflict, at most four resolutions are possible: Train 1 before Train 2, vice versa, Train 1 takes a different path, Train 2 takes a different path. These resolutions are

[2] Many Thanks to Andrea D'Ariano and Marcella Samá from the Roma Tre University for sharing AGLib with SBB for evaluation purposes.

[3] https://www.crowdai.org/challenges/train-schedule-optimisation-challenge

[4] https://github.com/deuxnids/sbb_challenge

[5] Links to the code repositories are embedded in the following discussion thread on the challenge site: https://www.crowdai.org/topics/what-tools-did-you-use/discussion

[6] https://github.com/iquadrat/sbb-train-scheduler

again weighted and the preferred resolution is greedily applied. If a resolution leads to infeasibilty, the next one is tried. If all lead to infeasibilities, backtrack to previous conflict.

This greedy approach is similar in spirit to some Travel Advance Heuristics (TAH), such as the *critical first*-TAH in Khadilkar (2017), but it takes into account routing alternatives and the inherent flexibility of service intentions.

The algorithm is a satisfiability solver, not an optimizer. It stops after the first feasible solution has been found. The maximum allowed train-wise and total delays can be specified via parameters.

4 Tests and Results

4.1 Test Setup

We evaluate the different solvers on the problem instances of the Crowd Sourcing Challenge[7]. These instances are derived from the actual timetable on the triangle of lines Zurich - Thalwil, Thalwil - Chur and Thalwil - Lucerne. They differ in the time window they cover and the number of routing alternatives available to the trains. The complexity of the instances generally increases with their number. Instance 01 is too trivial to be interesting and is excluded. Instance 02 contains 58 trains with virtually no routing alternatives for a total of roughly 4300 route sections. Instance 09 contains 287 trains and with all routing alternatives a total of over 34000 route sections.

Most instances are solvable with zero delay. In this sense, they represent a 'normal' planning scenario where trains are scheduled without disturbances and with enough capacity on the network available. Instance 05 is special in that it simulates the effect of the closure of one track on a double track line and it is not solvable without significant delay; its optimal objective value is 32.98.

The solvers are run on pods in a cloud platform configured to have a CPU limit of 8 cores and 40GB of RAM, with the exception of the crowdsourced solver IQUADRAT, which is run on an office notebook with Intel i5-6300U CPU @ 2.4GHz and 12 GB of RAM. To simulate 8 usable cores, IQUADRAT, as a single-threaded solver, is run with the 8 different configurations listed in 1.

The solver MILP-ND (as in 'no-delay') uses the MILP model as described in 3.1, but assumes that no delay is allowed for any train. This drastically reduces the model size and search space compared to the MILP solver, where a maximum delay of 30 minutes for each train is assumed.

Timeout for all solvers and instances is 900 seconds.

4.2 Results and Discussion

Table 2 summarizes the results. We observe the following patterns.

The MILP solver shows decent results for smaller and medium-sized instances. In particular, it can prove optimality for instances 02-04. With bigger instances with more routing alternatives, computation times quickly explode. The no-delay feasiblity checker MILP-ND can solve much larger instances, also 06-08, provided they are feasible.

Instance 05 with the closed track, which is not solvable with zero delay, is best solved by IQUADRAT and AGLib. AGLib impresses by mostly keeping up with the other solvers

[7] https://github.com/crowdAI/train-schedule-optimisation-challenge-starter-kit/tree/master/problem_instances

Table 1: Settings used for the IQUADRAT solver

max_penalty	max_penalty_per_intention	connetion_badness_factor	provides best solution for instance
0	0	5	01, 02, 03, 06, 07, 08, 09
5	5	5	04
10	5	5	-
20	5	5	-
20	10	5	-
40	10	5	-
40	15	5	05
60	15	5	-

while only using one CPU core. Instance 05, in particular, highlights how AGLib is very efficient in improving the initial train routes in the right place and at the right time, namely where and when it is most critical. AGLib has problems with instance 09 since the initial route given to each train in the input file is not feasible in terms of computing a conflict-free train schedule. A future version of AGLib is expected to be able to recover from such infeasibilities.

In the set of instances used for this comparison, the algorithm by IQUADRAT works exceptionally well as a feasibility checker for a provided bound. For the medium and larger instances that are solvable with zero delay, it finds solutions much quicker than all other approaches. Further study with a larger variety of instances should be performed to assess the validity of the approach in general.

CPO is an all-rounder in the wide range of scenarios 02-08, but is outperformed by the specialists for with/without delay instances. ASP shows decent results for smaller and medium-sized instances but, same as bare MILP, does not inherenty scale. Model building, grounding and solving would have to be wrapped into an iterative framework that solves complicated instances step by step.

5 Conclusions and Outlook

In this paper, we have compared several different solution methods for the specific train scheduling problem (routing and timetabling) that we face at SBB. We can conclude that solving routing and scheduling in one MILP-instance seems to be too much for larger instances. A MILP model remains in any case our benchmark for solving smaller instances exactly. None of the approaches alone will work for the system we need to build, in terms of size but also in terms of business requirements (the 9 instances analyzed in this paper are still small compared to what we have to solve in reality). Therefore, we are of the opinion that decomposition methods for solving the problem are unavoidable.

This comparison also gave us some input to continue our work towards increasing the tractable scenario sizes. We plan to pursue the following research directsion:

- Implement iterative MILP solvers using row and column generation, thereby reducing

Table 2: Computation times for the problem instances. Numbers represent average computation time in seconds. Numbers in parentheses represent the objective value of the best solution found. Note that AGLib uses a different objective function, 3.4. 'infeas.' means the problem is infeasible. 'no sol' means the solver provided no feasible solution within the time limit.

Instance	MILP	MILP-ND	CPO	ASP	AGLib	IQUADRAT
02	8.8 (0)	1.4 (0)	7.6 (0.6)	23 (0)	4.3 (0.2)	4.0 (0)
03	49 (0)	3.0 (0)	15 (1.4)	52 (0)	8.9 (0.5)	8 (0)
04	164 (0.08)	8.8 (1)	20 (15.8)	80 (1)	18 (7.5)	16 (4.7)
05	900 (57.83)	infeas.	900 (84.4)	900 (44.85)	321 (124.6)	391 (39.95)
06	no sol.	219 (0)	211 (161.2)	no sol.	409 (105.9)	75 (0)
07	no sol.	613 (0)	276 (218)	no sol.	699 (133.2)	128 (0)
08	no sol.	266 (0)	100 (99.6)	no sol.	338 (141.3)	44 (0)
09	no sol.	no sol.	no sol.	no sol.	no sol.	277 (0)

model size.

- The strategy of first solving a no-delay feasibility problem with reduced search space is promising for each of the solvers.
- Further investigation of routing and scheduling heuristics of AGlib and evaluate possibilities for parallelization.

Furthermore, we see much promise in combining different solvers by sharing solutions across them. For example, we plan to use initial solutions from quick heuristic solvers, such as AGLib, ans using them as starting solutions for the MILP-solvers, or even inject them during a solve run.

References

M. Fischetti and M. Monaci. Using a general-purpose MILP solver for the practical solution of real-time train rescheduling. Technical report, 2015.

M. Gebser, R. Kaminski, B. Kaufmann, and T. Schaub. *Answer Set Solving in Practice*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers, 2012.

R. Kaminski, T. Schaub, and P. Wanko. A tutorial on hybrid answer set solving with clingo. In G. Ianni, D. Lembo, L. Bertossi, B. Faber, W.and Glimm, G. Gottlob, and S. Staab, editors, *13th International Summer School of the Reasoning Web*, volume 10370 of *Lecture Notes in Computer Science*, page 167–203. Springer-Verlag, 2017.

H. Khadilkar. Scheduling of vehicle movement in resource-constrained transportation networks using a capacity-aware heuristic. In *Proceedings of the American Control Conference*, 2017.

H. Khadilkar. A Scalable Reinforcement Learning Algorithm for Scheduling Railway Lines. *IEEE Transactions on Intelligent Transportation Systems*, pages 1–11, 2018.

P. Laborie, J. Rogerie, P. Shaw, and P. Vilím. IBM ILOG CP optimizer for scheduling: 20+ years of scheduling with constraints at IBM/ILOG. *Constraints*, 2018.

P. Pellegrini, G. Marlière, and J. Rodriguez. Real time railway traffic management modeling track-circuits. In *ATOMOS 2012, 12th Workshop on Algorithmic Approaches for Transportation Modelling, Optimization, and Systems*, page 12p, France, Sept. 2012.

M. Samà, A. D'Ariano, D. Pacciarelli, and F. Corman. A variable neighborhood search for fast train scheduling and routing during disturbed railway traffic situations. *Computers & Operations Research*, 01 2015.

K. Schupbach, G. Caimi, and J. Jordi. Towards Automated Capacity Planning in Railways. In *IRSA 2017: Proceedings: 1st International Railway Symposium Aachen, Germany, 28-30 November 2017*, pages 310–325, Aachen, Nov 2018. 1st International Railway Symposium Aachen, Aachen (Germany), Publication Server of the University Library, RWTH Aachen University.

U. Weidmann, M. Laumanns, M. Montigel, and X. Rao. Dynamische kapazitätsoptimierung durch automatisierung des bahnbetriebs. *ETR Eisenbahntechnsiche Revue*, 12, 12 2014.

Exploring the Potential of GPU Computing in Train Rescheduling

Sai Prashanth Josyula [1], Johanna Törnquist Krasemann [2], Lars Lundberg [3]
Department of Computer Science, Blekinge Institute of Technology, Sweden
[1] E-mail: sai.prashanth.josyula@bth.se
[2] E-mail: johanna.tornquist.krasemann@bth.se
[3] E-mail: lars.lundberg@bth.se

Abstract
One of the crucial factors in achieving a high punctuality in railway traffic systems, is the ability to effectively reschedule the trains when disturbances occur. The railway traffic rescheduling problem is a complex task to solve both from a practical and a computational perspective. Problems of practically relevant sizes have typically a very large search space, making it a challenge to arrive at the best possible solution within the available computational time limit. Though competitive algorithmic approaches are a widespread topic of research, limited research has been conducted in exploring the opportunities and challenges in parallelizing them on graphics processing units (GPUs). This paper presents a *conflict detection* module for railway rescheduling, which performs its computations on a GPU. The aim of the module is to improve the speed of solution space navigation and thus the solution quality within the computational time limit. The implemented GPU-parallel algorithm proved to be more than twice as fast as the sequential algorithm. We conclude that for the problem under consideration, using a GPU for conflict detection likely gives rise to better solutions at the end of the computational time limit.

Keywords
Real-time decision support, Train rescheduling, Conflict detection, Parallel algorithms, Graphics processing units.

1 Introduction

Scheduling is a frequently employed crucial operation in several sectors, e.g., manufacturing sector, railway transport sector, etc. In railway traffic network management, the ability to efficiently schedule the trains and the network maintenance, significantly influences the punctuality of trains and Quality of Service (QoS). The importance is reflected in the goal set by the Swedish railway industry stating that by year 2020, 95% of all trains should arrive at the latest within five minutes of the initially planned arrival time (Trafikverket, 2017).

In 2017, punctuality of rail passenger services in Sweden was recorded as 90.3% (Trafikverket, 2017). The punctuality of trains is primarily affected by (1) the occurrence of disturbances, (2) the robustness of the train timetables and the associated ability to recover from delays, along with (3) the ability to effectively reschedule trains within an allowable time interval, whenever disturbances occur, so that their consequences (e.g., delays) are minimized. This paper focuses on improving the ability to effectively reschedule trains during

disturbances.

Day-to-day train services in the rail sector are based on preplanned railway timetables. These timetables are planned to ensure that the services are feasible, i.e., the applicable constraints are respected. Typically, such constraints enforce safety by requiring a minimum time separation between consecutive trains passing through the same railway track. A disturbance in a railway network is an unexpected event that renders the originally planned timetable infeasible by introducing 'conflicts'. A conflict is considered to be a situation that arises when two trains require an infrastructure resource during overlapping time periods in a way such that one or more system constraints are violated.

Disturbances occur due to (1) incidents such as over-crowded platform(s) that possibly lead to unexpectedly long boarding times and minor delays, or (2) larger incidents such as power shortages, signalling system failures, train malfunctions that cause more significant delays. Train timetables are planned with appropriate time margins in order to recover from minor delays. Hence, in case of a minor disturbance, the affected train(s) may be able to recover from the effects of the disturbance provided there is sufficient buffer in the original timetable. In case of a disturbance that causes a significant delay to one or more trains, conflicts arise in the original timetable and it becomes operationally infeasible.

In order to resolve a conflict, the following rescheduling tactics are frequently employed: (1) Retiming, i.e., allocating new arrival and departures times to one or more trains, (2) local rerouting, i.e., allocating alternative tracks to one or more trains, (3) reordering, i.e., prioritizing a train over another, (4) globally rerouting the trains, or (5) partially/fully cancelling the affected train services. Detecting conflicts (i.e., checking the feasibility of the timetable) and resolving them (i.e., applying rescheduling tactics to obtain a feasible timetable) during operations, constitutes real-time railway traffic rescheduling.

During a disturbance scenario, given sufficiently large computation time, the best alternative rescheduled timetable can be chosen rather unambiguously, based on the goals of the decision-maker. However, in practice, the time interval available to reschedule the railway traffic and obtain a conflict-free rescheduled timetable at the time of a disturbance is quite narrow, e.g., 10–20 seconds (Bettinelli et al., 2017). Hence, it is a challenge to quickly explore the alternative desirable solutions and consequently reach the best alternative within the available time.

According to a recent survey (Fang et al., 2015), heuristic algorithmic approaches are most frequently employed by researchers to solve real-time railway rescheduling problems. Josyula et al. (2018) present a fast heuristic search algorithm based on iteratively detecting conflicts and resolving them using chosen rescheduling tactics. While solving the real-time railway rescheduling problem, the algorithm searches the *solution space* and produces feasible revised schedules of increasing quality with passage of time.

Though faster navigation of the solution space alone does not improve the quality of the final solution obtained by a heuristic algorithm, it very likely improves the quality of the final solution obtained within a computational time limit[1]. One way to improve the speed of solution space navigation is by designing parallel algorithms (e.g., Josyula et al. (2018)) suited for parallel hardware.

This paper presents a fast conflict detection algorithm for GPUs, which in turn results in a faster navigation of solution space. By speeding up the computation of alternative revised schedules, the most desirable schedule can be obtained by the end of the computational time

[1] assuming that the computational time limit $<$ time taken by the algorithm to obtain its final solution.

limit, thus resulting in efficient real-time railway rescheduling. The GPU-based conflict detection algorithm serves as a 'building block' for parallel train rescheduling algorithm(s).

The paper is organized as follows. The next section describes the problem at hand in more detail while overviewing the related research work. Section 3 presents a basic introduction to GPUs and explores the benefits and challenges of using them. It also presents a description of the algorithm for conflict detection (the CD algorithm) and its adaptation to GPUs (the CD-GPU algorithm). Section 4 includes the following: (i) description of the experiment used to evaluate the effects of incorporating GPUs in train conflict detection, and (ii) obtained results that comprise recorded execution times of conflict detection on central processing unit (CPU) and GPU. Section 5 analyzes and discusses the results of the experiments in order to infer valid conclusions.

2 Problem description and Related work

Optimization problems of practically relevant sizes often demand significant computational resources. Real-time railway rescheduling is one such problem that requires substantial computing capabilities to be solved to completion within an acceptable time. One of the key challenges in efficient rescheduling is to quickly explore the alternative desirable solutions in the solution space and consequently reach the best alternative within the permitted time.

Recent advances in computer hardware have made powerful chips, such as multi-core CPUs and GPUs, quite affordable and available even on commonplace computers. However, in order to employ such hardware in solving optimization problems, relevant and suitable algorithms (particularly designed and implemented for such hardware) are required. Typically, parallel algorithms are designed to employ (1) multiple processing units constituting modern CPU(s), and/or (2) GPU(s). In real-time railway (re)scheduling, the potential of parallel algorithms employing multi-core CPUs has been investigated in Mu and Dessouky (2011); Iqbal et al. (2013). More recently, Bettinelli et al. (2017); Josyula et al. (2018) report significant improvements in speed (without compromising solution quality) as a result of parallelization on CPUs.

Josyula et al. (2018) devise a train rescheduling algorithm that constructs and simultaneously navigates the branches of a search tree in parallel, as illustrated in Figure 1. The search tree is represented with conflicts as the nodes and rescheduling decisions as the edges. Each node also has a revised timetable associated with it; the root node corresponding to the original, disturbed timetable. The timetable of a subsequent child node is obtained by applying the rescheduling decision represented by its incoming edge on the parent node's timetable. The conflict represented by each node is obtained by (1) generating the node's timetable, (2) detecting the conflicts (using the CD algorithm) in the timetable, and (3) selecting the earliest of the detected conflicts. For a more detailed description of the parallel algorithm, see Josyula et al. (2018).

From Figure 1, it can be seen that conflict detection is a crucial operation that is frequently performed throughout the search tree exploration. Hence, attempts to speed up such an operation to attain faster search tree explorations, are well-justified. Initial trials to speed up conflict detection in the existing parallel algorithm by creating additional CPU threads proved unfavorable. The reason is that this resulted in the algorithm creating a large, non-optimal number of total CPU threads. However, other techniques to speed up conflict detection by employing alternatives to multi-core CPUs (e.g., GPUs) remain yet to be investigated.

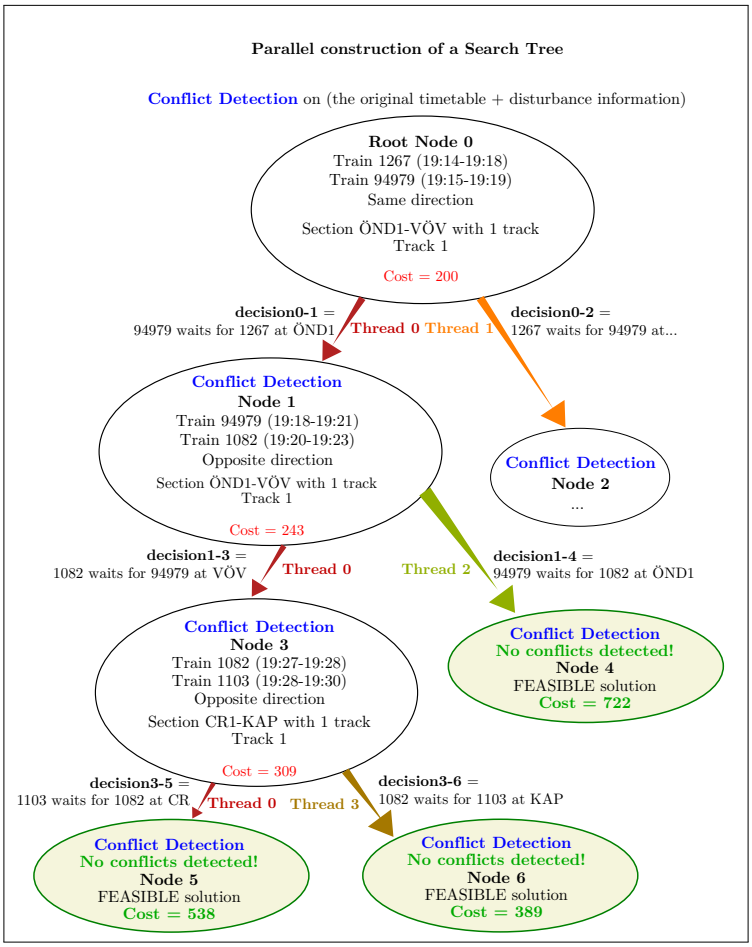


Figure 1: Illustration of the parallel algorithm designed by Josyula et al. (2018) through an example. The four parallel threads (0, 1, 2, and 3) explore the tree in parallel.

A parallel algorithm employing a GPU can either perform: (1) all of its computations on the GPU, while requiring little or no interaction with the CPU, e.g., Gmys et al. (2016), or (2) part of its computations on the GPU, while requiring significant CPU-GPU interactions. Several algorithms have been parallelized on GPUs for well-known optimization problems, such as the flow shop (Melab et al., 2012; Dabah et al., 2016), flexible job shop (Bożejko et al., 2010; Bożejko et al., 2012) and routing problems (Schulz et al., 2013). Inspired by the greedy algorithm in Törnquist Krasemann (2012), Petersson (2015) devised a building block for train rescheduling, which employs the GPU to explore multiple branches of the search tree in parallel. However, this building block spends significant time in exploring redundant solutions due to the design choices made in the search tree representation.

Very little attention has been given to employ GPUs to improve real-time railway re-

scheduling. Though commercial optimization solvers, such as Gurobi and CPLEX, make use of multi-core CPUs to solve a formulated model (e.g., a Mixed Integer Programming (MIP) formulation of the train rescheduling problem), currently, such solvers are not well-suited for GPUs (Glockner, 2015). For example, while solving a MIP model, each node of the search tree requires very different calculations (Glockner, 2015), whereas GPUs are designed for efficiently performing identical calculations on different data. The main objective of this research is to explore the potential of GPUs in solving the railway rescheduling problem. We did not come across research studies that answer the following research question:

How can a GPU be employed to improve computational decision support for real-time railway rescheduling?

This work contributes towards filling this research gap. The main contributions of the work presented in this paper are as follows:

(i) a building block for conflict detection on GPUs.

(ii) an evaluation of the effects of incorporating GPUs in railway rescheduling.

3 Exploring the benefits and challenges of using GPUs

A typical computer consists of a CPU as well as a GPU, both with significantly different architectures (Figure 2). A CPU is typically optimized for serial tasks, whereas a GPU is optimized for several parallel tasks. For example, consider the job of converting a color image to grayscale (Figure 3) wherein each color pixel described by a triplet of values (R, G, B) is to be converted to a corresponding grayscale pixel described by a single value that is computed by $\frac{(R+G+B)}{3}$. A GPU is highly efficient at completing this job by converting in

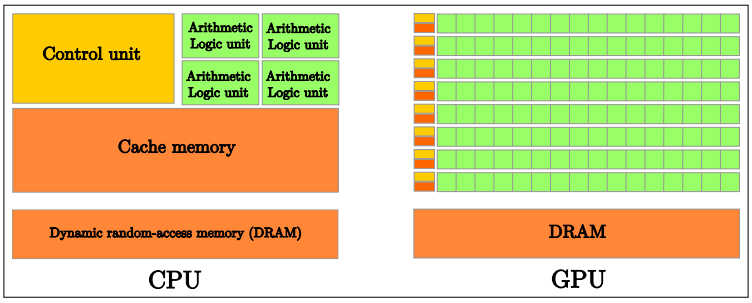


Figure 3: Conversion of a picture from color to grayscale.

context of optimization, see Brodtkorb et al. (2013). In the context of search trees, the computing power of a GPU can be utilized either for (i) parallel construction/exploration of the search tree (e.g., Petersson (2015)), or (ii) computations during tree construction/exploration (e.g., Melab et al. (2012)). The latter approach is well-motivated as the structure of the explored search tree is typically irregular, thus making tree exploration likely unfavourable for parallelization on GPUs.

In order to identify the computations worth parallelizing on a GPU, the performance reports of a previously profiled[3] heuristic algorithm for train rescheduling (Josyula et al., 2018) are examined. The results of profiling show that significant time is spent in conflict detection (the CD algorithm). While employing the algorithm to solve a rescheduling problem of moderate size[4] (i.e., a case study scenario in Josyula et al. (2018)), the conflict detection operation occurs around half a million times. Therefore, with an aim to speed up the detection of conflicts, we design a parallel algorithm for conflict detection on GPUs (the CD-GPU algorithm). Appendix A presents a code snippet[5] from the corresponding GPU program (also known as a 'kernel' in GPU terminology) implemented using the CUDA® framework (Fang et al., 2011).

Figure 4 gives an overview of the conflict detection on CPU (employing the CD algorithm) and on GPU (employing the CD-GPU algorithm) through an example. The railway infrastructure and timetable chosen for the example are illustrated in the figure. The graph adjacent to the timetable depicts that the latter is operationally infeasible and has three conflicts (labelled 1, 2, and 3). In order to detect these conflicts on a CPU, the *track event lists* are generated from the timetable, after which the CD algorithm is employed. When detecting these conflicts on GPU (by employing the CD-GPU algorithm), we instead generate *concatenated track event lists*. Then, the GPU threads, in parallel, detect the conflicts in the timetable (e.g., in Figure 4, ten threads, in parallel, detect three conflicts). In the next section, the effects of incorporating GPUs in train conflict detection are evaluated.

4 Experimental description

In order to explore the potential of GPU in solving the real-time rescheduling problem, we conduct experiments through which the speed of conflict detection on GPU is measured.

[3] using Intel® VTune™ performance profiler.
[4] 59 sections, 3-hour time window, initial delay due to disturbance = 25 minutes.
[5] The entire kernel is uploaded online and is publicly available (Josyula, 2019).

Algorithm 1: The CD algorithm for conflict detection on CPU

Input: Timetable T
Output: Set of detected conflicts

```
1  Generate track event lists from the timetable (see Figure 4).
2  foreach section j do
3      foreach track i of section j do
4          foreach pair of consecutive train events allocated to the track i do
5              if both the trains are in the same direction and
6              the section is a multi-block section then
7                  if Headway time constraint is violated then
8                      Conflict detected between the two train events on section j!
9              else
10                 if Clear time constraint is violated then
11                     Conflict detected between the two train events on section j!
```

Algorithm 2: The CD-GPU algorithm to detect conflicts on GPU (abridged version)

Input: Timetable T
Output: Set of detected conflicts

```
1  Sort the timetable array to generate concatenated track event lists (see Appendix B).
2  Create n threads to be executed in parallel, where n = length of the array T.
3  i = ID of the thread, i ∈ {0, 1, 2 ... n − 1}.
4  foreach thread except the last thread do
5      Event e_i = i^th element of the sorted array T.
6      Event e_{i+1} = i + 1^th element of the sorted array T.
7      if e_i and e_{i+1} are allocated to the same track of the same section then
8          if the trains are in the same direction and
9          the section is a multi-block line section then
10             if Headway time constraint is violated then Conflict detected!
11         else
12             if Clear time constraint is violated then Conflict detected!
```

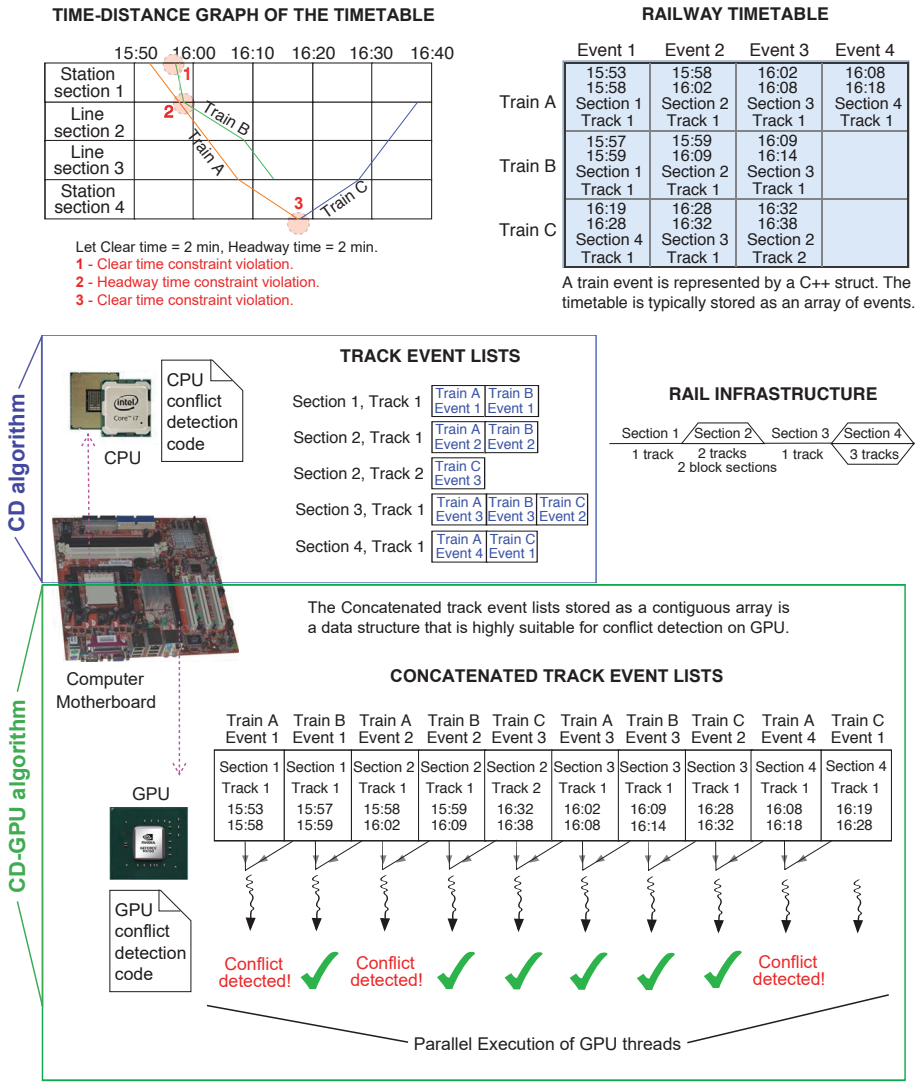


Figure 4: Summary of conflict detection on CPU vs GPU.

Prior to describing the experiments in depth, it is crucial to realize the following steps that are involved in the execution of a program that employs a GPU:

1. Allocation of required resources (e.g., global memory) on the GPU.
2. Transfer[6] of input data from CPU to the allocated memory in GPU.
3. Invocation of the GPU kernel that works on the input data and outputs results.
4. Transfer of results from the memory in GPU to the CPU.

4.1 Input data

Given an initial timetable T_{init} subject to a disturbance, the algorithm outlined in Josyula et al. (2018) generates, in parallel, alternative rescheduling solutions which are computed by iterating between conflict detection and conflict resolution (i.e., rescheduling of trains). We denote an intermediary rescheduling solution that is subject to conflict detection, T. Hence, the algorithm computes in parallel a set $\boldsymbol{T}$ of alternative rescheduling solutions. For instance, in the example run of the parallel algorithm (Josyula et al., 2018) shown in Figure 1, four rescheduling solutions are being generated in parallel (i.e., four branches of the tree are being explored in parallel). Therefore, corresponding to this example, the set $\boldsymbol{T}$ consists of four timetables. In other words, $|\boldsymbol{T}| = 4$.

The purpose of the experiments is to apply the GPU-based conflict detection on the set of alternative rescheduling solutions denoted $\boldsymbol{T}$. This is accomplished through the following three steps:

(i) transferring the set $\boldsymbol{T}$ from the CPU to the GPU,

(ii) detecting in parallel, conflicts in each timetable T, on the GPU,

(iii) transferring the results from GPU to CPU.

The potential of GPU can be best measured when the above steps (i)–(iii) are carried out a considerable number of times (e.g., 5000 times). This is taken into consideration while recording the execution times.

The size of results transferred in step (iii) is proportional to the size of the input data transferred in step (i); it is not related to the number of conflicts detected by the CD-GPU algorithm. The reason is that the results comprise values that correspond to each train event of the input data. These values indicate the presence/absence of a conflict along with its type (conflict due to violation of headway time constraint or clear time constraint). Similarly, the time taken for step (ii) (the CD-GPU algorithm) depends on the number of train events, not the number of conflicts in the input timetable(s). For instance, the CD-GPU algorithm requires equal execution time in the following two cases:

- to determine that an input feasible timetable has zero conflicts,
- to determine the number of conflicts in an input infeasible timetable.

[6]Typically, CPU communicates with GPU via high-speed bus called PCI express.

Due to the above reasons, the input data used throughout the experiments is generated in the following way. A feasible timetable T_{init} consisting of 740 train events is randomly chosen. When subject to a random disturbance of five minutes, 13 conflicts arise in T_{init}. This disturbed timetable consisting of 13 conflicts is used for populating the set $\boldsymbol{T}$ throughout the experiments. The railway infrastructure consists of 59 sections (including stations) and extends from Karlskrona to Tjörnarp.

4.2 Experimental variables

Variable	Description	Type (Independent, Controlled or Dependent)
$\lvert\boldsymbol{T}\rvert$	Number of timetables in the set $\boldsymbol{T}$.	This is an independent variable, the value of which is systematically changed.
t	Total number of times steps (i)–(iii) are executed. The value of t = 10,000.	The value of this variable is intentionally kept constant in order to clearly isolate the relationship between the other variables. This is the controlled variable.
c	Total number of times the conflict detection is performed ($\lvert\boldsymbol{T}\rvert \times t$).	This value is systematically changed to see its effect on the recorded measurements. This is the independent variable.
t_{gpu}	Time taken by GPU to perform conflict detection c times.	The value of this variable is observed and recorded. This is the dependent variable.

Table 1: Variables used in the experiments.

Table 1 lists the experimental variables and describes them in detail. As a benchmark for the recorded values of t_{gpu}, the associated conflict detection computations on the CPU are performed by:

(I) detecting conflicts in the chosen timetable T,

(II) recording the execution time (t_{cpu}) taken by the CPU to perform step I c times.

$$\text{Speedup } (S) = \frac{\text{Time taken by CPU to perform conflict detection } c \text{ times}}{\text{Time taken by GPU to perform conflict detection } c \text{ times}} = \frac{t_{cpu}}{t_{gpu}}$$

Note that each value of $|\boldsymbol{T}|$ in the performed experiments is intended to represent the number of branches of the search tree that a train rescheduling algorithm explores in parallel. Hence, the values are limited to $|\boldsymbol{T}| = \{1, 2, 4, 8, \ldots, 256\}$[7]; for practical problem scenarios, it is quite realistic to explore up to 256 branches of the search tree in parallel. The measurements for $|\boldsymbol{T}| = 512$ are recorded only to notice the trend of speedup.

4.3 Platform description

The experiments are performed on a laptop equipped with an Intel Core i7-8550U CPU and an Nvidia® GPU with compute capability 6.1. The GPU consists of 3 streaming multiprocessors (SMs), each with 128 cores. For detailed specifications of the GPU, see Appendix C.

[7]For the sake of convenience, we use only powers of 2.

The underlying operating system is 64-bit Windows® 10 Education and the available random-access memory is 16 GB[8]. The CPU code has been compiled using Microsoft® C++ optimizing compiler V19.14.26431, with whole program optimization (`/GL` flag) and maximum optimization favouring speed (`/O2` flag). The GPU code has been compiled using Nvidia CUDA compiler V9.2.148.

4.4 Kernel launch parameters

In an Nvidia GPU, the basic unit of execution is a *warp*, which is a collection of several threads. For devices with compute capability 6.1, a warp consists of 32 threads. All the threads in a warp are executed simultaneously by an SM; multiple warps can be executed on an SM at once.

A *block* of threads is a CUDA programming abstraction; all the threads in a block can communicate with each other (via shared memory, synchronization primitives, etc.) to co-operatively solve a problem in parallel.

In order to execute the conflict detection kernel on GPU, the *number of threads per block* and the *total number of blocks* need to be specified. These are known as kernel launch parameters. A frequently employed heuristic to select the number of threads per block is to aim for a high *occupancy*.

$$\text{Occupancy} = \frac{\text{number of warps running concurrently on an SM}}{\text{maximum number of warps that can run concurrently on the SM}} \tag{1}$$

The CUDA occupancy calculator (Nvidia, 2019) allows computation of the occupancy of a GPU by a given CUDA kernel.

For the GPU used in the experiments, the denominator of Equation 1 is 64. Compiling the conflict detection kernel with the compilation flag `--ptxas-options=-v` shows that it uses 25 registers per thread and 18960 bytes of shared memory per block. When this kernel resource usage is given as input to the occupancy calculator, Figure 5 is obtained as output. Based on this figure, the number of threads per block is chosen to be 512 in order to achieve 100% occupancy. The number of blocks to be launched is calculated using the following formula:

$$\text{Number of blocks } (b) = \frac{\text{Total number of threads}}{\text{Number of threads per block}} = \frac{\text{Total number of threads}}{512}$$

From Algorithm 2 and Figure 4, notice that the total number of GPU threads is equal to the total number of events involved in conflict detection. In the experiments, the latter number is supposed to be the number of events in set $\boldsymbol{T}$, which is $|\boldsymbol{T}| \times 740$. However, since $|\boldsymbol{T}| \times 740$ is not always an integral multiple of 512, the number of blocks are determined using the following formula:

$$\text{Number of blocks } (b) = \left\lfloor \frac{\text{Total number of events in set } \boldsymbol{T}}{512} \right\rfloor \tag{2}$$

Consequently, throughout the experiments, conflict detection on the GPU is not performed on all the events in the set $\boldsymbol{T}$. The last x events, where $x = (|\boldsymbol{T}| \times 740) \,\%\, 512$, are not sent as input to the GPU, and hence are not involved in conflict detection. The same events are excluded while performing conflict detection on the CPU.

[8] 1 kilobyte (KB) = 2^{10} bytes, 1 megabyte (MB) = 2^{10} KB, 1 gigabyte (GB) = 2^{10} MB.

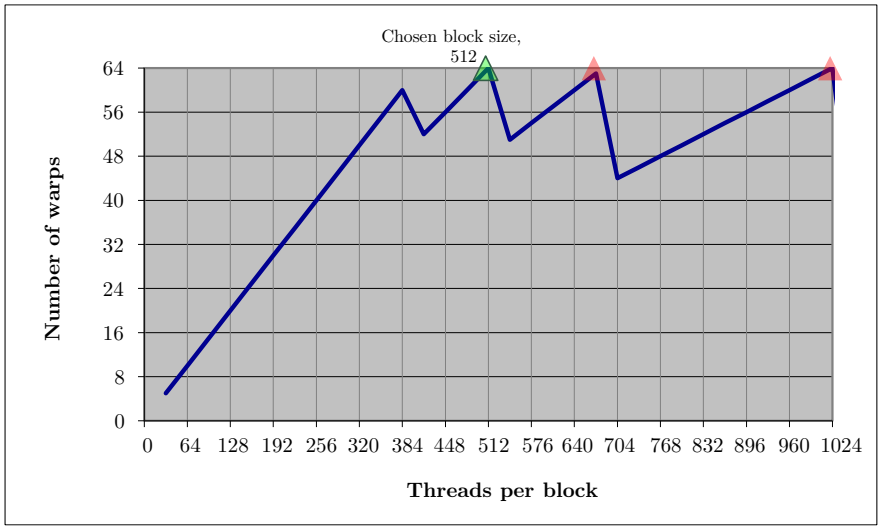


Figure 5: Impact of varying block sizes on multiprocessor occupancy.

4.5 Recorded results

The results of the experiments, summarized in Table 2, show that employing the GPU for conflict detection during real-time railway rescheduling can make the process more than twice as fast. Each recorded value of t_{gpu} and t_{cpu} is the average of five observations.

Explanation of the decrease in speedup value in Table 2: The total data[9] d transferred between CPU and GPU is proportional to $|\boldsymbol{T}|$. Through profiling the kernel, it was observed that the data transfer speed d_{speed} is not constant across different values of $|\boldsymbol{T}|$; for smaller values of $|\boldsymbol{T}|$ (consequently, smaller values of d), the d_{speed} is greater.

For example, for $|\boldsymbol{T}|$ = 1, d = 123 MB and d_{speed} = 6.3 GB/sec. For $|\boldsymbol{T}|$ = 2, d = 246 MB and d_{speed} = 5.7 GB/sec. For $|\boldsymbol{T}|$ = 256, 512, d = 45 GB and 90 GB, whereas the data transfer speeds are 3 GB/sec and 2.6 GB/sec respectively. This explains the fall in speedup (from 2.77 to 2.43) when the value of $|\boldsymbol{T}|$ is increased from 256 to 512.

5 Discussions and Conclusion

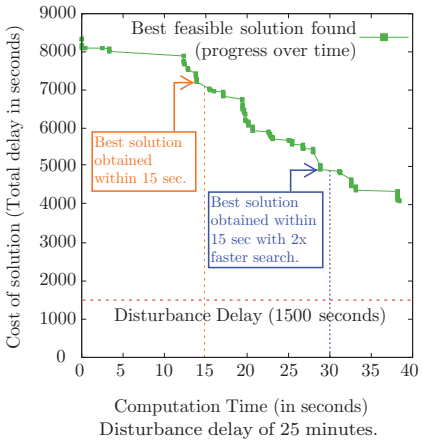
We present two examples (Figure 6) to illustrate the potential improvement (or the lack thereof) in the quality of solution due to faster search tree navigation. As can be seen in Figure 6a, a twofold faster search tree navigation leads us to obtain better solutions within a given computational time limit of, e.g., 15 seconds. However, in the disturbance scenario in Figure 6b, a twofold faster search tree navigation does not lead to a better solution within a time limit of 15 seconds.

GPUs possess the potential to speedup real-time railway rescheduling, thus improving

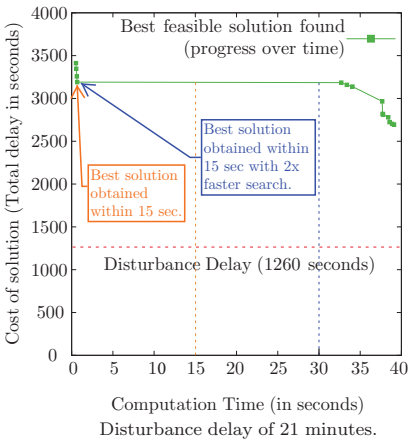
[9]Size of total data transferred = (Size of input data + Size of results) $\times$ 10^3

Number of			Number of		Time (sec)		
Time-tables in set $\boldsymbol{T}$	Times conflict detection (c)	Events in set $\boldsymbol{T}$	Events used for conflict detection	Blocks (b)	t_{gpu}	t_{cpu}	Speedup
1	1×10^3	1×740	1×2^9	1	1.23	0.22	0.18
2	2×10^3	2×740	2×2^9	2	1.45	0.42	0.29
4	4×10^3	4×740	5×2^9	5	1.47	0.88	0.60
8	8×10^3	8×740	11×2^9	11	1.66	1.87	1.13
16	16×10^3	16×740	23×2^9	23	2.49	3.14	1.26
32	32×10^3	32×740	46×2^9	46	3.50	7.33	2.10
64	64×10^3	64×740	92×2^9	92	6.02	15.06	2.50
128	128×10^3	128×740	185×2^9	185	10.81	29.17	2.70
256	256×10^3	256×740	370×2^9	370	19.03	52.75	2.77
512	512×10^3	512×740	740×2^9	740	40.73	99.20	2.43

Table 2: Results of conflict detection on CPU and GPU. For each measurement of t_{gpu}, steps (i)–(iii) are carried out 10^3 times. The number of events per timetable = 740, and the number of threads per block = 2^9.



(a) Example 1: Solutions obtained by a real-time train rescheduling algorithm during a disturbance scenario.



(b) Example 2: Solutions obtained by a real-time train rescheduling algorithm during another disturbance scenario.

Figure 6: Examples to illustrate potential improvement in quality of obtained 'best' solution due to faster search tree navigation.

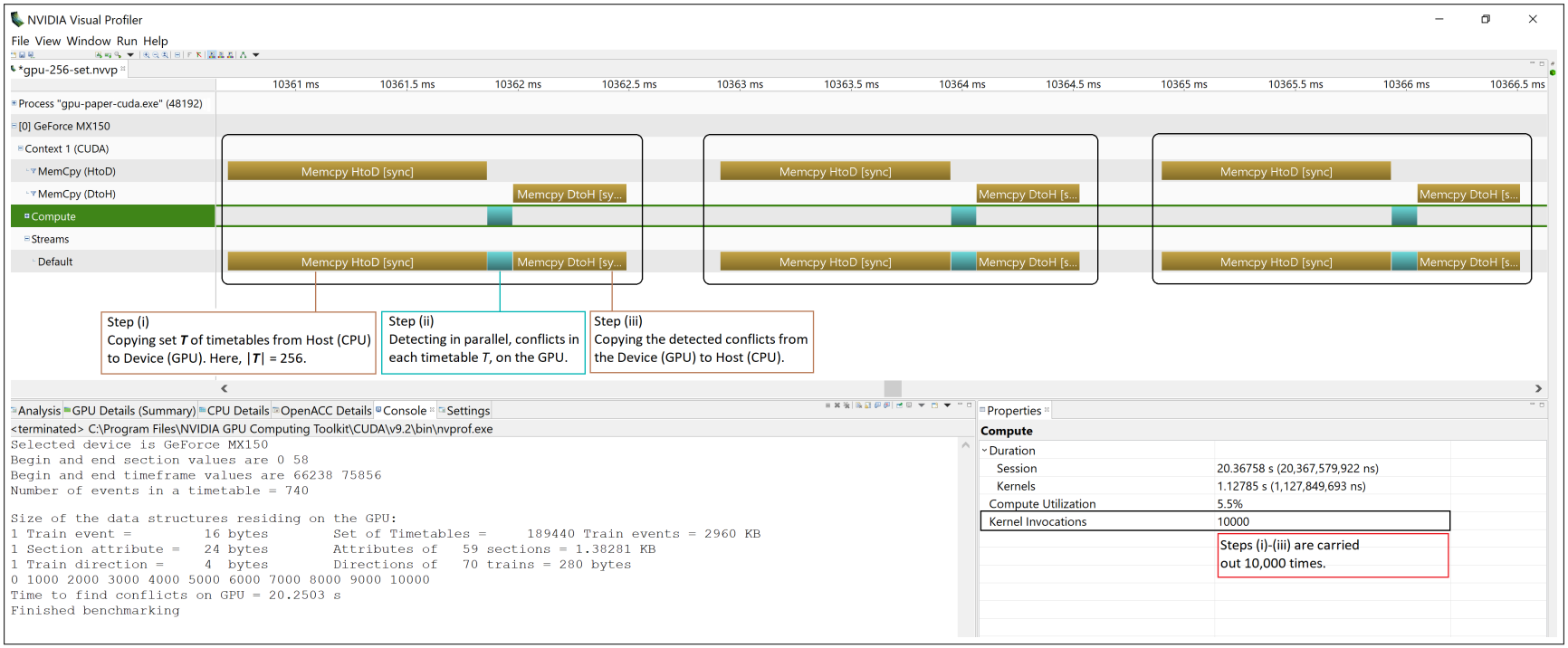


Figure 7: Output from Nvidia Visual Profiler when number of timetables in the set $\boldsymbol{T}$ = 256.

the likelihood of arriving at a better solution within the computational time limit. However, results of the experiments (tabulated in Table 2) show that this potential speedup (resulting from faster conflict detection using GPUs) requires several rescheduled timetables (i.e., $\geq$ 8) to be sent to the GPU in one transfer.

Profiling[10] the parallel program with the Nvidia Visual Profiler® shows (Figure 7) that for $\boldsymbol{T}$ = 256, only 5.5% of the recorded time (indicated by the parameter t_{gpu} in Table 2) is actually spent detecting conflicts. A major portion of the recorded time is spent on transferring data between the CPU and GPU, which is a demanding side-effect of using a GPU in frequent interaction with a CPU. Since Table 2 shows that the speedup of using the GPU (including communication time) for $\boldsymbol{T}$ = 256 is 2.77, the speed up attained in conflict detection on the GPU (excluding communication time) is $\approx \frac{2.77}{0.055}$, which is ≈ 50. Hence, conflict detection on GPUs is far more efficient than reflected by the speedup values in Table 2. This indicates that massive speedups could be achieved through solution approaches that execute the entire train rescheduling algorithm on a GPU (in contrast to the presented approach of executing only the conflict detection on the GPU). Such approaches would drastically reduce the CPU-GPU memory transfers which are significant bottlenecks in the presented approach.

Thus, we conclude that it is worthwhile to investigate modifications to existing real-time railway rescheduling algorithms (e.g., Josyula et al. (2018)) such that (i) several timetables are sent to a GPU for parallel conflict detection, or (ii) the algorithm is executed entirely on a GPU.

6 Acknowledgements

The research presented in this paper has been conducted within the research project TRANSFORM, which is funded by grants from the municipality of Karlshamn as well as the Swedish Research Council (FORMAS) via ERA-NET. The project has also received support from Trafikverket (The Swedish Transport Administration) and Blekingetrafiken.

The authors would like to thank Dr. Andrew Moss, Dr. Martin Boldt and Dr. Veronica Sundstedt for providing feedback that greatly improved the manuscript. Thanks to Dr. Lawrence Henesey for providing valuable feedback on the extended abstract of this manuscript.

References

Bettinelli, A., Santini, A., and Vigo, D. (2017). A real-time conflict solution algorithm for the train rescheduling problem. *Transportation Research Part B: Methodological*, 106:237 – 265.

Bożejko, W., Uchroński, M., and Wodecki, M. (2010). Parallel hybrid metaheuristics for the flexible job shop problem. *Computers & Industrial Engineering*, 59(2):323–333.

Bożejko, W., Hejducki, Z., Uchroński, M., and Wodecki, M. (2012). Solving the flexible job shop problem on multi-gpu. *Procedia Computer Science*, 9:2020 – 2023. Proceedings of the International Conference on Computational Science, ICCS 2012.

[10]Nvidia Visual Profiler is a cross-platform performance profiling tool which provides vital feedback to developers for optimizing CUDA C/C++ programs.

Brodtkorb, A. R., Hagen, T. R., Schulz, C., and Hasle, G. (2013). Gpu computing in discrete optimization. part i: Introduction to the gpu. *EURO Journal on Transportation and Logistics*, 2(1):129–157.

Dabah, A., Bendjoudi, A., El-Baz, D., and Aitzai, A. (2016). GPU-Based Two Level Parallel B B for the Blocking Job Shop Scheduling Problem. In *2016 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*, pages 747–755.

Fang, J., Varbanescu, A. L., and Sips, H. (2011). A comprehensive performance comparison of cuda and opencl. In *Parallel Processing (ICPP), 2011 International Conference*, pages 216–225. IEEE.

Fang, W., Yang, S., and Yao, X. (2015). A survey on problem models and solution approaches to rescheduling in railway networks. *IEEE Transactions on Intelligent Transportation Systems*, 16(6):2997–3016.

Glockner, G. (2015). Parallel and distributed optimization with gurobi optimizer. http://www.gurobi.com/resources/seminars-and-videos/parallel-and-distributed-optimization. [Last accessed 24-September-2018].

Gmys, J., Mezmaz, M., Melab, N., and Tuyttens, D. (2016). A gpu-based branch-and-bound algorithm using integer–vector–matrix data structure. *Parallel Computing*, 59(Supplement C):119 – 139. Theory and Practice of Irregular Applications.

Iqbal, S. M. Z., Grahn, H., and Törnquist Krasemann, J. (2013). Multi-strategy based train re-scheduling during railway traffic disturbances. In *Proceedings of the 5th International Seminar on Rail Operations Modeling and Analysis (RailCopenhagen 2013, pp. 387-405), Technical University of Denmark, Denmark*.

Josyula, S. P. (2019). Cuda code for conflict detection on gpus. https://github.com/sai-j/gpu-conflict-detection.

Josyula, S. P., Törnquist Krasemann, J., and Lundberg, L. (2018). A parallel algorithm for train rescheduling. *Transportation Research Part C: Emerging Technologies*, 95:545–569.

Melab, N., Chakroun, I., Mezmaz, M., and Tuyttens, D. (2012). A gpu-accelerated branch-and-bound algorithm for the flow-shop scheduling problem. In *Cluster Computing (CLUSTER), 2012 IEEE International Conference on*, pages 10–17. IEEE.

Mu, S. and Dessouky, M. (2011). Scheduling freight trains traveling on complex networks. *Transportation Research Part B: Methodological*, 45(7):1103–1123.

Nvidia (2019). Cuda occupancy calculator. https://developer.download.nvidia.com/compute/cuda/CUDA_Occupancy_calculator.xls.

Petersson, A. (2015). Train re-scheduling : A massively parallel approach using cuda. Master's thesis, Blekinge Institute of Technology, Department of Computer Science and Engineering.

Schulz, C., Hasle, G., Brodtkorb, A. R., and Hagen, T. R. (2013). Gpu computing in discrete optimization. part ii: Survey focused on routing problems. *EURO Journal on Transportation and Logistics*, 2(1):159–186.

Törnquist Krasemann, J. (2012). Design of an effective algorithm for fast response to the re-scheduling of railway traffic during disturbances. *Transportation Research Part C: Emerging Technologies*, 20(1):62–78.

Trafikverket (2017). The Swedish Transport Administration Annual Report. https://trafikverket.ineko.se/se/the-swedish-transport-administration-annual-report-2017. [Last accessed 24-September-2018].

A Fragment of code from the GPU program

Figure A: Code snippet from the GPU kernel

```
// 1D grid and 1D blocks
auto threadsPerBlock = blockDim.x;
// Local thread number in the block
auto l = threadIdx.x;
auto blockNumInGrid = blockIdx.x;
// Global thread number
auto i = blockNumInGrid * threadsPerBlock + l;

// Shared memory data structures for speed
__shared__ tr_event sh_concat_tracklists[1025];
__shared__ int sh_directions[128];
__shared__ sec_attribs sh_section_attr[128];

// Private variable (per GPU thread) to record the conflict
 ↪  event
int2 conflict;
conflict.x = -1;
conflict.y = -1;

// Copy the section attributes to the block's shared memory
if (l < numb_sections)
    sh_section_attr[l] = section_attr[l];

// Copy the train directions to the thread block's shared
 ↪  memory
if (l < numb_trains)
    sh_directions[l] = directions[l];

// Copy a 'block' of sorted section lists to shared memory
sh_concat_tracklists[l] = concat_tracklists[i];
// For the last thread in the block
if(l == threadsPerBlock - 1)
    sh_concat_tracklists[l+1] = concat_tracklists[i+1];

// Ensure all writes to shared memory are completed
__syncthreads();

// Other code not included in this snippet

// Coalesced copy the detected conflict to global memory
conflicts[i] = conflict;
```

B Efficient generation of concatenated track event lists

Concatenated track event lists (for use by the GPU) can be efficiently generated from a timetable by sorting it using the following logic.

Figure A: Sorting logic for the train events comprising a timetable

```
sort (event1, event2)
{
  if(event1.section == event2.section)
  {
    if(event1.track == event2.track)
      // Sort based on begin times.
    else
      // Sort based on track numbers.
  }
  else
    // Sort based on section numbers.
}
```

C Detailed specifications of the GPU used in the experiments.

Property	Value
Number of streaming multiprocessors	3
CUDA cores per multiprocessor, total cores	128, 384
Number of threads per warp	32
Maximum warps per multiprocessor	64
Maximum blocks per multiprocessor	32
Maximum threads per multiprocessor	2048
Maximum threads per block	1024
Register size, registers per multiprocessor	32 bit, 65536
Maximum registers per block	32 bit, 65536
Maximum registers per thread	255
Register allocation unit size	256
Register allocation granularity	warp
Shared memory allocation unit size	256
Warp allocation granularity	4
Maximum shared memory per block	48 KB
Shared memory per multiprocessor	96 KB
Constant memory	64 KB
Global memory	2048 MB

Table 3: Physical limits of the GPU used in the experiments.

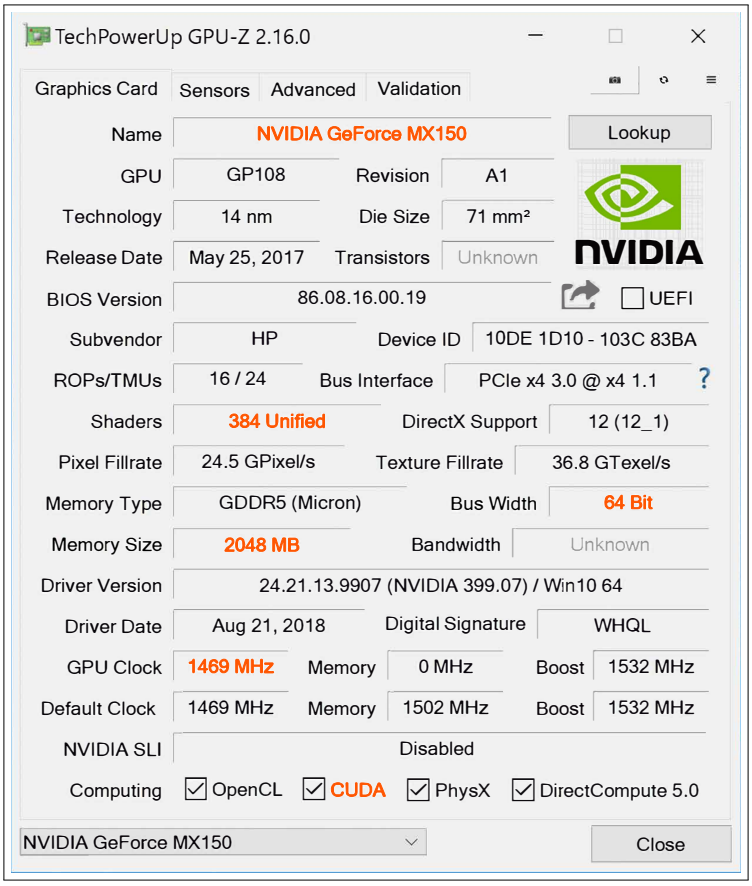


Figure 8: Further detailed specifications of the GPU.

A Graph Application for Design and Capacity Analysis of Railway Junctions

Predrag Jovanovic [a,1], Norbert Pavlovic [a], Ivan Belosevic [a] Sanjin Milinkovic [a]
[a] Faculty of Transport and Traffic Engineering, University of Belgrade
Vojvode Stepe 305, 11000 Belgrade, Serbia
[1] p.jovanovic@sf.bg.ac.rs, Phone: +381 63 475829

Abstract
In this paper, we developed an analytical model for strategic decision making, for selection of the best solution of the junction layout according to the maximum theoretical infrastructure capacity, completely independent of the timetable. Model achieves triple effects as it enables the selection of the most favorable route sequence, as well as the theoretical capacity calculation. The model uses well known combinatorial problems on graphs, Weighted Vertex Coloring Problem (WVCP) and Traveling Salesman Problem (TSP) to determine the minimum time of the infrastructure occupancy. The model is tested on three different junction layouts.

Keywords
Railway Junction, Capacity, Weighted Vertex Coloring Problem, Traveling Salesman Problem,

1 Introduction

In the recent years, the capacity utilization on the main railway lines and corridors has been increasing. Modern trends in strategic policy such as the opening of a railway market and the appearances of new railway operators led to increase in the number of trains and the capacity of the railway infrastructure has become a bottleneck for the entire railway system. Consequently, there is a decline in the quality of transport service due to the occurrence of train delays.

Railway infrastructure is the most expensive subsystem of the entire railway system. However, the maximum utilization of railway infrastructure capacity should not be the ultimate aim. A high value of the infrastructure capacity utilization coefficient leads to train delays, as well as an exponential increase in these delays (Yuan and Hansen (2004), Landex (2008)). Furthermore, train delays cause a drastic reduction in the quality of transport services. As a result, there is a demand for the construction of new railway lines, as well as for the reconstruction and modification of existing ones.

The term ”railway infrastructure capacity”, in academic and especially in professional publications, mainly refers to the capacity of railway lines. Existing methods, such as UIC 406 (Union International des Chemins de Fer - UIC (2013)), focus on the calculation of railway track capacity, while capacity issues addressing railway nodes are considered as specific cases. However, junctions and stations as nodes in railway networks are essential to the entire railway line capacity evaluations. The capacity of junctions is a complex param-

eter and its calculation is a difficult task primarily due to various train movements that are allowed to be set through a switching area. In such situations, some train routes are compatible and can be executed simultaneously, whereas other train movements are not compatible and have to be separated by a time interval. The minimum time intervals between two successive but incompatible train movements differ depending on the sequences of train route realizations.

Permanent development in computer science and technologies put forwards simulation methods as a reliable approach for evaluating railway capacity. Simulation methods enable the representation of dynamic behavior of a rail traffic system duplicating its real-world operations. Basically simulation models are categorized as macroscopic (e.g. Kecman et al. (2013)) or microscopic (e.g. Nash and Huerlimann (2004) or Radtke and Hauptmann (2004)) models. However, simulation methods have to be adapted to each specific application environment requiring a large amount of preprocessing input data. It could be extremely difficult to collect all required input data, especially for conception solutions characterized with imprecisely defined infrastructure (either regarding track layout or interlocking components) or timetable data. In contrast, analytical methods present a convenient approach aimed to preliminary evaluate capacity of different conception solutions and to identify bottlenecks. Analytical methods utilize mathematical expressions to obtain theoretical upper bound on capacities. Main advantages of analytical methods are fast and simple calculations that provide sufficiently accurate results.

Analytical methods that address capacity evaluations of railway nodes are presented in Malavasi et al. (2014) referring to the mathematical expressions given by Potthoff (1980), Corazza and Musso (1991) and guidelines provided by German railways from 1979. In addition to these simple analytical approaches, Huisman et al. (2002) proposed an analytical approach for the analysis of railway nodes based on the queuing theory. Yuan and Hansen (2007) proposed a stochastic model for train delay propagation that could be used to estimate capacity utilization. Lindner (2011) presented the application of UIC 406 method for station capacity evaluations. The UIC approach was adopted by Landex and Jensen (2013) to analyze capacity at stations with simple track layouts. Also, authors proposed additional measures to analyze and describe track complexity and robustness of train operations. The similar topic on understanding the relationships between capacity utilization and performances of railway stations and junctions is analyzed by Armstrong and Preston (2017). Finally, Jensen et al. (2017) expanded the UIC approach to calculate infrastructure utilization in networks, considering different sequences of a train route realization and their dependence on the infrastructure occupation. As authors stated, the approach is ideal for strategic planning providing the evaluation of different infrastructure solutions.

In this paper, we developed an analytical model applicable for design and capacity analysis of railway junctions. The proposed method determines the sequence of train routes that guarantees the lowest capacity utilization. Based on the proposed approach, it is possible to compare different junction layouts determining the capacity utilization coefficient for each of them. The model is developed as a reverse approach to the graphic Potthoff model. Its main advantages are simplicity and the fact that the model does not require train schedules (timetables). For input data, the model requires only conceptual solutions with defined sets of feasible train routes characterized with the average duration of train routes and mean time intervals between each of them.

2 Problem description and model formulation

The term capacity of the railway infrastructure includes the number of train movements that can be realized in the considered time. The calculation process of the line capacity between two stations involves determining the exact line occupation time by all trains. The time obtained in this manner is used to calculate the utilization coefficient of the railway infrastructure. However, during the calculation process of the capacity of junctions, this procedure becomes significantly complicated, primarily because some train routes can be realized simultaneously with some other routes.

The model proposed in this paper requires the construction of a route compatibility matrix in the first step, as in most of the previously described models. In addition, the model uses a graphical interpretation similar to the Potthoff model. After the construction of the route compatibility matrix, the graph should be constructed such that every possible train movement should be presented as a vertex. An example of junctions used for a detailed description of the model is taken from (Pachl (2004)) as shown in Figures 1 and 2. In these figures, the letters represent the start and end points of the considered routes.

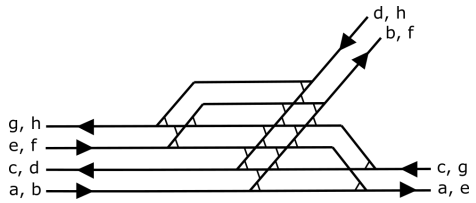


Figure 1: ”Inferior” design of the example junction

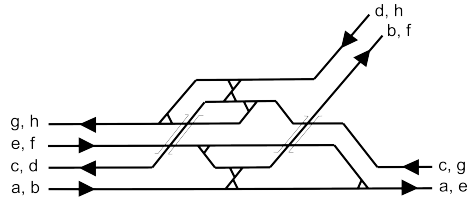


Figure 2: ”Improved” design of the example junction

Based on the provided example junction, in the first step, the matrices of compatible train routes should be constructed. The compatibility matrix is formed by assigning a ”+” sign to the element of matrix $c_{i,j}$ if routes *i* and *j* are compatible with each other. Conversely, the ”-” sign is assigned to the element of matrix $c_{i,j}$ if routes *i* and *j* are incompatible with each other. At the same time, the matrix of minimum time intervals should be created, in such a way that for each element in the compatibility matrix with sign ”-”, for each pair of routes, one calculate and enter the value of the minimum time interval since previous route releases the last joint infrastructure element, until the moment when a consecutive route can start.

Now, the model is developed on the basis of a simple variation in graphical interpretation of the Potthoff method: each possible route is represented as a vertex of the graph, and the edges link the vertices that represent *mutually incompatible* routes, i.e., those train movements that cannot be executed simultaneously. Thus, the graph $G = (V, E)$ is constructed, where V represents a set of vertices, and with E a set of edges are marked. The graph defined in this manner is complementary to that defined by the original Potthoff method (Pachl (2004)). For the junctions presented in 1 and 2, the constructed graphs are shown in Figures 3 and 4 for the "inferior" and "improved" layouts, respectively.

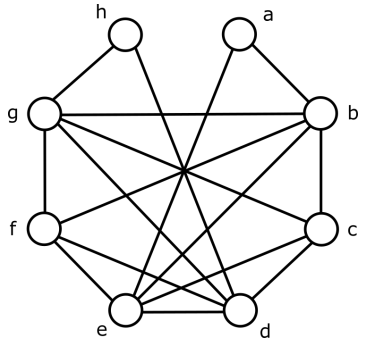


Figure 3: Graph of incompatible train routes for "inferior" layout of the example junction

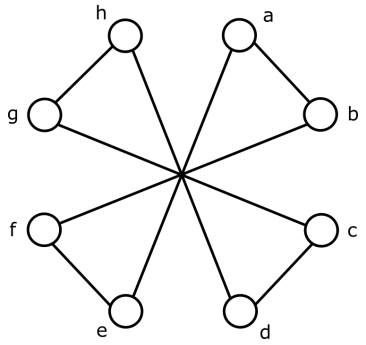


Figure 4: Graph of incompatible train routes for "improved" layout of the example junction

Keeping in mind the rule that in one moment in time, one infrastructure segment can be allocated to only one train movement, the next question can be asked: how to execute all intended routes in such a way that each train movement must be performed at least once and that there is no collision between any two train routes?

Let S denote the set of all infrastructure segments in the switching area and V the set of all possible train routes through the considered switching area. For any train route x, S_x is a set of infrastructure segments that will be occupied during the realization of route x, at least in one moment. If y denotes another route, then we will call x and y incompatible routes if

they cannot be executed simultaneously, i.e., if they must be separated in time, if and only if it is valid

$$S_x \cap S_y \neq \emptyset \tag{1}$$

Nodes of graph G, which are linked by an edge, represent train routes that require at least one "common" element of the infrastructure.

2.1 Weighted Vertex coloring-based approach to junction design analysis

In graph theory, the coloring of a graph is a simple marking of the graph's elements. Similar to the coloring of edges, researches have dealt with the problem of vertex coloring, the problem that we use in our model. The vertex coloring problem (VCP) assumes that each vertex (node) is attributed by a certain marking (color), such that two neighboring vertices, i.e., vertices connected by an edge, cannot have the same marking (color). Formally, if we denote *K=(1,...,m)* as a set of markings (colors), the problem of the vertex coloring for graph G, with m colors, is mapping $C : V \rightarrow K$. The graph is correctly colored for

$$c(i) \neq c(j), \forall \{i, j\} \in E. \tag{2}$$

The smallest number of colors that is sufficient for a graph to be correctly colored is defined as a chromatic number of graph G and is marked as $\chi(G)$. Graph G is $k-colored$ if it is not $(k-1)-colored$. The graph coloring is optimal if all vertices are colored and if k is a minimal number of colors that can be used to color the graph. Although, complexity of the chromatic number computation is known to be NP-hard, for every $k > 3$, a $k-coloring$ of a graph exists by the so called "four color theorem", and it is possible to find such a coloring in polynomial time.

VCP can be modeled by integer linear programming. First, we define two sets of binary variables:

- x_{ij} - a variable that defines whether the marking (color) j is assigned to vertex i; the variable has value 1 if and only if color j is assigned to vertex i,
- y_j - a variable that defines whether the marking (color) j is used in the process of mapping; the variable has a value 1 only if color j is assigned to at least one of the vertices.

The goal is coloring all vertices of the graph using the minimal number of colors; that is, to establish a chromatic number of the graph, the objective function is defined as

$$min \sum_j y_j \tag{3}$$

with a set of constraints

$$\sum_j x_{ij} = 1,\ i \in V \tag{4}$$

$$x_{ij} + x_{kj} \leq 1,\ \forall (i,k) \in E,\ j = 1, ..., n \tag{5}$$

$$x_{ij} \in \{0, 1\},\ i \in V,\ j = 1, ..., n \tag{6}$$

$$y_j \in \{0,1\},\ j = 1, ..., n. \tag{7}$$

If we apply a VCP on previously described graphs of incompatible routes, the chromatic number of a graph, i.e., the number of used colors for an optimal coloring of incompatible routes graph, will represent a minimal number of the groups of routes that should be formed so that each route is performed exactly once. All vertices that are marked with the same color belong to a set of routes that are mutually compatible and can be executed simultaneously. Colored graphs of "inferior" and "improved" designs of the switching area are shown in Figures 5 and 6.

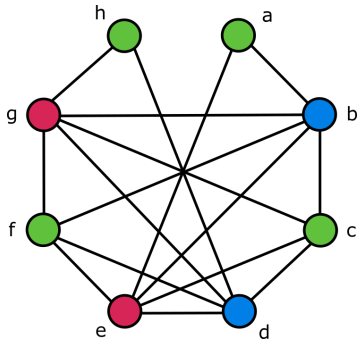


Figure 5: Colored graph of incompatible routes for "inferior" design of switching area

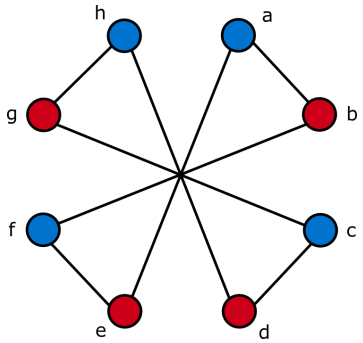


Figure 6: Colored graph of incompatible routes for "improved" design of switching area

For the realization of each set of mutually compatible routes, one after another, in several iterations, each of the defined routes will be completed. Now, it can be confirmed that through the analysis of "inferior" and "improved" designs of the switching area, all routes for the "improved" design can be executed in two iterations, while for the "inferior" design, for completing all routes, we need to form at least three sets of mutually compatible routes.

Based on such a simplified approach for presenting a problem, a model will allow a creative analysis for the layout of the switching area, according to a possible number of required sets for exactly one execution of each route.

In the previously described model, graph coloring does not consider time for route execution, but only their mutual compatibility. A consequence of such model application leads to the generation of the so-called ”unproductive” times. ”Unproductive” time represents a time elapsed from the end of one route within one set of mutually compatible routes (within vertices in one color) until the end of the longest route of the same set. In situations where it is possible to color a graph in more than one way, the time difference between the moments of finished routes and that when the route that needs maximum time to finish is over and belongs to the same set of compatible routes, it is considered as an unproductive time. Even with a previously introduced constraint which imposes that all routes from the next set start their execution simultaneously, after the competition of all defined routes, there is a ”lost” time. To fully understand unproductive and lost time, let us assume that we observe some junction and it is possible to define five routes and that these routes can be grouped in several ways – in Figure 7, there is a diagram of the time distribution.

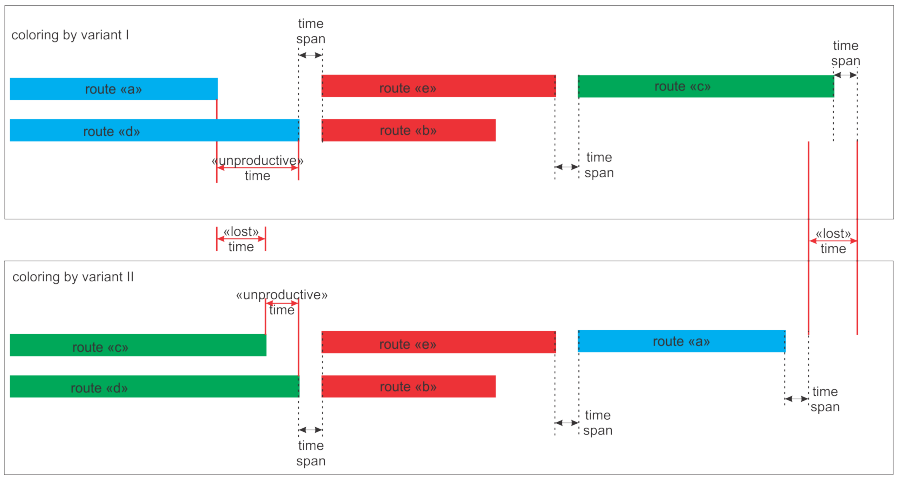


Figure 7: Two alternatives of the Gantt diagram of train routes when graph coloring for incompatible routes is possible in many ways

As presented in Figure 7, ”lost” time is the difference between ”unproductive” times within different sets. Due to the constraint imposed by the simultaneous start of the routes within the next set, ”unproductive” time cannot be eliminated and ”lost” time is generated as an extension of total time of the switching area occupied by all routes.

To reduce the produced negative effects, in the process of coloring the incompatibility graph, it is necessary to group the routes where the time difference between the longest route and a previous route is the smallest within the same set. This can be achieved by assigning each vertex j of graph G a nonnegative value w_j^v. The value of w_j^v is a weight of vertex j, and in the model, it represents the execution time of a route j.

The weighted vertex coloring problem (WVCP) is an extension of the basic graph VCP, where the basic principles of graph coloring are the same. Connected vertices of the graph should be assigned different colors, by defining a minimization of the sum of the cost for the used colors as an objective function. The cost of the used colors is the maximum value

of the vertex weight coefficients that were assigned the same color (Malaguti et al. (2009); Furini and Malaguti (2012)). WVCP is known to be NP-hard.

The model is based on the assumption that the graph vertex weight coefficients w_j^v, $\forall$ $j \in V$, are nonnegative integer values. However, without lack of generalization, we can consider them as real values, ordered by descending values. The model is then shaped as mixed integer programming, as we define the following two sets of variables (Malaguti et al. (2009); Malaguti (2009)):

- x_{ij} - a binary variable with a value of 1 if and only if the color j is assigned to vertex i,
- z_j - a real variable that has a value of the cost for color j.

Now, we can define a basic model with the objective function

$$min \sum_j z_j \tag{8}$$

and constraints

$$z_j \geq w_j^v \cdot x_{ij},\ i \in V,\ j = 1, ..., n \tag{9}$$

$$\sum_j x_{ij} = 1,\ i \in V \tag{10}$$

$$x_{ij} + x_{kj} \leq 1,\ (i,j) \in V,\ j = 1, ..., n \tag{11}$$

$$x_{ij} \in \{0, 1\},\ i \in V,\ j = 1, ..., n. \tag{12}$$

In the defined model, relation (8) is an objective function, constraint (9) defines a cost for each color, and (10) formulates a demand that all vertices must be assigned a color. Constraint (11) represents a basic limitation of the graph VCP, i.e., the neighboring vertices cannot be assigned the same color, while (12) defines a binary variable x (Malaguti (2009); Malaguti et al. (2009)).

As opposed to the basic graph VCP, the solution for WVCP does not have to provide an optimal graph coloring, according to a chromatic number of the graph, $\chi(G)$. Hence, it is possible to group mutually compatible routes in a larger number of groups than it would be minimum necessary, with an assumption that vertex weights are defined as a time to perform certain routes represented by vertices. The model objective function gives the shortest occupation time for the junction *only by time for the completion of a routes*. By each increase in the number of different sets of compatible routes, the total occupation time of the junction is increased by a necessary time interval between each newly added set and its predecessor set of compatible routes. Therefore, through the application of the WVCP model, improvement is evident only if the solution is optimal by the defined objective function (8) as well as by the objective function (3). For this reason, the final number of groups is adopted from the results of VCP. After that, in the case of a different manner of combining routes obtained by VCP and WVCP, in order to improve the results we accept the WVCP solution.

An improvement that is imposed by the application of the WVCP model is a consequence of the comparison of grouped compatible routes with the longest route within the same set while ignoring the "short" routes within a set. However, besides in extreme situations, this will not affect the result.

2.2 Weighted Vertex coloring-based approach for capacity determination

To determine the capacity of a junction, it is necessary to define the time needed for the realization of all routes assuming that each route is realized at least once. Furthermore, we assume that the realization of all routes within a single group is simultaneous and that it starts once all infrastructural and rail operational conditions are met. The assumption that all routes within the same group of mutually compatible routes begin its realization simultaneously allows the formation of a simplified graph, $D(V', E')$. In this simplified graph, vertices are groups of mutually compatible routes, defined by the solution of the WVCP model (relations (8)-(12)). In such a graph, "compatible groups" cannot exist because they would be returned as a joined group by the WVCP model. Thus, the graph created is a complete graph with edges between all pairs of vertices. Now the weight coefficient of the edge is introduced as the maximum value of the required interval between the longest route in group i and all routes within group j of mutually compatible routes, $\tau_{i,j}$:

$$w_{ij}^{e} = max\ \tau_{ij}, \forall (i,j) \in V',\ i \neq j. \tag{13}$$

However, as the minimum necessary time interval between incompatible routes does not have to be equal and most often is not, there are two possibilities. First, a higher value is chosen for the weight coefficient of the edge:

$$w_{ij}^{e} = max\ (w_{ij}^{e}, w_{ji}^{e}). \tag{14}$$

The second possibility, which is used in this paper, imposes the formation of independent edges for each of these two intervals. In this way, the model defines a graph of "incompatible groups of routes" creating a complete digraph, i.e., a directed graph with a pair of edges between all pairs of vertices.

Besides the weight coefficients of the edges, those of the vertices can be assigned to graph D as the maximum realization time of the routes that are grouped together. Bearing in mind the assumption that all routes within one group start simultaneously, the duration of the realization of all routes within one group of mutually compatible routes will be equal to that of the longest route within that group. If we assume that t_{rj} is the duration of a route j in group r, the realization time of all routes from that group will be the same:

$$w_{j}^{r} = \max_{j}\ t_{rj}. \tag{15}$$

To determine the most favorable sequence in which the routes will be executed, it is necessary to first determine the order of the groups of mutually compatible routes. In addition, to determine the capacity of the entire switching area, it is necessary to determine the total time of occupation of the switching area through the realization of all routes when each of them is realized exactly once. Given the characteristics of the defined graph D, both problems can be solved by finding the shortest Hamilton cycle in graph D. The problem of finding the shortest Hamilton cycle, if there is one, is known as the traveling salesman problem (TSP), the famous combinatorial problem, from the NP-complete class. In order to allow periodic repetition of the most favorable sequence throughout observation period, we need to determine Hamiltonian cycle, i.e. Hamiltonian path would not be sufficient for total occupation time determination.

The most favorable sequence in which the routes will be executed is gained by determining the order of realization of groups of mutually compatible routes, as a solution to the

shortest allowed Hamilton cycle, while the total time of occupying the switching area, T_g^s, will be equal to the sum of the solution of TSP problem and the sum of realization times of the longest routes within each group. According to relation (8), the sum of the realization times of the longest routes within each group equals

$$\sum_c w_c^v = T_{WVCP} = min \sum_j z_j. \tag{16}$$

Thus, the total occupation time of the switching area T_g^s by all routes and all necessary time intervals between them equals

$$T_g^s = T_{WVCP} + T_{TSP}. \tag{17}$$

The coefficient of utilization is defined as the ratio of the total occupation time T_g^s and observation time U

$$\eta = \frac{T_g^s}{U}. \tag{18}$$

On the other hand, the total theoretical number of routes N_r that can be executed during a certain period U is defined as

$$N_r = \frac{U}{T_g^s} \cdot \nu \tag{19}$$

where ν signifies the total number of defined routes in the switching area, i.e., the sum of all routes from all groups.

In this way, the model can be used not only for the design analysis of switching areas but also for determining the most favorable sequence of route realization and for approximate capacity determination. The approximate capacity of the switching area, i.e., the maximum number of routes in the observed switching area, can be determined exclusively with the assumption that the traffic pattern, i.e., the specified order of route realization, is unchangeable.

The formed direct graphs, after applying WVCP on the aforementioned examples for ”inferior” and ”improved” track layout designs, are shown in Figures 8 and 9, respectively. The determination of vertex weight coefficients as the maximum duration of route realization within each group is shown in red text, while the procedure of determining the weights of the edges is shown next to each edge.

Considering the developed model, it is easy to compare the two junction layouts, both in terms of the number of simultaneous routes and from the aspect of determining the most favorable sequence of route realization and determining the total capacity.

2.3 Model expansion to achieve demanded route sequences and to deal with heterogeneity

In the case of a timetable with an unequal number of routes from and for different directions, i.e., when some of the routes should be executed more often than other train movements, these routes must be presented as distinct vertices in the graph. Moreover, they have to be connected by edges with all vertices that their base routes are connected with, including the additional edge to the base route. All such ”additional” routes entered into the graph

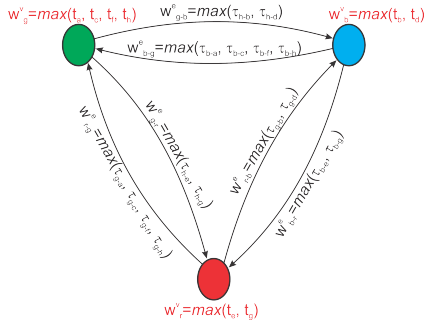


Figure 8: Reduced direct graph coloring of incompatible rides for ”inferior” design of the switching area

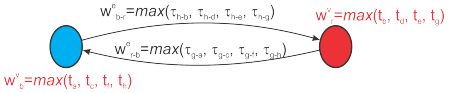


Figure 9: Reduced direct graph coloring of incompatible rides for ”improved” design of the switching area

as distinct vertices and have all characteristics of their base routes. Moreover, they have to respect the same compatibility rules with other routes, with which they are also in conflict. A graph for a case with an unequal number of routes for/from different directions (a and c represent base routes that should be realized twice as often as the rest) and for the ”inferior” design of the switching area is shown in Figure 10. Since the execution of all routes, including additional routes a' and c' represents a cycle, the order of routes in the cycle can be changed, i.e., in the vertex coloring process, additional routes are equal to their base routes, so it is possible to change the execution sequence, as shown in Figure 11.

On the other hand, a case may arise where, with the change in the frequency of certain route realizations, certain limitations concerning the order of their execution are required. Namely, when a certain base route has a higher realization frequency than others, e.g., route a in Figures 10 and 11, there is no logic to allowing successive realization of two, or even more, same routes, especially in case of passenger trains. Actually, it is necessary to introduce additional restrictions in TSP, preventing the procurement of an optimal solution with the adjacent vertices of the same route. At the lowest level, this can be achieved by the removal of edges from digraph $D(V', E')$ that connect ”critical” groups of routes.

Besides the abovementioned case, the requirements for the successive execution of individual routes may occur, especially in the case of passenger trains, in order to obtain connections for the transfer of passengers from one train to another. As in the previous case, simply by modifying the digraph $D(V', E')$, it is possible to impose the successive realization of the two groups of routes, but this time, by forcing the path, from one vertex into another, i.e., through the existence of obligation of a particular edge in the TSP problem solution.

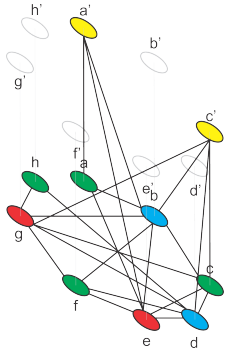


Figure 10: Incompatibility graph for additional routes and different frequency – alt. I

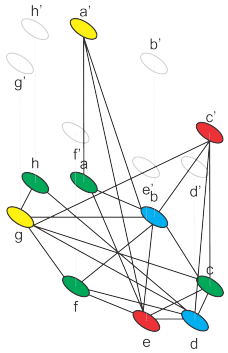


Figure 11: Incompatibility graph for additional routes and different frequency – alt. II

The process of determining the total occupancy time of the junction for a cycle period remains completely unchanged - if there is a change in the number of groups of simultaneous routes, they are equal with other groups, so the algorithm should be applied entirely. Ideally, routes with a higher frequency can be realized simultaneously with the routes of another group, so the graph will accordingly be colored.

In cases where it is predicted that an identical route is carried out by trains whose paths in the timetable are different, i.e., in the case of heterogeneous traffic, as well as in the case of different route frequencies, the vertices of routes using identical parts of the infrastructure but different technical parameters (running speed, train length, etc.) are added to the graph of mutually incompatible routes, while the mutual relations with remaining routes in the incompatibility matrix do not change.

3 Case study and result analysis

For complete application and testing of the defined model, we created three different track layout alternatives for flaying (or grade separated) railway junction. The examined railway

junction has a track configuration in which two main double track railway lines cross each other by a bridge to avoid conflicts of their 4 main routes (a, b, c, d). Furthermore, all three alternatives have track connections that enable additional 8 routes for crossing trains over both railway lines in both directions (e, f, g, h, i, j, k, l). However, the alternatives differ in the complexity of their track layouts expressed either in the number of installed switches, diamond crossings or bridges. The applied track layout directly influences the compatibility of train routes.

Alternative 1 - a basic layout that provides single track connections required to enable trains to cross over railway lines. The track layout consists of two main double track lines, 4 single track connections with installed 24 switches. The layout provides 52 compatibilities among the observed 12 routes. This junction layout is shown in Figure 12

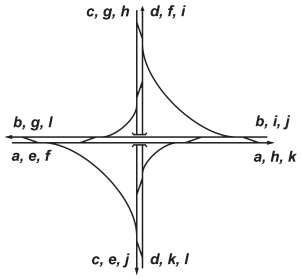


Figure 12: Alternative I of the conceptual solution of test junction

Alternative 2 - a layout that provides double track connections between main railway lines (Figure 13). Double track connections enable two heading trains to cross between main lines in parallel. In addition to two main double track lines, the layout consists of 4 double track connections with installed 16 switches and 8 fixed diamond crossings. The layout provides 60 compatibilities among the observed 12 routes.

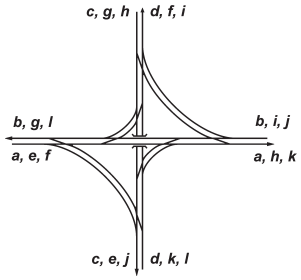


Figure 13: Alternative II of the conceptual solution of test junction

Alternative 3 - a layout that additionally reduce route conflicts providing grade separated track connections instead of fixed diamond crossings. In addition to main double track lines, the layout consists of 4 double track connections with installed 16 switches and 8 bridges. The layout provides 84 compatibilities among the observed 12 routes. This layout is shown

in Figure 14.

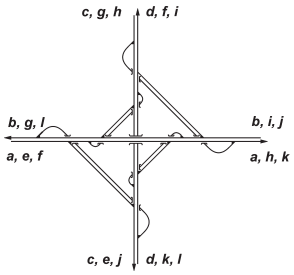


Figure 14: Alternative III of the conceptual solution of test junction

In addition to the base traffic pattern with exactly one train run per route, we analyze two variants where we increased number of trains on some routes. All routes, together with the estimated duration time for each route, are shown in Table 1.

Table 1: Assumed routes and their duration in minutes

Route symbol	Route duration [min.]
a	1.72
b	1.78
c	1.69
d	1.71
e	2.13
f	2.35
g	2.07
h	2.22
i	2.14
j	2.23
k	2.20
l	2.27

To demonstrate how the developed model responds to traffic heterogeneity, we analyze two more variants where we increased number of trains on some routes. The number of trains on each route in observed traffic pattern variants is shown in Table 2.

Table 2: The number of trains on each route in one cycle

Traffic pattern	a	b	c	d	e	f	g	h	i	j	k	l
variant I	1	1	1	1	2	4	2	4	2	4	2	4
variant II	1	1	1	1	2	4	4	2	2	4	4	2

Following a defined method, for every variant, an incompatibility graph was formed and then we applied VCP and WVCP on them. With finding the optimal solutions of WVCP for each defined variant, we obtained the minimum junction occupation times only by route realization, for each alternative separately. The obtained results are shown in Table 3.

After the minimum occupation times by route realization were established, graph reduction was executed. The reduced digraphs were used as an input to TSP and the solutions were obtained using OPL models. The results are shown in Tables 3 and 4. In this manner, we obtained junction occupation time only by minimal necessary time intervals between the groups of mutually incompatible routes, as well as the best feasible sequences of the groups, for each alternative and each variant separately.

Table 3: Acquired results, by variant

	N_{route}	N_{inc}	N_c	T_{WVCP}	T_{TSP}
alternative I	12	92	4	8.60	2.60
alternative II	12	84	3	6.40	2.05
alternative III	12	60	3	6.33	2.00
alt. I - variant I	28	564	11	24.66	6.80
alt. I - variant II	28	500	9	24.58	6.76
alt. II - variant I	28	508	7	20.26	5.86
alt. II - variant II	28	420	11	15.72	4.67
alt. III - variant I	28	308	7	15.58	4.69
alt. III -variant II	28	308	7	15.58	4.78

In the Table 3, column names represent:

- N_{route} - Number of routes,
- N_{inc} - Number of incompatibilities between the routes,
- N_c - Number of colors,
- T_{WVCP} - Total running time [min.] (solution of WVCP) and
- T_{TSP} - Total time intervals [min.] (solution of TSP).

Table 4: Junction capacity, by alternative and by variant

	U	N^h_{route}	η	N_r
alternative I	11.20	64	18.70[%]	1542
alternative II	8.45	85	14.10[%]	2044
alternative III	8.33	86	13.90[%]	2074
alt. I - variant I	31.46	53	52.40[%]	1281
alt. I - variant II	31.34	53	52.20[%]	1286
alt. II - variant I	26.12	64	43.50[%]	1543
alt. II - variant II	20.39	82	34.00[%]	1977
alt. III - variant I	20.27	82	33.80[%]	1989
alt. III -variant II	20.36	82	33.90[%]	1980

Column names in the Table 4 represent:

- U - Total utilization time [min.],
- N^h_{route} - Theoretical maximum number of routes, per hour,

- η - Utilization coefficient for one hour [%] and
- N_r - Theoretical maximum number of routes, per day

By analyzing the obtained results, we can conclude that the best design solution is alternative III, according to the maximum theoretical capacity. As the second-best solution, alternative II was selected.

Obtained results clearly indicate that in a defined model segment of determination of minimum occupation time by route realization, obtaining WVCP solution, is equally important as a segment of determination of minimum occupation time by necessary time intervals between the routes and the best feasible sequence of the routes.

4 Conclusions

Although, thus far, considerable software has been developed for a precise determination of infrastructure capacity, the existence of simple, analytical methods has always had its advantages, especially when quick solutions with satisfactory accuracy are required. A simulation model, although very fast, often requires long-term preparation for precise data acquisition and storing them in a database.

The developed model provides the possibility of a relatively simple junction capacity determination when there are no details regarding train sequence and no timetable. It's extremely useful when it is necessary to quickly obtain solutions for the comparison of several different junction designs, particularly conceptual solutions, considering that all elements are not yet determined. In addition, the model provides the possibility of precise determination of capacity utilization in the time period and determination of the best sequence of train routes.

Although all combinatorial problems used in the paper belong to the NP class (VCP in the scope of decision problem is NP-complete, WVCP is NP-hard, while TSP is also NP-complete), the application of the developed model in practice will be possible, since it is almost impossible to find a junction with so many possible routes, which would make the model too extensive for the application.

In the case study, our developed method was strictly applied on theoretical junction designs, which could be classified as of medium-heavy complexity, or, at the very least, not of easy one. Quality results were obtained, especially since the effects of different conceptual designs were immediately noticeable, even in the case of very small changes in layout. In addition, it was determined that by adopting a better design of the future junction, the utilization coefficient could be reduced by almost 5%, comparing the most favorable and most unfavorable alternatives and equal number of routes. With different train frequencies, this improvement is even more noticeable.

The model has no implemented buffer times, in order to maintain timetable robustness and stability. The implementation of these times should represent the next step in the proposed model development.

Finally, it must be noted that the construction or modernization of a junction is an investment project with various criteria, and hence, the proposed model should be incorporated into a comprehensive decision support system, where infrastructure capacity would be only one criterion.

References

Armstrong, J., Preston, J., 2017. “Capacity utilisation and performance at railway stations”, *Journal of Rail Transport Planning & Management*, vol. 7, pp. 187-205.

Corazza, G., Musso, A., 1991. ”La circolazione e gli impianti ferroviari. La verifica a lungo termine”, *Ingegneria ferroviaria*, vol. 10, pp. 607–618.

Furini, F., Malaguti, E., 2012. ”Exact weighted vertex coloring via branch-and-price”, *Discrete Optimization*, vol. 9, pp. 130–136.

Huisman, T., Boucherieb, R. J., Dijka, N. M. V., 2002. “A solvable queueing network model for railway networks and its validation and applications for the Netherlands”, *European Journal of Operational Research*, vol. 142, pp. 30-51.

Jensen, L. W., Landex, A., Nielsen, O. A., Kroon, L. G., Schmidt, M., 2017. ”Strategic assessment of capacity consumption in railway networks: Framework and model”, *Transportation Research Part C: Emerging Technologies*, vol. 74, pp. 126–149.

Kecman, P., Corman F., D’Ariano, A., Goverde, R., 2013. ”Rescheduling models for railway traffic management in large-scale networks”, *Public Transport*, vol. 5, pp. 95-123.

Landex, A., 2008. *Methods to estimate railway capacity and passenger delays*, Ph.D. thesis, Technical University of Denmark.

Landex, A., Jensen, L. W., 2013. ”Measures for track complexity and robustness of operation at stations”, *Journal of Rail Transport Planning & Management*, vol. 3, pp. 22-35.

Lindner T., 2011. ”Applicability of the analytical UIC Code 406 compression method for evaluating line and station capacity”, *Journal of Rail Transport Planning & Management*, vol. 1, pp. 49-57.

Malaguti, E., 2009. *The vertex coloring problem and its generalizations*, Ph.D. thesis, Universitá di Bologna.

Malaguti, E., Monaci, M., Toth, P., 2009. ”Models and heuristic algorithms for a weighted vertex coloring problem”, *Journal of Heuristics*, vol. 15 (5), pp. 503–526.

Malavasi, G., Molkova, T., Ricci, S., Rotoli, F., 2014. ”A synthetic approach to the evaluation of the carrying capacity of complex railway nodes”, *Journal of Rail Transport Planning & Management*, vol. 4, pp. 28–42.

Nash, A., Huerlimann, D., 2004 ”Railroad simulation using OpenTrack”, In: Allan, J. et al. (eds.), *Computers in Railways VII*, pp. 45-59, WIT Press, Southampton.

Pachl, J., 2004. *Railway Operation and Control*, 2nd Edition. VTD Rail Publishing, Mountlake Terrace WA 98043 USA.

Potthoff, G., 1980. *Verkehrsströmungslehre, Band 1 - Die Zugfolge auf Strecken und in Bahnhöfen*, 3rd Edition. Transpress Verlag, Berlin.

Radtke, A., Hauptmann, D., 2004. “Automated planning of timetables in large railway networks using a microscopic basis and railway simulation techniques”, In: Allan, J. et al. (eds.), *Computers in Railways VII*, pp. 615-625, WIT Press, Southampton.

Union International des Chemins de Fer - UIC, 2013. *UIC Code 406 - Capacity*, 2nd edition. UIC, Paris.

Yuan, J., Hansen, I., 2004. ”Analysis of Scheduled and Real Capacity Utilization At a Major Dutch Railway Station”, *WIT Transactions on The Built Environment*, vol. 74, pp. 593–602.

Yuan, J., Hansen, I. A., 2007. “Optimizing capacity utilization of stations by estimating knock-on train delays”, *Transportation Research Part B*, vol. 41, pp. 202-217.

Train Unit Shunting : Integrating rolling stock maintenance and capacity management in passenger railway stations

Franck Kamenga [a,b,1], Paola Pellegrini [c], Joaquin Rodriguez [b],
Boubekeur Merabet [a], Bertrand Houzel [a],
[a] Direction Générale Exploitation Système, SNCF Réseau
174 Avenue de France, 75013 Paris, France
[1] E-mail: franck.kamenga@reseau.sncf.fr, Phone: +33 (0)7 71 54 42 29
[b] IFSTTAR-ESTAS, Université Lille Nord de France
13 rue Elisée Reclus, 59666 Villeneuve d'Ascq, France
[c] IFSTTAR-LEOST, Université Lille Nord de France
13 rue Elisée Reclus, 59666 Villeneuve d'Ascq, France

Abstract
In passenger railway stations, train units preparation is crucial for service quality. This preparation includes maintenance check, cleaning, coupling and uncoupling. Such operations require parking train units on shunting yards located close to platforms. Therefore trains have to be moved between platform and shunting tracks. Taking over train units between their arrival and their departure in a station constitutes shunting. The Generalized Train Unit Shunting problem (G-TUSP) is the problem of shunting operations planning. The problem is to assign arriving train units to departing train units, shunting tracks and paths, to schedule shunting movements and to assign crews to maintenance operations. The aim of the paper is to provide an optimization approach for the G-TUSP. The contribution presents an integrated problem with a mixed-integer linear programming (MILP) formulation. The formulation is based on a microscopic model of the infrastructure and formal train units in order to consider coupling and uncoupling. The model is solved exactly using the commercial solver CPLEX. It is tested on instances based on Metz-Ville station in France. The results are promising and show the suitability of the model.

Keywords
Train Unit Shunting, Train Maintenance Scheduling, Track Allocation, Routing, Railway Station Capacity

1 Introduction

Rolling stock planning must manage *train units* between an arriving trip and a departure trip in a station. This specific part of rolling-stock management is called *shunting* Inside stations, train units are prepared for departure and possibly stored for several hours if they are not needed immediately. More precisely, they are cleaned and have maintenance checks. Moreover, train units can be coupled or uncoupled to match train configuration required for departure. This is done on siding tracks located around platform tracks. Parallel siding tracks form shunting *yards*. Some of these tracks have specific amenities such as train-wash

for external cleaning or pits for maintenance checks. To be stored in yards, train units need first of all to be moved from their arrival platform. Then, they can possibly need to be moved there from one yard to another. Finally they need to be moved to their departure platform. Movements arriving or departing from a yard are called *shunting movements* and must respect traffic safety rules imposed by signalling system and by ground-agents instructions. Indeed, shunting movements must not create conflicts with the rest of train traffic in the station.

Shunting operations planning includes several decisions. First, arriving train units must be assigned to departures, which constitutes a matching decision. This matching must take into account rolling stock features required for departures. Another decision concerns train units location: they must be parked at one or several shunting tracks depending on amenities required by maintenance operations. Similarly, movements are set to achieve the parking locations. For these movements, route planning decisions are to be made, since paths are assigned to train units and movements are scheduled based on running times and potential conflicts. Finally, depending on maintenance crews availability, maintenance operations must be scheduled. Although all these decisions are often taken separately, they are usually strongly interdependent. For instance, some matching plans make train units parking or maintenance scheduling impossible.

The Generalized Train Unit Shunting problem (G-TUSP) is the problem of shunting operations planning. It integrates four sub-problems:

- The Train Matching Problem (TMP), the problem of matching arriving and departing train units.
- The Track Allocation Problem (TAP), the problem of choosing train units location.
- The Shunting Routing Problem (SRP), the problem of determining train units routing during shunting movement.
- The Shunting Maintenance Problem (SMP), the problem of defining train units maintenance scheduling.

The G-TUSP considers a station and a timetable with arriving and departing trains that need to be shunted. It is a pre-operational problem, it is solved from 6 days to 4 hours before operations. The problem aims to minimize departure delays and cancellations if timetable perturbations are expected, as well as maintenance call off. Moreover, the minimization of the number of coupling and uncoupling operations is also sought.

The aim of the paper is to provide a formal model of the G-TUSP. Specifically, the contribution consists in formulating an integrated problem as a mixed-integer linear program (MILP) formulation. The formulation is based on a microscopic representation of the infrastructure and on consideration of dummy train units in order to manage coupling and uncoupling. The rest of the paper is organized as follows. Section 2 reports a summary of the literature on shunting operations planning problems. Section 3 proposes the MILP formulation of the G-TUSP. Section 4 describes the experiments carried out as proof of concept of the applicability of the formulation. Section 5 concludes the paper.

2 Related works

Several contributions introduce problems dealing with various aspects of shunting for passenger transportation.

A part of the literature focuses on the TAP without train matching. A first variant tackled concerns TAP for maintenance. In this problem, it is considered that a train unit may be parked successively on different tracks to use various equipments necessary for its maintenance. The objective is to do so as efficiently as possible. Tomii and Zhou (2000) tackle the SMP and the TAP. Here, the operations scheduling is performed through a PERT network and resource assignments are chosen thanks to a genetic algorithm. Other papers consider TAP for maintenance with a fixed maintenance schedule. Arrival and departure time on shunting tracks can be data of the problem (Li et al. (2017)) or decision variables thanks to a discrete time model (Jacobsen and Pisinger (2011)). A second variant is based on pure TAP. The combinatorial difficulty comes from the fact that several trains can be parked on the same track. When a train leaves a shunting track, it must not be blocked by another train parked in front of it. A constraint based on this requirement is called a *crossing constraint*. Also, the length of trains parked on a shunting track does not exceed the track length. A constraint based on this requirement is called a *length constraint*. Di Stefano and Koči (2004) provide significant theoretical results for TAP without length constraints. Gilg et al. (2018) propose an integer linear programming (ILP) formulation for the TAP with a robust extension and a stochastic version tested on real instances.

A second part of the literature deals with combining TAP and TMP. This combination corresponds to the Train Unit Shunting Problem (TUSP). Winter and Zimmerman (2000) study several algorithms to solve the corresponding problem in tram depots. For what concerns railway, this problem is first introduced by Freling et al. (2005) and solved with a two phases approach. MP is tackled with linear programming solver and then a column generation is used for TAP. Haijema et al. (2006) also consider a two phase approach. It is implemented with a dynamic programming based heuristic. Kroon et al. (2008) give an integrated ILP formulation which gathers TMP and TAP. Haahr et al. (2017) solve the same problem with column generation. This approach is compared with greedy algorithms and a constraint programming method. Lentink et al. (2006) propose an additional step in which they solve SRP thanks to an A* algorithm. Ramond and Marcos (2014) describe a TUSP extension to SMP for ROADEF/EURO challenge. Conflicts between shunting movements are tackled with a macroscopic representation of the infrastructure.

In this paper, we propose an integrated formulation for G-TUSP, while the literature always tackles separately one or few sub-problems.

3 Formulation

3.1 Modeling principles

In our formulation of the G-TUSP, we consider that train units can be coupled or uncoupled to form trains. Three formal sets of trains are introduced to model this: arriving, intermediate and departing trains. Arriving trains are moved from a platform track to the shunting yard. Once there, they are uncoupled if needed, and they become intermediate trains, which are moved in the yard and submitted to maintenance. Finally, intermediate trains are coupled if necessary and become departing trains to be moved to the suitable platform track. Trains move on an infrastructure modeled microscopically through a *track-circuit* scale representation. A track-circuit is a portion of track on which the presence of a train unit is automatically detected. Thanks to this infrastructure model, detailed characteristics of *interlocking systems* are taken into account and train safety is ensured through suitable

separation..

Figure 1 represents a simple example in which an orange, a green and a blue path are shown with their respective track-circuits named z followed by a number. Both orange and blue paths use track-circuit z_{15}, therefore they cannot utilize it at the same time. The train with the orange path is an intermediate train whose path starts at shunting track 21. This train results from the arriving train using the green path and has to be cleaned. It is parked at the shunting track 29 for cleaning. The train with the blue path is a departing train which uses platform A.

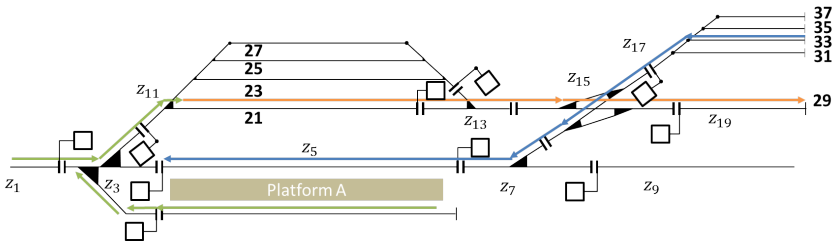


Figure 1: Simple example. Station layout with signals represented by squares. The green arriving train whose path is represented with a green line becomes the orange train at the shunting track 21. The orange intermediate train's path is represented in orange. The blue departing train leaves the shunting track and is moved to platform A. This train uses the blue path.

Trains

We denote T_T the set of arriving trains. Each arriving train can be splitted into several intermediate trains. For an arriving train t', $T_I(t')$ is the set of its intermediate trains. The set of departing trains is denoted T_S. For a departing train we denote $T_I(t)$ the set of intermediate trains which are compatible with t. Those are intermediate trains which can be coupled to obtain t. In this definition intermediate trains in $T_I(t)$ must arrive before t's departure.

Every train is composed of one or several train units. Train units are divided into types so that same type train units get interchangeable. Every arriving train entering the shunting disappear and one or more intermediate trains appear. All intermediate trains do not disappear to become departing trains. Some intermediate trains may remain in the shunting yard at the end of the planning period. For trains that are stored in the station before the planning period, a trivial train is introduced. This arriving train enters the station at the beginning of the planning period on the associated siding.

Besides, by definition, the sets $T_I(t)$ are disjoints. For readability, we introduce $T_I = \cup_{t \in T_T} T_I(t)$ that is the set of intermediate trains. We can remark that a departing train t and an arriving train t' use the same set of train unit if and only if $T_I(t) = T_I(t')$. In Figure 2, three types of train units are considered: hashed ones, full colored ones and white ones. For each arriving train, the set of its intermediate trains is represented by a thick lined dashed box. For each departing train, the set of compatible intermediate trains is represented with a tight lined dashed box. Arrows represent a possible combination of coupling and uncoupling to use the train units available to compose the two departing trains. Here, The arriving train

t_1 is uncoupled in order to obtain train t_A and two intermediate trains are coupled to obtain train t_B.

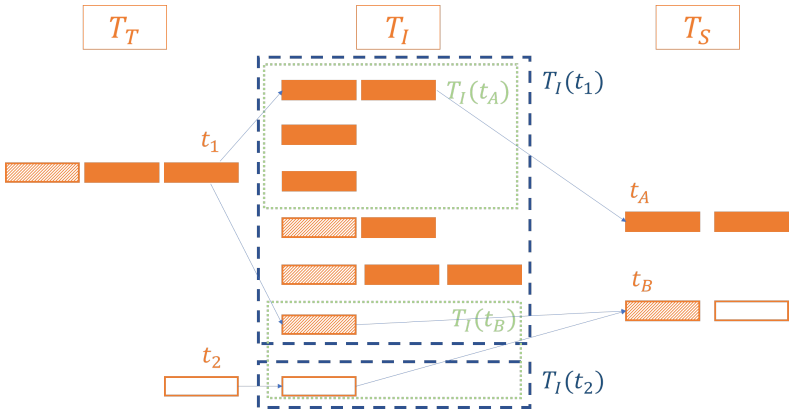


Figure 2: Train matching. Arriving trains T_T on the left are used for the departing trains T_S on the right thanks to intermediate trains T_I. A possible matching is represented with arrows.

We also consider trains which stop at the station without being shunted. Those are *passing trains*. The set of passing trains is denoted T_P.

Infrastructure

A *track-circuit scale* model is used in order to get a rigorous capacity occupation. In the station area, a train follows a path which is a track-circuits succession. As trains can turn around, a path may go twice through a track-circuit. Therefore, we introduce formal track-circuits to precise passing direction. For every real track-circuit, we consider a set of corresponding formal track-circuits. These sets contain up to two formal track-circuits, since there is a formal track-circuit per direction.

We distinguish the notion of path from that of route. Routes are individually handled and defined by signalling control. A path is the concatenation of routes and may include turnarounds. In the turnarounds, a first route is defined up to the turnaround place where a second route starts.

Capacity occupation is based on track-circuit reservation. When a train t needs to go through a track-circuit tc, the signal which allows t to move into the *block section* where tc is located must have a green aspect. A block section is a sequence of track-circuits which can be utilized by at most one train at a time. Thanks to the interlocking system, the green aspect can be obtained once the path r that leads t to tc is formed. This is why we introduce formation times, which depends on block sections characteristics. However r can only be formed if all conflicting routes are released. A block section locked by a train is released shortly after this train clears the last track-circuit it is using in the block section itself.

A path can imply parking on a shunting track. Paths are set such that shunting tracks are at the beginning or the end of the path. For a path r, we define Ps^r the set of shunting tracks where r starts and Pe^r the set of shunting tracks where r ends. Ps^r and Pe^r can contain one shunting track or be empty. Every train has a set of usable paths. Arriving

trains paths terminate at a shunting track and go through a platform, while departing trains paths begin at a shunting track and go through a platform. Exactly one path is assigned to arriving, departing and passing trains. Intermediate trains paths begin and terminate on a shunting track. When an intermediate train needs to be parked at several tracks, several paths are assigned to it. In order to define a sequence of paths, two fictive paths r_0 and r_∞ are assigned to intermediate trains. r_0 is at beginning of the sequence while r_∞ terminates it.

The set of exit points of a shunting track p is denoted $Ex(p)$. This set contains at most two elements which indicate a geographical location. We use the locations left and right, respectively denoted L and R. A train enters in (or exits from) a shunting track p with the path r by the exit point $Es(r,p) \in Ex(p)$ (or $Ee(r,p) \in Ex(p)$). A path r_2 can only follow a path r_1 if r_1 ends at the shunting track where r_2 begins: $Ps^{r_2} \cap Pe^{r_1} \neq \emptyset$. In the example of Figure 1, the green path is denoted r_1 and the orange one is denoted r_2. r_2 follows r_1 at the siding track 21. Indeed $Ps^{r_2} = Pe^{r_1} = \{21\}$.

Maintenance operations

Cleaning or maintenance operations may be included in the rolling-stock plan. They are considered to be made on intermediate trains. The operations carried out on an intermediate train $t \in T_I$ form set O_t. An operation $o \in O_t$ can only be performed on shunting tracks with specific facilities. The sequence of operations is given. We introduce P^o set of shunting tracks where o can be carried out. In addition, an operation requires the use of specific human resources. We consider that an operation o requires a crew among the set HR^o of crews which can be assigned to o. Each crew is available from its shift start time to its shift end time.

We also note that when an operation is in progress, the shunting track where it is carried out must be protected to ensure staff safety. Thus, during this period, no other train can enter this shunting track or leave it.

3.2 MILP formulation

In the MILP, we use the following notations:

T_T,T_I,T_S,T_P	set of arriving trains, intermediate trains, departing trains, passing trains
$T = T_T \cup T_I \cup T_S \cup T_P$	set of trains
$T^* = T_T \cup T_I \cup T_S$	set of shunted trains
$T_I(t)$	set of intermediate trains compatible with the arriving or departing train $t \in T_T \cup T_S$
TU,$m_{t,tu}$	set of train unit types, number of train units of type $tu \in TU$ in the train $t \in T^*$
index t	index of train $t \in T$
ty_t,l_t, a_t,d_t	type of train $t \in T$, length of train $t \in T$, arrival time of train $t \in T_T \cup T_P$, departure time of train $t \in T_S \cup T_P$
B_t,Q_t	cancellation cost of train $t \in T_S$, cost associated to the delay of train $t \in T_S \cup T_P$
A_t,Q_R	cost of one time unit duration of a shunting movement performed on the intermediate train $t \in T_I$ and cost of the assignment of a route to the intermediate train $t \in T_I$

$\omega_{t,t'}$	weight associated to the assignment of intermediate train $t' \in T_I(t)$ to departing train $t \in T_S$
Q_C,Q_H	coupling cost, uncoupling cost
b_{t_1,t_2}	indicator function: 1 if $t_1 \in T_I(t)$ (with $t \in T_T$) is placed to the left of $t_2 \in T_I(t)$ with index $t_1 <$ index t_2, 0 otherwise
$i(t,t')$	indicator function: 1 if train $t \in T^*$ is reused for train $t' \in T^*$, $i(t,t') = 1 \iff (t \in T_T, t' \in T_I(t)) \vee (t = t' \in T_I) \vee (t' \in T_S, t \in T_I(t'))$
mp	minimum parking time
R_t,TC_t,Z_t	set of paths, formal track-circuits and real track-circuits which can be used by a train $t \in T$
$TC(z)$	set of formal track-circuits corresponding real track-circuit $z \in \cup_{t \in T} Z_t$
Z^r, TC^r	set of real and formal track-circuits the path $r \in \cup_{t \in T} R_t$
M_R	maximum number of paths which can be assigned to an intermediate train
$OTC_{ty,r,tc}$	set of consecutive formal track-circuits preceding $tc \in TC^r$ which are occupied by a train of type ty traveling along path $r \in \cup_{t \in T} R_t$ if its head is on tc, depending on train and track-circuit length
$pc_{r,t}$, $sc_{r,t}$	formal track-circuits preceding and following $tc \in TC^r$ along path $r \in R_t$
$rt_{ty,r,tc}$, $ct_{ty,r,tc}$	running and clearing time of $tc \in TC^r$ along $r \in \cup_{t \in T} R_t$ for a train of type ty
$ref_{r,tc}$	reference formal track-circuit for reservation of $tc \in TC^r$ along $r \in \cup_{t \in T} R_t$
$bs_{r,t}$	block section including formal track-circuit $tc \in TC^r$ along $r \in \cup_{t \in T} R_t$
for_{bs},rel_{bs}	formation and release time for block section bs
Ps^r,Pe^r, P^r	set of shunting tracks where $r \in \cup_{t \in T} R_t$ begins, set of shunting tracks where $r \in \cup_{t \in T} R_t$ ends, set of tracks in r $P^r = Ps^r \cup Pe^r$
$Z(p)$,$Ex(p)$	set of real track-circuits and set of exit points composing a shunting track p
L_p	length of shunting track p
$tcp_{r,p}, tce_{r,p}$	reference formal track-circuit for parking at shunting track $p \in P^r$ along $r \in \cup_{t \in T} R_t$, first formal track after shunting track $p \in Ps^r$ along $r \in \cup_{t \in T} R_t$
$Es(r,p)$, $Ee(r,p)$	entrance and exit point of $r \in \cup_{t \in T} R_t$ at shunting track $p \in Pe^r$ and $p \in Ps^r$
O_t	set of operations to carry on $t \in T_I$
pR^o,ω_o	duration and cancellation cost of operation $o \in \bigcup_{t \in T_I} O_t$
HR^o,P^o	set of crews and shunting tracks which can be assigned to operation $o \in \bigcup_{t \in T_I} O_t$
E_t	set of successive operations on $t \in T_I$. $(o, o') \in E_t$ if and only if the operation o' follows the operation o

sR_{hr},eR_{hr}	shift start time and shift end time of crew hr
M,τ_M	large constant compared to event times, end of planning period

In the formulation, we introduce non-negative continuous variables:

- $oc_{t,r,tc}$, $\phi_{t,r,tc}$, $sU_{t,r,tc}$, $eU_{t,r,tc}$, with $t \in T, r \in R_t, tc \in TC^r$: time at which t starts the occupation of tc along r, additional running time of t on tc along r, time at which tc starts being utilized by t along r, time at wich tc ends being utilized by t along r
- $sO_{o,r,r',p,hr}$, with $t \in T_I, o \in O_t, r, r' \in R_t, p \in P^r \cap P^o, hr \in HR^o$: time at which o starts at shunting track p between paths r and r' with crew hr
- D_t, with $t \in T_S \cup T_P$: delay suffered by train t when exiting the control area

Moreover, we introduce binary variables:

- xT_t, with $t \in T_I$, is equal to 1 if t is created and 0 otherwise
- $xS_{t,t'}$, with $t \in T_S, t' \in T_I(t)$, is equal to 1 if t' is assigned to t and 0 otherwise
- $xR_{t,r}$, with $t \in T$, $r \in R_t$, is equal to 1 if t uses r and 0 otherwise
- $xO_{o,r,r',p,hr}$, with $t \in T_I, o \in O_t, r, r' \in R_t, p \in Pe^r \cap Ps^{r'} \cap P^o, hr \in HR^o$, is equal to 1 if o is carried out at shunting track p between paths r and r' with crew hr and 0 otherwise
- qS_t, with $t \in T_S$, is equal to 1 if t is cancelled and 0 otherwise
- $yR_{t,t',r,r',tc,tc'}$ with $t, t' \in T$, $r \in R_t$, $r' \in R_{t'}$, $z \in Z^r \cap Z^{r'}$, $tc, tc' \in TC(z)$, $tc, tc' \in TC^r \cap TC^{r'}$, index $t <$ index t', is equal to 1 if t uses tc along r before t' uses tc' along r' and 0 otherwise
- $k_{t,r,r'}$, with $t \in T_I, r, r' \in R_t, (Ps^{r'} \cap Pe^r \neq \emptyset) \vee (r = r_0) \vee (r' = r_\infty)$ (i.e. r' can follow r), is equal to 1 if t uses r followed by r' and 0 otherwise
- $y_{o,o',hr}$ with $t, t' \in T$, $o \in O_t$, $o' \in O_{t'}$, $hr \in HR^o \cap HR^{o'}$, index $t <$ index t', is equal to 1 if hr performs o before o' and 0 otherwise
- $ysO_{o,t,r_1,r_2,r,p}$, with $t \in T_I$, $t' \in T_I$, $t \neq t'$, $o \in O_{t'}$, $r_1, r_2 \in R_{t'}$, $r \in R_t$, $p \in P^o \cap Pe^{r_1} \cap Ps^{r_2} \cap Pe^{r'}$, is equal to 1 if operation o is carried out at shunting track p between path r_1 and r_2 before t enters shunting track p through r and 0 otherwise
- $yeO_{o,t,r_1,r_2,r,p}$, with $t \in T_I$, $t' \in T_I$, $t \neq t'$, $o \in O_{t'}$, $r_1, r_2 \in R_{t'}$, $r \in R_t$, $p \in P^o \cap Pe^{r_1} \cap Ps^{r_2} \cap Ps^{r'}$, is equal to 1 if operation o is carried out at shunting track p between path r_1 and r_2 before t leaves shunting track p through r and 0 otherwise

We also introduce the following integer variables:

- u_t, with $t \in T_T$ gives the number of uncoupling operations on t

- v_t, with $t \in T_S$ gives the number of coupling operations on t

The objective function to minimize integrates several penalties (1). First, it takes into account the cost of departure cancellations and delays. The function includes uncoupling and coupling operations cost. Then, penalties for intermediate trains assignment to departing trains are added. Moreover, we minimize the number of shunting movements for an intermediate train and the duration of these movements. Finally maintenance operations cancellation costs are introduced. We note that we can have a penalty only if the intermediate train concerned by the operation is actually created.

$$\min \sum_{t \in T_S} B_t \cdot qS_t + \sum_{t \in T_S \cup T_P} Q_t D_t + \sum_{t \in T_T} Q_C \cdot u_t + \sum_{t \in T_S} Q_H \cdot v_t + \sum_{t \in T_S} \sum_{t' \in T_I(t)} \omega S_{t,t'} xS_{t,t'} + \sum_{t \in T_I, o \in O_t} \left(xT_t - \sum_{\substack{p \in P^O \cap Pe^r \cap Ps^{r'} \\ r,r' \in R_t, hr \in HR^O}} xO_{o,r,r',p,hr} \right) + \sum_{t \in T_I} \sum_{r \in R_t, p \in Ps^r} Q_R xR_{t,r} + A_t(oc_{t,r,tc_\infty} - oc_{t,r,tce_{r,p}}) \quad (1)$$

Matching constraints

The MILP formulation must consider TMP constraints. First, we need to check train compositions. We introduce constraints for the number of train units of a specific type in trains. For each type, each arriving train must have the same number of train units as intermediate trains created after uncoupling (2). Also, each departing train must have the same number of train units as the intermediate trains assigned to it for coupling (3). As intermediate trains can not be splitted, each of them can be assigned at most to one departing train. If the intermediate train is not created, it can not be assigned to a departing train (4). A departure train is cancelled if no intermediate train is assigned to it (5). Then, the number of uncoupling operations on an arriving train or coupling operations on a departing train is equal to the number of intermediates trains assigned minus one (6), (7).

$$m_{t,tu} = \sum_{t' \in T_I(t)} m_{t',tu} xT_{t'} \ \ \forall t \in T_T, tu \in TU \quad (2)$$

$$m_{tu,t} = \sum_{t' \in T_I(t)} m_{t',tu} xS_{t,t'} \ \ \forall t \in T_S, tu \in TU \quad (3)$$

$$\sum_{t' \in T_S : t \in T_I(t')} xS_{t',t} \leq xT_t \ \ \forall t \in T_I \quad (4)$$

$$1 - qS_t \leq \sum_{t' \in T_I(t)} xS_{t,t'} \ \ \forall t \in T_S \quad (5)$$

$$u_t \geq \sum_{t' \in T_I(t)} xT_{t'} - 1 \ \ \forall t \in T_T \quad (6)$$

$$v_t \geq \sum_{t' \in T_I(t)} xS_{t,t'} - 1 \ \ \forall t \in T_S \quad (7)$$

Routing constraints

An arriving or a passing train cannot be operated before its arrival time (8). The start time of track-circuit occupation by a train along a path is zero if the path itself is not used (9). A train starts occupying a track-circuit along a path after spending in the preceding track-circuit its running time and an additional running time, if the path is used (10). An arriving or a passing train uses exactly one path (11). These sets of constraints are inspired by the RECIFE-MILP model of Pellegrini et al. (2015). A departing train uses exactly one path if it is created and zero otherwise (12). An intermediate uses at most M_R paths if it is created and zero otherwise (13). If an intermediate is created, it uses the dummy paths r_0 (14) and r_∞ (15).

$$oc_{t,r,tc} \geq a_t \cdot xR_{t,r} \;\; \forall t \in T_T \cup T_P, r \in R_t, tc \in TC^r \tag{8}$$

$$oc_{t,r,tc} \leq M \cdot xR_{t,r} \;\; \forall t \in T, r \in R_t, tc \in TC^r \tag{9}$$

$$oc_{t,r,tc} = oc_{t,r,pc_{r,tc}} + \phi_{t,r,pc_{r,tc}} + rt_{t,r,pc_{r,tc}} \cdot xR_{t,r} \;\; \forall t \in T, r \in R_t, tc \in TC^r \tag{10}$$

$$\sum_{r \in R_t} xR_{t,r} = 1 \;\; \forall t \in T_T \tag{11}$$

$$\sum_{r \in R_t} xR_{t,r} = 1 - qS_t \;\; \forall t \in T_S \tag{12}$$

$$\sum_{r \in R_t} xR_{t,r} \leq M_R \cdot xT_t \;\; \forall t \in T_I \tag{13}$$

$$xR_{t,r_0} = xT_t \;\; \forall t \in T_I \tag{14}$$

$$xR_{t,r_\infty} = xT_t \;\; \forall t \in T_I \tag{15}$$

Two constraints model the sequence of path used by an intermediate train. If a path is used by an intermediate train:

- exactly one path follows it (16),
- exactly one path precedes it (17).

$$\sum_{r' \in R_t : (Pe^r \cap Ps^{r'} \neq \emptyset) \vee r' = r_\infty} k_{t,r,r'} = xR_{t,r} \;\; \forall t \in T_I, r \in R_t \setminus \{r_\infty\} \tag{16}$$

$$\sum_{r' \in R_t : (Pe^{r'} \cap Ps^{r} \neq \emptyset) \vee r' = r_0} k_{t,r',r} = xR_{t,r} \;\; \forall t \in T_I, r \in R_t \setminus \{r_0\} \tag{17}$$

A delay is at least equal to the difference between the actual exit time from the infrastructure and the scheduled departure time (18).

$$D_t \geq \sum_{r \in R_t} oc_{t,r,tc_\infty} - d_t \;\; \forall t \in T_S \cup T_P \tag{18}$$

The formulation includes constraints that take into account train matching decisions and the sequence of paths used by an intermediate train. These constraints consider two trains t and t' which use the same rolling-stock. A minimum parking time must be ensured between t's arrival (at the end of t's path) and t''s departure on the shunting track. It happens when an arriving train t becomes an intermediate train t' (19), when an intermediate train uses two path in a row (20) and when an intermediate train becomes an departing train (21).

$$\begin{aligned} oc_{t',r',tce_{r',p}} \geq \sum_{r \in R_t : p \in Pe^r} & [oc_{t,r,pc_{r,tc_\infty}} + (rt_{t,r,pc_{r,tc_\infty}} + mp) \cdot xR_{t,r}] \\ & -M(1 - k_{t,r_0,r'}) \;\; \forall t \in T_T, t \in T_I(t), r' \in R_{t'}, p \in Ps^{r'} \end{aligned} \tag{19}$$

$$\begin{aligned} oc_{t,r',tce_{r',p}} \geq oc_{t,r,pc_{r,tc_\infty}} & + rt_{t,r,pc_{r,tc_\infty}} + mp - M(1 - k_{t,r,r'}) \\ & \forall t \in T_I, r, r' \in R_{t'} : p \in Pe^r \cap Ps^{r'} \end{aligned} \tag{20}$$

$$\begin{aligned} \sum_{r \in R_{t'} : p \in Ps^{r'}} oc_{t,r,tce_{r',p}} \geq oc_{t,r,pc_{r,tc_\infty}} & + (rt_{t,r,pc_{r,tc_\infty}} + mp) \cdot xR_{t,r} \\ & -M(1 - xS_{t',t}) \;\; \forall t' \in T_S, t \in T_I(t'), r \in R_t, p \in Pe^r \end{aligned} \tag{21}$$

Moreover, we need to ensure spatial coherence. It means that when an arriving train t becomes an intermediate train t', t uses a path which ends at the same shunting track as the path used by t' (22), (23). The same happens when an intermediate train t' becomes a departing train t (24), (25).

$$\begin{aligned} \sum_{r \in R_t : p \in Pe^r} xR_{t,r} \leq \sum_{r \in R_{t'} : p \in Ps^r} & k_{t',r_0,r} + M_R(1 - xT_{t'}) \\ & \forall t \in T_T, t' \in T_I(t), p \in \bigcup_{r \in R_t, r' \in R_{t'}} (Pe^r \cup Ps^{r'}) \end{aligned} \tag{22}$$

$$\begin{aligned} \sum_{r \in R_{t'} : p \in Ps^r} k_{t',r_0,r} \leq \sum_{r \in R_t : p \in Pe^r} & xR_{t,r} + M_R(1 - xT_{t'}) \\ & \forall t \in T_T, t' \in T_I(t), p \in \bigcup_{r \in R_t, r' \in R_{t'}} (Pe^r \cup Ps^{r'}) \end{aligned} \tag{23}$$

$$\begin{aligned} \sum_{r \in R_{t'} : p \in Pe^r} k_{t',r,r_\infty} \leq \sum_{r \in R_t : p \in Ps^r} & xR_{t,r} + M_R(1 - xS_{t,t'}) \\ & \forall t \in T_S, t' \in T_I(t), p \in \bigcup_{r \in R_t, r' \in R_{t'}} (Ps^r \cup Pe^{r'}) \end{aligned} \tag{24}$$

$$\sum_{r \in R_t : p \in Ps^r} xR_{t,r} \leq \sum_{r \in R_{t'} : p \in Pe^r} k_{t',r,r_\infty} + M_R(1 - xS_{t,t'})$$
$$\forall t \in T_S, t' \in T_I(t), p \in \bigcup_{r \in R_t, r' \in R_{t'}} (Ps^r \cup Pe^{r'}) \tag{25}$$

Otherwise, track-circuits of shunting tracks must remain in use when a train is parked. Thus, when an arriving train t becomes an intermediate train t', t' starts using the first track-circuit of its path before t finishes using the last track-circuit of its path (26). The same happens when an intermediate train uses two paths in a row (27) and when an intermediate train becomes a departing train (28).

$$sU_{t',r',sc_{r',tc_0}} \leq eU_{t,r,pc_{r,tc_\infty}} - M(2 - k_{t',r_0,r'} - xR_{t,r})$$
$$\forall t \in T_T, t' \in T_I(t), r \in R_t, r' \in R_{t'}, Pe^r \cap Ps^{r'} \neq \emptyset \tag{26}$$

$$sU_{t,r',sc_{r',tc_0}} \leq eU_{t,r,pc_{r,tc_\infty}} - M(1 - k_{t,r,r'})$$
$$\forall t \in T_T, r, r' \in R_t, Pe^r \cap Ps^{r'} \neq \emptyset \tag{27}$$

$$sU_{t,r,sc_{r',tc_0}} \leq eU_{t,r',pc_{r,tc_\infty}} - M(2 - k_{t',r',r_\infty} - xR_{t,r})$$
$$\forall t \in T_S, t' \in T_I(t), r \in R_t, r' \in R_{t'}, Ps^r \cap Pe^{r'} \neq \emptyset \tag{28}$$

An additional set of constraints deals with formal track-circuit tc reservation. A train's utilization of a track-circuit along a route starts as soon as the train starts occupying the reference formal track-circuit $ref_{r,tc}$ for the reservation of tc minus the formation time (29). A train's utilization of a track-circuit along a route ends when the track-circuit has been physically cleared plus the release time (30). Thus, the equality considers running time, additional running time and clearing time on the track-circuit tc along the path r. Finally, it incorporates possible additional running time on following track-circuits if the train t is long enough to occupy more than one track-circuit at a time. Then, there exists tc' so that tc is physically occupied by t while the head of t reaches the end of track-circuit tc', i.e. $tc \in OTC(t, r, tc')$. There are also disjunctive constraints (31)(32) so that that two trains can not utilize a track-circuit at the same time. These constraint does not affect track-circuits of common shunting tracks.

$$sU_{t,r,tc} = oc_{t,r,ref_{r,tc}} - for_{bs_{r,tc}} xR_{t,r} \quad \forall t \in T, r \in R_t, tc \in TC^r \tag{29}$$

$$eU_{t,r,tc} = oc_{t,r,tc} + ((rt_{t,r,tc} + ct_{t,r,tc} + rel_{bs_{r,tc}})xR_{t,r} + \phi_{t,r,tc})$$
$$+ \sum_{tc' \in TC : tc \in OTC(t,r,tc')} \phi_{t,r,tc'} \quad \forall t \in T, r \in R_t, tc \in TC^r \tag{30}$$

$$eU_{t,r,tc} - M(1 - yR_{t,t',r,r',tc,tc'}) \leq sU_{t',r',tc'}$$
$$\forall t, t' \in T, \text{index } t < \text{index } t', r \in R_t, r' \in R_{t'}, z \in Z^r \cap Z^{r'} \setminus \bigcup_{p \in P^r \cap P^{r'}} Z(p), \tag{31}$$
$$tc \in TC(z) \cap TC^r, tc' \in TC(z) \cap TC^{r'}$$

$$\begin{aligned} eU_{t',r',tc'} - M \cdot yR_{t,t',r,r',tc,tc'} \leq sU_{t,r,tc} \\ \forall t, t' \in T, \text{index } t < \text{index } t', r \in R_t, r' \in R_{t'}, z \in Z^r \cap Z^{r'} \setminus \bigcup_{p \in P^r \cap P^{r'}} Z(p), \\ tc \in TC(z) \cap TC^r, tc' \in TC(z) \cap TC^{r'} \end{aligned} \tag{32}$$

Maintenance scheduling constraints

For maintenance operations, we specify the inequalities that must be verified at the beginning of the tasks. This must take into account the availability of crew and shunting tracks.

If an intermediate train t is obtained, any operation carried on on t uses only one crew and one shunting track along a given path (33). An operation performed by crew hr must start after the shift start time of hr (34) and before its shift end time (35). An operation carried on on train t at shunting track p between paths r and r' needs to start after t's arrival on p through r. t's arrival time on p through r is given by the expression $sP_{t,r,p}$ (39). If $r \neq \{r_0\}$, $sP_{t,r,p}$ is the moment when t starts using the reference track-circuit for parking at p (37). Else, $r = r_0$ and we need to consider the arriving train which uses the same rolling-stock. Then an intermediate train arrives at its first shunting track when its corresponding arriving train arrives (38). Besides, an operation carried on on train t at shunting track p between paths r and r' needs to finish before t's departure from p through r'. t's departure time from p through r' is given by the expression $eP_{t,r',p}$ (36). If $r \neq \{r_\infty\}$, $sP_{t,r,p}$ is the moment when t starts using the reference track-circuit for parking at p (40). Else, $r = r_\infty$ and we need to consider the departing train which uses the same rolling-stock. Then an intermediate train leaves its first shunting track when its corresponding departing train leaves (41),(42). If no departing train is assigned to t, then t stays at its last shunting track until the end of the planning period (43). Otherwise, if an operation o' follows an operation o, then o' starts after the end of o (44).

$$\sum_{hr \in HR^o, r, r' \in R_t, p \in Pe^r \cap Ps^{r'} \cap P^o} xO_{o,r,r',p,hr} \leq xT_t \ \ \forall t \in T_I, o \in O_t \tag{33}$$

$$\begin{aligned} sO_{o,r,r',p,hr} \geq sR_{hr} \cdot xO_{o,r,r',p,hr} \\ \forall t \in T_I, o \in O_t, r, r', \in R_t, p \in Pe^r \cap Ps^{r'} \cap P^o, hr \in HR^o \end{aligned} \tag{34}$$

$$\begin{aligned} sO_{o,r,r',p,hr} + pR^o \leq eR_{hr} \cdot xO_{o,r,r',p,hr} \\ \forall t \in T_I, o \in O_t, r, r', \in R_t, p \in Pe^r \cap Ps^{r'} \cap P^o, hr \in HR^o \end{aligned} \tag{35}$$

$$\begin{aligned} sO_{o,r,r',p,hr} \geq sP_{t,r',p} - M \sum_{p \in P^o} (1 - xO_{o,r,r',p,hr}) \\ \forall t \in T_I, o \in O_t, r, r' \in R_t, p \in Pe^r \cap Ps^{r'} \cap P^o, hr \in HR^o \end{aligned} \tag{36}$$

$$sP_{t,r,p} = sU_{t,r,tcp_{t,p}} \ \ \forall t \in T_I, r \in R_t \setminus \{r_0, r_\infty\}, p \in Pe^r \tag{37}$$

$$sP_{t,r_0,p} = \sum_{r' \in R_{t'} : p \in Pe^{r'}} sU_{t',r',tcp_{r',p}} \ \ \forall t' \in T_T, t \in T_I(t'), p \in \bigcup_{r \in R_t} Ps^r \tag{38}$$

$$sO_{o,p,r,r',hr} + pR^o \leq eP_{t,r',p} + M \sum_{p \in P^o} (1 - xO_{o,p,r,hr}) \\ \forall t \in T_I, o \in O_t, r, r' \in R_t, p \in Pe^r \cap Ps^{r'} \cap P^o, hr \in HR^o \tag{39}$$

$$eP_{t,r,p} = eU_{t,r,tcp_{r,p}} \ \ \forall t \in T_I, r \in R_t \setminus \{r_0, r_\infty\}, p \in Ps^r \tag{40}$$

$$eP_{t,r_\infty,p} \geq \sum_{r' \in R_{t'}: p \in Pe^{r'}} eU_{t,r,tcp_{r',p}} - M(1 - xS_{t',t}) \\ \forall t' \in T_S, t \in T_I(t'), p \in \bigcup_{r \in R_t} Ps^r \tag{41}$$

$$eP_{t,r_\infty,p} \leq \sum_{r' \in R_{t'}: p \in Pe^{r'}} eU_{t,r,tcp_{r',p}} + M(1 - xS_{t',t}) \\ \forall t' \in T_S, t \in T_I(t'), p \in \bigcup_{r \in R_t} Ps^r \tag{42}$$

$$eP_{t,r_\infty,p} \geq \tau_M - M \left(1 - \sum_{t' \in T_S: t \in T_I(t')} xS_{t',t}\right) \\ \forall t \in T_I, p \in \bigcup_{r \in R_t} Ps^r \tag{43}$$

$$sO_{o',r'_1,r'_2,p',hr'} \geq sO_{o,r_1,r_2,p,hr} + pR^o \\ \forall t \in T_I, \forall (o,o') \in E_t, r_1, r'_1, r_2, r'_2 \in R_t, p \in P^o \cap Pe^{r_1} \cap Ps^{r_2}, \\ p' \in P^{o'} \cap Pe^{r'_1} \cap Ps^{r'_2}, hr \in HR^o, hr' \in HR^{o'} \tag{44}$$

As two operations can not use a crew at the same time, there are disjunctive constraints (45), (46).

$$sO_{o',r'_1,r'_2,p',hr} \geq sO_{o,r_1,r_2,p,hr} + pR^o - M(1 - y_{o,o',hr}) \\ \forall t, t' \in T, o \in O_t, o' \in O_{t'}, hr \in HR^o \cap HR^{o'}, r_1, r_2 \in R_t, r'_1, r'_2 \in R_{t'}, \\ p \in P^o \cap Pe^{r_1} \cap Ps^{r_2}, p' \in P^{o'} \cap Pe^{r'_1} \cap Ps^{r'_2}, \text{index } t < \text{index } t' \tag{45}$$

$$sO_{o,r_1,r_2,p,hr} \geq sO_{o',r'_1,r'_2,p',hr} + pR^o - My_{o,o',t,t',hr} \\ \forall t, t' \in T, o \in O_t, o' \in O_{t'}, hr \in HR^o \cap HR^{o'}, r_1, r_2 \in R_t, r'_1, r'_2 \in R_{t'}, \\ p \in P^o \cap Pe^{r_1} \cap Ps^{r_2}, p' \in P^{o'} \cap Pe^{r'_1} \cap Ps^{r'_2}, \text{index } t < \text{index } t' \tag{46}$$

Finally, there is the protection of the garage tracks during an operation. A disjunction sets that trains must enter a shunting track before the beginning (47) or after the end (48) of an operation. An other disjunction sets that trains must leave a shunting track before the beginning (49) or after the end (50) of an operation.

$$\begin{aligned} sP_{t,r,p} \geq sO_{o,r_1,r_2,p,hr} + pR^o + M(1 - ysO_{o,t,r_1,r_2,r',p}) \\ \forall t \in T_I, t' \in T_I, t \neq t', o \in O_{t'}, r_1, r_2 \in R_{t'}, r \in R_t, \\ p \in P^o \cap Pe^{r_1} \cap Ps^{r_2} \cap Pe^{r'} \end{aligned} \tag{47}$$

$$\begin{aligned} sO_{o,r_1,r_2,p,hr} \geq sP_{t,r,p} + MysO_{o,t,r_1,r_2,r',p} \\ \forall t \in T_I, t' \in T_I, t \neq t', o \in O_{t'}, r_1, r_2 \in R_{t'}, r \in R_t, \\ p \in P^o \cap Pe^{r_1} \cap Ps^{r_2} \cap Pe^{r'} \end{aligned} \tag{48}$$

$$\begin{aligned} eP_{t,r,p} \geq sO_{o,r_1,r_2,p,hr} + pR^o + M(1 - yeO_{o,t,r_1,r_2,r',p}) \\ \forall t \in T_I, t' \in T_I, t \neq t', o \in O_{t'}, r_1, r_2 \in R_{t'}, r \in R_t, \\ p \in P^o \cap Pe^{r_1} \cap Ps^{r_2} \cap Ps^{r'} \end{aligned} \tag{49}$$

$$\begin{aligned} sO_{o,r_1,r_2,p,hr} \geq eP_{t,r,p} + MyeO_{o,t,r_1,r_2,r',p} \\ \forall t \in T_I, t' \in T_I, t \neq t', o \in O_{t'}, r_1, r_2 \in R_{t'}, r \in R_t, \\ p \in P^o \cap Pe^{r_1} \cap Ps^{r_2} \cap Ps^{r'} \end{aligned} \tag{50}$$

Parking constraints

Parking constraints are based on constraints which involve precedence between events. In a second step, these precedence variables are used to express the parking constraints.

A first set of variables indicates if two trains use a shunting track at the same time. Thanks to these variables length constraints are set.

For crossing constraints, we introduce two set of binary variables. The first one indicates the relative position of two trains when they enter a shunting track and the second one indicates the relative position of two trains when they leave the track. Two trains must have the same relative position on a shunting track when they enter and when they leave it. These positioning variables are deduced with a disjunction. This disjuntion is based on two assertions:

- if train t enters shunting track p through route r after t' through route r', t is placed on $Es(r,p)$ side of t'
- if train t leaves shunting track p through route r before t' through route r', t is placed on $Ee(r,p)$ side of t'

Table 1 presents a disjunction for entrance relative position variable. This variable is defined with intermediate trains $t, t' \in T_I$, routes $r \in R_t, r' \in R_{t'}$ and shunting track $p \in Pe^r \cap Pe^{r'}$.

$Es(r,p)$	$Es(r',p)$	t enters before t'	t' enters before t
L	L	0	1
L	R	1	1
R	L	0	0
R	R	1	0

Table 1: Values of the entrance relative position variable, with $t \in T_I$, index $t <$ index t', $r \in R_t, r' \in R_{t'}$, $p \in Pe^r \cap Pe^{r'}$. Variable equal to 1 if t is placed on left side of t' and 0 if t is placed on right side of t'

4 Experiments

In this section, we report on experiments that test the model on a panel of instances. The model is coded in Java and solved exactly using the commercial solver CPLEX. As in principle we shall deploy our solution method for G-TUSP in dispatching centers, it must be able to run on a computer of standard configuration. Therefore, it is executed on a 32 bit operation system equipped with a 2.1 GHz Intel®Core™i3-51010U processor and 4GB RAM. We study Metz-Ville station infrastructure. It is a major hub for Eastern France railway traffic. We tackle real scenarios which include disturbances such as arrival delay or track closure.

4.1 Case study

We consider traffic in Metz-Ville infrastructure and its passengers shunting yards represented in Figure 3. It is a major junction where the Nancy-Luxembourg and Metz-Strasbourg lines intersect. The station mainly hosts regional trains. Many of these trains start or end their service in Metz-Ville. The area is 3.8 km long and has 10 platforms including a dead-end one. The yards F1 and F2 are controlled from the signal box, while switches are directly handled by a ground-agent in yards F3 and F4. The infrastructure is composed of 138 track-circuits, 68 signals, 421 block sections and 405 routes.

The set of path R_t that can be used by a train is computed thanks to breadth-first search (BFS). In preprocessing, this BFS is based on a graph, whose vertices are signals or signs. Its edges are routes between signs and signals or represent turnarounds.

We consider a regular week day and two disturbed week days in 2018. One disturbed day includes several delays form Luxembourg between 16:30 and 19:40. During the other disturbed day, one of the two north side shunting necks is closed. This shunted neck circled in the red (Figure 3) and the available one is circled in green. Here trains perform turnarounds when necessary. A first set of scenarios studies trains between evening peak hour (18:30) and next morning peak hour (07:30). These are scenarios where trains have to be shunted for the night. Trains enters in yards in the evening to leave in the morning. A second set of scenarios considers trains between morning peak hour (07:00) and evening peak hour (19:00). In those scenarios trains are stored during the day. As we need to focus on those trains, we do not have to consider passing trains in the whole time horizon. Indeed, conflicts between shunting movements and passing trains occur during rush hours only. During off-peak time, Metz-Ville dispatchers can trivially find conflict-free shunting routes. Therefore, we only consider passing trains during peak hours (6:30 - 9:00 and 17:00

- 19:30).

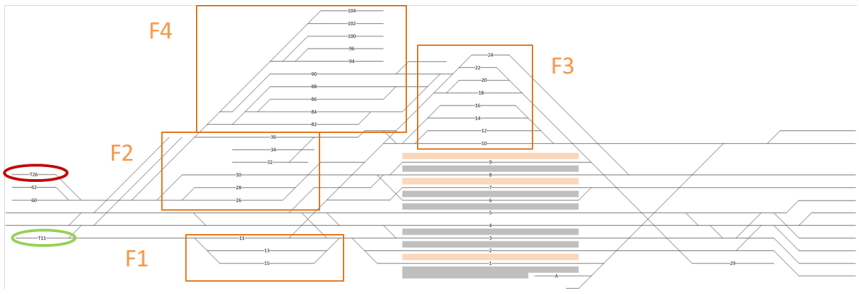


Figure 3: Layout of Metz-Ville station

Name	Day/ Night	Disturbance	$\|T_P\|$	$\|T^*\|$	# of continuous variables	# of binary variables	# of constraints
D1	Day	None	27	6	87 768	891 946	2 402 227
D2	Day	track closure	25	7	91 207	1 345 509	3 134 528
D3	Day	arrival delays	25	6	74 437	804 814	2 182 065
N1	Night	None	22	9	131 084	2 423 404	3 785 881
N2	Night	track closure	24	8	119 640	1 786 528	2 834 102
N3	Night	arrival delays	22	10	153 742	2 659 013	4 257 630

Table 2: Details on the instances tackled in the experimental analysis ($|T_P|$: number of passing trains, $|T^*|$:number of shunted trains)

Table 2 reports the details on the six instances tackled. In each of them there are 7 types of trains on which 4 different operations can be performed: arrival check, internal cleaning, WC cleaning and external cleaning. The track closure scenario reduces the set of possible shunting paths and imply the occurrence of conflicts. Indeed, if a train has to be moved from yard F2 to yard F4, it has to cross main tracks. In instance N3, as trains arrive late in the evening peak hour, their operation can not start on time. In this scenario, in reality as cleaning crews shift ended too early, some cleaning operations were actually postponed to the morning or cancelled. In Table 2, we report the number of passing trains $|T_P|$ and shunted trains $|T^*|$ as well as the number of continuous and binary variables created in the model. Despite the limited set of trains, we get large number of variables. This is essentially because of precedence variables yR which indicate the using order of a track-circuit.

4.2 Results

CPLEX running time is deliberately limited to 3600 seconds. Beyond that duration there is no practical interest for operational planning. Table 3 reports results obtained on the 6 instances described in Section 4.1. It shows the number of coupling and uncoupling required on shunted trains as well as the number of modifications to the planned train matching. It also reports the average number of routes allocated to an intermediate train by our solution

and the average number of routes actually allocated by dispatchers. Moreover, we indicate delays taken by departures performed by trains coming from shunting yards. However passing trains departures can also be delayed in addition to shunted trains delays, then the total delay reported in Table 3 comes from these two contributions. The table also shows the actual total delay recorded on traffic database. We remark that the solver does not reach an optimal solution or a proof of optimality in the allotted time. In particular, the gaps exceed 20 % in arrival delay scenarios.

Instance	D1	D2	D3	N1	N2	N3
running time (sec)	3600	3600	3600	3600	3600	3600
# cancelled operations	0	0	0	0	0	1
act. # cancel. op.	**0**	**0**	**0**	**0**	**0**	**2**
# coupling	1	2	0	1	2	2
# uncoupling	2	1	1	2	0	3
# modif. match.	0	0	0	0	2	3
av. # shunt. paths	2.5	3.09	2.67	2.89	3,38	3.10
act. av. # shunt. paths	**2.17**	**2.43**	**2.33**	**2.56**	**2.75**	**2.40**
# shunt. dep. del.	0	1	0	0	0	1
act. # shunt. dep. del.	**0**	**1**	**0**	**0**	**1**	**1**
tot. shunt. mov. time (min)	166.82	287.29	130.04	357.02	434.60	397.55
total delay (min)	0	11.87	54.51	0	3.43	26.32
act. total delay (min)	**0**	**12.5**	**68.5**	**0**	**8.0**	**25.0**
integer solution value	1584.11	1645.80	972.80	1978.45	986.71	2257.32
gap (%)	16.12	7.77	20.56	9.32	13.64	24.30

Table 3: Experiments results (act. # cancel. op.: number of cancelled operations by rolling stock managers, modif. match.: modifications to the planned train matching, av. # shunt. paths: average number of shunting paths allocated to intermediate trains by our solution, act. av. # shunt. paths: average number of shunting paths allocated to intermediate train by dispatchers, shunt. dep. del.: shunted departure trains delayed, act. # shunt. dep. del.: number of shunted departure trains delayed by dispatchers solution, tot. shunt. mov. time: total shunting movement time, act. total delay: total delay in dispatchers solution)

There is no total delay on D1 and N1 instances. However, more shunting movements are performed in our solution than in the one implemented by dispatchers. The solution of D2 brings a departure delayed as in the actual traffic data. It is in both cases the same train, nevertheless it suffers from a 8.34 minute delay in our solution while it was 10 minutes in reality. In instance D3, despite a significant gap, the solution obtained reduces the total delay. The solution of N2 switches two trains in order to reduce the delay. For N3, the result is notably different from the actual decisions. The solution switches three trains in order to cancel fewer operations. However, total delay gets higher. In summary, the implementation of our MILP formulation call the attention on relevant alternatives to the choices of dispatchers in the tested instances. In particular, they highlight the significant effect of changes in the train matching for G-TUSP.

However, we remark that Metz-Ville has a large number of sidings compared to the number of shunted trains. Indeed, it is not necessary to park several trains on the same

track except for coupling or uncoupling. Then, in our solutions, trains are always parked on different tracks. It makes a part of TAP constraints useless for these instances.

5 Conclusion

In this paper, we provided a formal description for the G-TUSP, which is the integrated problem of managing shunting operations planning in passenger trains. We tackled a large decision problem that includes many specific operational constraints. We presented a MILP formulation for allocations and continuous time scheduling.

The model copes with both rolling-stock management and capacity management. We extended some literature approaches which combine TAP with TMP. Moreover, we introduced microscopic-scale routing features based on a MILP formulation for real-time traffic management and maintenance scheduling aspects. Maintenance aspects led us to consider that the trains can be successively parked on several tracks which is typically not considered in TUSP literature. The proof of concept carried out on the Metz-Ville instance validates the model relevance. Indeed, it confirms the interest of implementing an integrated approach for improving the operating performance of a station. Even if we can not prove the optimality of the solutions, they are very satisfying compared to the decisions made by dispatchers.

Our study highlights practical issues we will like to tackle in future research. We first need to reduce calculation time. A heuristic phase may provide a first integer solution to the MILP solver, which typically has a major impact on performance. Improvements of the MILP formulation based on valid inequalities may be proposed. In principle, We may also reduce the number of variables, especially precedence ones, by reducing the number of routes to consider. The choice of the remaining routes is in this case critical, and a suitable approach must be found. Other solution techniques such as decomposition can be applied in future works. Moreover, to increase the practical relevance of the formulation, the weights used in the objective function needs to be set in a very accurate way. They are currently quite arbitrary, and they may not properly mimic the need of compromises in real-life situations.

References

G. Di Stefano and M. L. Koči. A graph theoretical approach to the shunting problem. *Electronic Notes in Theoretical Computer Science*, vol. 92,pp. 16–33, 2004.

R. Freling, R. M. Lentink, L. G. Kroon, and D. Huisman. Shunting of passenger train units in a railway station. *Transportation Science*, 39(2):261–272, 2005.

B. Gilg, T. Klug, R. Martienssen, J. Paat, T. Schlechte, C. Schulz, S. Seymen, and A. Tesch. Conflict-free railway track assignment at depots. *Journal of rail transport planning & management*, 8(1):16–28, 2018.

J. T. Haahr, R. M. Lusby, and J. C. Wagenaar. Optimization methods for the train unit shunting problem. *European Journal of Operational Research*, 262(3):981–995, 2017.

R. Haijema, C. Duin, and N. M. Van Dijk. Train shunting: A practical heuristic inspired by dynamic programming. *Planning in Intelligent Systems: Aspects, Motivations, and Methods*, pages 437–475, 2006.

P. M. Jacobsen and D. Pisinger. Train shunting at a workshop area. *Flexible services and manufacturing journal*, 23(2):156–180, 2011.

L. G. Kroon, R. M. Lentink, and A. Schrijver. Shunting of passenger train units: an integrated approach. *Transportation Science*, 42(4):436–449, 2008.

R. M. Lentink, P.-J. Fioole, L. G. Kroon, and C. v. Woudt. Applying operations research techniques to planning of train shunting. *Planning in Intelligent Systems: Aspects, Motivations, and Methods*, pages 415–436, 2006.

H. Li, M. Jin, S. He, Z. Ye, and J. Song. Optimal track utilization in electric multiple unit maintenance depots. *Computers & Industrial Engineering*, 108(Supplement C):81 – 87, 2017.

P. Pellegrini, G. Marlière, R. Pesenti, and J. Rodriguez. RECIFE-MILP: An effective MILP-based heuristic for the real-time railway traffic management problem. *IEEE Transactions on Intelligent Transportation Systems*, 16(5):2609–2619, 2015.

F. Ramond and N. Marcos. Roadef/euro 2014 challenge, trains don't vanish!-final phase, rolling stock unit management on railway sites. *Challenge ROADEF/EURO*, page 28, 2014.

N. Tomii and L. J. Zhou. Depot shunting scheduling using combined genetic algorithm and pert. *WIT Transactions on The Built Environment*, 50, 2000.

T. Winter and U. T. Zimmermann. Real-time dispatch of trams in storage yards. *Annals of Operations Research*, 96(1):287–315, 2000.

Sustainability of Railway Passenger Services – A Review of Aspects, Issues, Contributions and Challenges of Life Cycle Emissions

Marko Kapetanović [a,1], Niels van Oort [a], Alfredo Núñez [b], Rob M.P. Goverde [a]

[a] Department of Transport and Planning, Delft University of Technology
P.O. Box 5048, 2600 GA Delft, The Netherlands
[1] E-mail: M.Kapetanovic@tudelft.nl, Phone: +31 (0) 15 27 84914
[b] Section of Railway Engineering, Delft University of Technology
P.O. Box 5048, 2600 GA Delft, The Netherlands

Abstract
This paper presents a review of research and models regarding sustainability of railway passenger services. In order to take into account all relevant aspects in terms of environmental impacts of a railway passenger service, a holistic system perspective is required, that includes a whole life cycle assessment. A life cycle approach is important since comparison of for instance only the exhaust emissions of an electric vehicle with a petrol vehicle is misleading, due to neglecting the emissions of for instance electrical energy production process. Thus, all stages in energy carrier, vehicle and infrastructure life cycles are to be considered. Existing models are analyzed, as well as possible developments, focusing on diesel and electrical traction as the most common traction options in use, and on GHG emissions, especially on CO_2, which takes the greatest part in all emissions. Issues and challenges in improving the environmental impact of railway passenger services are addressed. Additionally, several areas are indicated where environmental aspects could be included in future assessment models. The main challenge is answering how the existing partial assessments can be brought together and, together with filling the identified gaps, allow to conduct a comprehensive LCA which will produce real-world emissions estimations. Results of this paper will be used as an input in developing a framework for quantifying and improving overall environmental impacts of a railway passenger service.

Keywords
Railway transport, Sustainability, Environmental pollution, CO_2, Life cycle assessment

1 Introduction

"Sustainable Transportation" is a widely discussed and researched topic. Starting from the report titled "Our Common Future" of the Brundtland Commission (UN, 1987), in which the sustainable development is defined as a "*development which meets the needs of current generations without compromising the ability of future generations to meet their own needs*", a number of initiatives and studies have been conducted in the transport industry. Reference is often made to the three 'dimensions' or 'pillars' of sustainability – namely the environment, the economy, and society/social equity. However, the majority of studies so far prioritized economic aspects.

The transport sector, as one of the largest contributors in global greenhouse gas (GHG) emissions, is especially affected by the increased concerns for the environment in the last

decade(s). Carbon dioxide (CO_2) takes the largest part in all GHG emissions from transportation, more than 95%, while other most represented GHGs include methane (CH_4), nitrous oxide (N_2O), sulphur hexafluoride (SF_6), hydrofluorocarbons (HFC) and perfluorocarbons (PFC) (EU, 2017). In quantifying the amount and the composition of emitted GHGs, in order to make different types of GHGs comparable, a so called CO_2 equivalence factor ($CO_{2\text{-eq}}$) is defined for each of them (IPCC, 2007). This factor expresses the global warming potential (GWP) of one unit of a GHG compared with one unit of CO_2. For instance, N_2O has a $CO_{2\text{-eq}}$-factor of 298, i.e. one ton of N_2O has the same global warming effect as 298 tons of CO_2 (EC, 2014). Globally, the railway sector was responsible for 1.9% of transport-related final energy demand, and for 4.2% of CO_2 emissions from the transport sector in 2015. Following the UN's Paris Climate Agreement from 2015 (UN, 2015), the EU's overall goal is to reduce GHG emissions from transport by 2050 to a level that is 60 % below that of 1990 (EEA, 2017). For the railway sector targets are set by the UIC (International Union of Railways) and CER (Community of European Railway and Infrastructure Companies), with the short term target on decreasing CO_2 emissions by 30% over the period 1990 to 2020, with a further decrease by 50% in 2030 (UIC, 2012).

Taking into account the global tendency in modal shift to railways, the environmental impact of this mode of transport should be given more attention. In their "5E" framework which is used to quantify the value of public transport using five E's (Effective mobility, Efficient city, Economy, Environment, and Equity), Van Oort et al. (2017) showed that one of the main potential benefits of modal shift to railways regards environmental aspects. Technological progress is also made in recent years with the introduction of alternative fuels. However, comprehensive studies which would encompass the whole life cycle and give the insights in total impact of the novel energy options for railways are lacking in the literature.

In this paper, a review that highlights and analyzes the contributions in environmental sustainability related to passenger railway services is presented. Existing models are reviewed, as well as possible developments, focusing on diesel and electrical traction, as the most common traction options in use, and on GHG emissions, especially on CO_2, which takes the greatest part in all emissions. Additionally, main issues and challenges are addressed and several areas are indicated where environmental aspects could be included in future assessment models.

Section 2 introduces existing emissions assessment approaches and outlines the differences between them. Section 3 reviews the literature on the direct emissions estimations for railways. Section 4 gives the review on railway Well-to-Wheel (WTW) analyses. Section 5 reviews the railway Life Cycle Assessment (LCA) studies. Discussion on the main findings is given in section 6. Finally, section 7 ends this literature review with the main conclusions and provides the future research directions.

2 Railway Emissions Assessment Approaches

Emissions as a consequence of railway service operation are closely related and are directly influenced by the energy consumption. Thus, in most railway emissions assessments energy use and emissions estimation are carried out simultaneously. In general, all the emissions from the railway service operation can be divided into direct emissions (e.g. from diesel consumption in the combustion engine, usually referred as the consumption phase) and indirect emissions (e.g. from energy carrier production and delivery and the construction/production and maintenance of infrastructure and vehicles).

A number of approaches in emissions assessment have been developed and applied, and

the selection of the adequate method is influenced by numerous factors and aspects, such as the goal and scope of the study, system boundaries, data availability, does the study represent *ex ante* or *ex post* evaluation, etc. In general, two main categories of research methods for calculating energy use and emissions per transport unit can be distinguished (Van Wee et al., 2005):

- 'Bottom-up' methods (BUMs), which explicitly include determinants such as weight, resistances, speed, etc.; and
- 'Top-down' methods (TDMs), which use aggregated data in calculations by dividing total energy use and emissions by the selected transport indicator, i.e. tons of $CO_{2\text{-eq}}$ / passenger-km.

Regarding the scope and system boundaries of the study, a number of studies limited their scope on direct emissions from the consumption phase (Papagiannakis and Hountalas, 2003; Lapuerta et al., 2008; Papagiannakis et al., 2010a; 2010b; Johnson et al., 2013). In order to take into account all relevant aspects in terms of environmental impacts of a

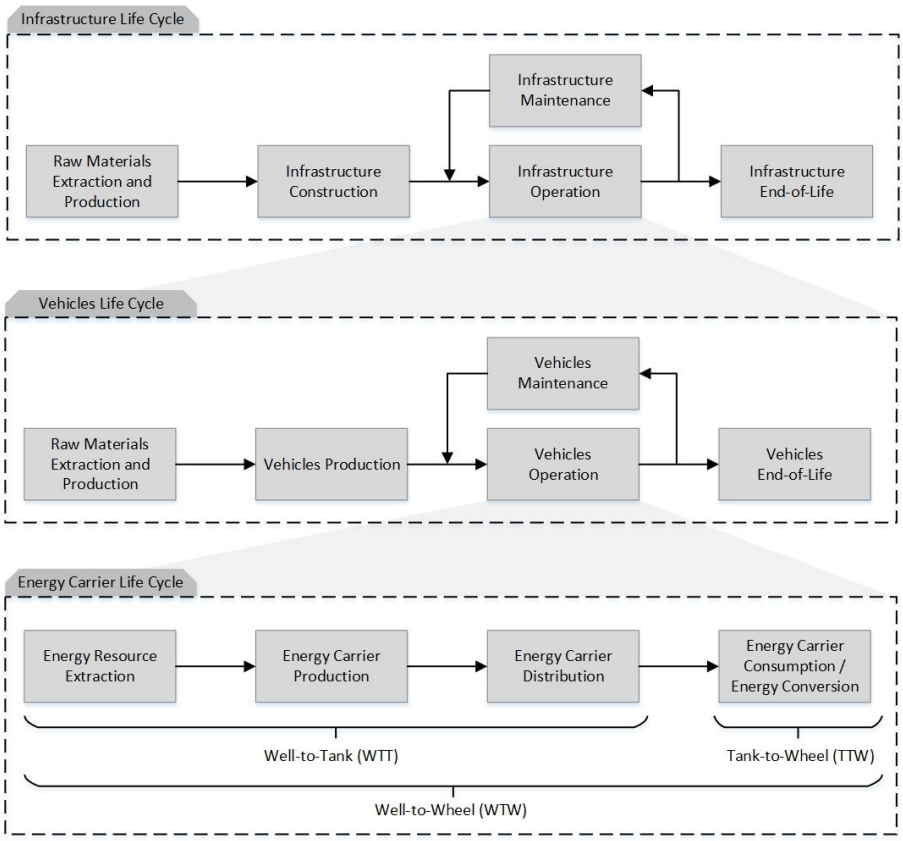


Figure 1: Infrastructure, Vehicles and Energy Carrier Life Cycle

passenger railway service a holistic system perspective that observes the whole life cycle (emissions from all stages in energy carrier, vehicle and infrastructure life cycles) has gained great importance in the recent years. The complete infrastructure, vehicles and energy carrier life cycles with the main corresponding processes are presented in Fig. 1.

The life cycle approach is important, because, for instance, comparison of only the exhaust emissions of an electric vehicle with a petrol vehicle is misleading, due to neglecting the emissions from electrical energy production, especially if the primary resource is i.e. coal. A holistic approach helps in better understanding of energy consumption and associated CO_2 emissions by analyzing these aspects throughout the whole life cycle of the system, instead of only considering the consumption phase.

Studies that observe the whole energy carrier pathway employ the so-called Well-to-Wheel (WTW) approach. WTW analyses are divided into two stages, as depicted in Fig. 1: (i) Well-to-Tank (WTT) stage, consisting of energy resource extraction, production and distribution processes; and (ii) Tank-to-Wheel (TTW) stage, also referred as the vehicle operation phase, or the consumption phase (Hoffrichter et al., 2012). The WTW approach neglects infrastructure construction and vehicles production, as well as infrastructure and vehicles end-of-life processes (recycling and disposal), and it represents a subclass of a wider-scope Life Cycle Assessment (LCA) approach (Orsi et al., 2016). LCA observes the complete infrastructure and/or vehicles pathway, and in most cases explicitly or implicitly encompasses all the processes included in WTW.

The organization of this review paper is based on the scope and system boundaries criteria, where the following three sections provide a review on: (i) studies and approaches focused on the direct emissions from the consumption phase; (ii) WTW analyses which observe energy carrier life cycle; and (iii) LCA studies which encompass infrastructure and/or vehicles life cycles and associated emissions.

3 Direct Emissions from the Consumption Phase

A number of studies has limited their research on the consumption phase, in particular on direct energy consumption and related emissions. Two different approaches for estimating emissions in this phase can be identified in the literature: (i) applying direct on-track or laboratory measurements, using modern equipment, sensors, etc.; and (ii) applying mathematical models and numerical calculations.

3.1 Emissions Obtained from Direct Measurements

Direct measurements in assessing emission levels is in most cases applied in testing engines powered by different liquid and gaseous fuels, such as diesel, bio-diesel, or natural gas using modern measuring equipment. These measurements are in most cases project-tailored and represent expensive and extensive experiments. Although usually case-specific, the results of these studies can be very useful in future research, either in the assessment models development or in results validation. Existing studies in the literature and their main findings are given chronologically in the remaining of this sub-section, as follows.

Papagiannakis and Hountalas (2003), Papagiannakis et al. (2010a, 2010b) conducted an experimental investigation to examine the effects of the emissions of a high speed, compression ignition engine where liquid diesel fuel is partially substituted by natural gas in various proportions, with the natural gas fumigated into the intake air. The experimental results disclose the effect of these parameters on nitric oxide (NO_x), carbon monoxide (CO), unburned hydrocarbons (HC) and particulate matter (PM) emissions, with the beneficial

effect of the presence of natural gas being revealed. They conclude that dual fuel combustion using natural gas as a supplement for liquid diesel fuel is a promising technique for controlling both NO_x (decrease up to 47%) and PM emissions on existing diesel ignition engines, requiring only slight modifications of the engine structure. The observed disadvantages are an increase in HC and CO emissions that can be possibly mitigated by applying modifications on the engine tuning, e.g. injection timing of liquid diesel fuel mainly at part loads.

In 2006 Rail Safety and Standards Board (RSSB) and the Association of Train Operating Companies (ATOC) investigated the use of bio-diesel on Britain's railways and published a report on August 2010 (RSSB, 2010). The effects on the engine's performance and exhaust emissions were tested using increasing biodiesel blending. The engines were tested under laboratory conditions on a range of blends of bio-diesel, from 5% bio-diesel (B5) in steps up to 100% bio-diesel (B100). Based on the results, it has been concluded that B20 (a 20% blend of bio-diesel mixed with 80% diesel) was sensibly the highest blend that could be accepted without significant expenditure to retune engines. The use of B20 did not appear to cause any significant engine wear, but the fuel consumption performance was worse. Generally when using bio-fuel: the fuel consumption increased; NOx levels tended to increase; the total HC emissions tended to decrease; CO and CO_2 emissions were less consistent throughout but tended to be lower than for diesel; the PM and exhaust smoke decreased.

Lapuerta et al. (2008) collected and analyzed papers published in scientific journals about diesel engine emissions when using bio-diesel fuels as opposed to conventional diesel fuels. The first section is dedicated to the effect of bio-diesel fuel on engine power, fuel consumption and thermal efficiency, while the second section focus on the comparison of engine emissions from bio-diesel and diesel fuels, paying special attention to the most concerning emissions: NO_x and PM, the latter not only in mass and composition but also in size distributions. In this case the highest consensus was found in the sharp reduction in PM emissions.

Xue et al. (2011) analyzed reports about bio-diesel engine performance and emissions, published by highly rated journals in scientific indexes since year 2000. The effects of biodiesel on engine power, economy, durability and emissions including regulated and non-regulated emissions were analyzed. It was found that the use of bio-diesel leads to substantial reduction in PM, HC and CO emissions accompanying with a small power loss, increase in fuel consumption and increase in NO_x emissions on conventional diesel engines.

Poompipatpong and Cheenkachorn (2011) modified a diesel engine for natural gas operation and evaluated the emission and power output effects of such modifications. They also mentioned that two of the advantages of natural gas are clean combustion and attractive price. They tested the emissions of CO, THC and NO_x for different compression ratios and compared the results.

Abdelaal and Hegab (2012) tested a single-cylinder direct injection (DI) diesel engine on regular operation and dual-fuel mode, with natural gas as the main fuel and diesel fuel as a pilot. Comparative results of exhaust emission were presented for several operating modes. They mentioned natural gas as a partial supplement for diesel fuel as a very promising solution for reducing pollutant emissions, particularly NO_x and PM. The results showed reduction in NO_x and CO_2 emissions, while CO emissions increased.

In 2012, Clean European Rail-Diesel (CleanER-D, 2012) delivered a report on the impact and performance of alternative fuels in rail applications. The main objective was to study the different types of fuel used in railway applications and their effect on engine parameters. It was found that bio-diesel blends up to 20% are technically feasible although

increasing fuel consumption compared to diesel.

Park et al. (2012) examined the PM characteristics of diesel locomotive engine exhaust at various engine ratings. Diesel engine exhaust was collected via a dilution tunnel and the concentration and size distribution of fine particles were measured by a scanning mobility particle sizer. The results showed that the maximum CO emission was reached at 59% of the maximum rating, after which emissions decreased.

Johnson et al. (2013) described and applied a technique for analyzing exhaust emission plumes from unmodified locomotives under real world conditions from railway trains servicing an Australian shipping port. The method utilized simultaneous measurements downwind of the railway line of the following pollutants: particle number, PM, mass fraction, SO_2, NO_x and CO_2, from which emission factors were then derived. Samples from 56 train movements were collected, analyzed and presented. The quantitative results for emission factors were noted and the findings were compared with previously published papers. Statistically significant correlations within the group of locomotives sampled were found between the emission factors for particle number, SO_2 and NO_x.

3.2 Emissions Obtained from Numerical Calculations

Obtaining emission levels by means of numerical calculations can be done using both TDMs and BUMs. TDMs are usually used for direct emissions calculation in WTW or LCA studies, since they use aggregated data and are easily incorporated in wider scope studies. Most commonly used BUMs in calculating the emissions of rolling stock in the consumption phase are through energy consumption calculations based on resistances. Since the large majority of the energy used by the train (≈80%) is to overcome resistances that the train is subject to when traveling along the track, once these resistances are known they can be multiplied by the distance traveled in determining total energy consumption (Network Rail, n.d.; SYSTRA, 2011). Once energy consumption needed for overcoming the resistances is calculated, it then can be multiplied by the emission factors in order to obtain the total emissions of the train.

All resistances can be split into two categories: (i) inertial/grade resistances, which account for the infrastructure characteristics, and are independent of the train; and (ii) running resistances, which depend on train characteristics and train speed (UIC, 2003). Running resistances of a train can be modelled using the standard Davis Equation (Davis, 1926):

$$R = A + Bv + Cv^2 \quad (1)$$

where R is resistance (N), v is speed (m/s), and A, B and C are coefficients specific to the train obtained from the experimental data, where A is proportional to the mass of the train and accounts for the bearing resistances, B accounts for the rolling resistance and C for the air resistance.

Esters and Marinov (2014) identified three different existing methods for energy consumption calculation based on resistances and applied them in calculating emissions of UK rolling stock. The three methods for energy consumption calculation are: (i) the International Union of Railways (UIC) method, (ii) the Rail Safety and Standard Board (RSSB) method, and (iii) the ARTEMIS rail emissions model. Although they all start from the standard Davis Equation given in (1), the coefficients and amount of data required for their implementation differs. The three methods are presents in sub-sections as follows.

International Union of Railways (UIC) Method

The UIC methodology (Garcia, 2010) factors the distance travelled into the equations and thus gives the energy consumption directly instead of resistances of the train. Total energy consumption is calculated as:

$$E = E_m + E_a \tag{2}$$

where E_m represents the energy due to mechanical resistances, and E_a the energy due to aerodynamic resistances. Energy consumption due to mechanical resistances depends on the mass of the train and arise due to the contact between the wheels of the train and the track:

$$E_m = (a + a_c) \cdot m \cdot l \tag{3}$$

where a is the coefficient depending on the rolling stock (N/t), a_c is the coefficient depending on the route - number of curves on a track and their length and radius (N/t), m is mass of the train (t), and l is the length of the route (m).

Energy due to aerodynamic resistance (E_a) is expressed as the sum of drag due to pressure forces (E_p) and drag caused by friction (E_f). Energy required to overcome pressure drag is given by:

$$E_p = c_p \cdot S_f \cdot \int T_f \cdot v^2 \cdot dl \tag{4}$$

where c_p is the pressure drag coefficient (N/(km/h)^{2}m^2), S_f is the cross-sectional frontal area of train (m^2), T_f is the tunnel factor, v is speed (km/h), and l is the length of the route (m).

Energy needed to overcome frictional drag is given by:

$$E_f = c_f \cdot S_m \cdot \int T_f \cdot v^2 \cdot dl \tag{5}$$

with c_f the frictional drag coefficient (N/(km/h)^{2}m^2), and S_m the wet surface area (m^2) where the train will feel shear stresses due to the forward motion of the train:

$$S_m = ((2H) + W) \cdot L_t \tag{6}$$

where H is the height of the train (m), W is the width of the train (m), and L_t is the length of the train (m).

Rail Safety and Standards Board (RSSB) Method

The RSSB methodology (RSSB, 2007) uses a specific version of the Davis Formula:

$$R = k \cdot M + (B_1 + B_2) \cdot v + C \cdot v^2 \tag{7}$$

where k is the constant of proportionality, M is the mass of the train (kg), B_1 is a constant which relates to the rolling resistance of the train and is linearly proportional to the mass of the train, B_2 is a constant representing the mass of cooling air and the mass of ventilation air, v is the train speed (m/s), and C is a constant used to describe the aerodynamics of the train, given by:

$$C = \frac{\rho}{2} \cdot C_d \cdot A_x \tag{8}$$

where ρ is the density of air (kg/m^3), A_x is the cross-sectional frontal area of the train (m^2), and C_d is the drag coefficient given by:

$$C_d = C_{dht} + C_{dl} + C_{db} + C_{di} + C_{de} \tag{9}$$

where C_{dht} is the head and tail drag coefficient and is determined by the pressure forces at the head and tail of the train, C_{dl} is the frictional drag coefficient and is linearly proportional to the length of the train, C_{db} is the bogie drag coefficient, C_{di} is the extra drag coefficient dependent on the number of vehicles, and C_{de} is the pantograph drag coefficient used to account for the pressure forces felt by the pantographs on an electric train. C_{dl} is given by:

$$C_{dl} = L \cdot L_f \tag{10}$$

where L is length of the train (m), and L_f is the length factor. C_{db} is given by:

$$C_{db} = 2N_v \cdot B_f \tag{11}$$

with N_v the number of vehicles, and B_f the bogie factor. C_{di} is given by:

$$C_{di} = 0.025(N_v - 1) \tag{12}$$

ARTEMIS Rail Emissions Model

The ARTEMIS rail emissions model (Lindgreen and Sorenson, 2005) uses a fundamental approach to calculating resistance, which is split into two parts. Summing the two resistive forces gives:

$$F_m = F_R + F_L \tag{13}$$

where F_m is the total resistance of the train (N), F_R is the rolling resistance (N), and F_L is the air resistance (N). Rolling resistance is given by:

$$F_R = f_R \cdot m \cdot g \tag{14}$$

where m is mass of the train (kg), g is gravitational acceleration (m/s2), and f_R is the rolling resistance coefficient given by:

$$f_R = C_0 + C_1 \cdot \left(\frac{v}{v_0}\right) + C_2 \cdot \left(\frac{v}{v_0}\right)^2 \tag{15}$$

where C_0, C_1 and C_2 are coefficients, v is the train speed (km/h), and v_0 is the speed constant equal to 100km/h. C_1 and C_2 are constant specific for different train types, and C_0 is given by:

$$C_0 = \frac{f_{sl} \cdot m_l + f_{sv} \cdot m_v}{m} \quad (16)$$

where f_{sl} is the rolling resistance coefficient for locomotive which depends on the number of axles of the locomotive, f_{sv} is the rolling resistance coefficient for carriages, m_l is the total mass of locomotives (kg), m_v is the total mass of carriages (kg), and m is the total mass of the train (kg). f_{sv} is a function of axle load and is given by:

$$f_{sv} = C_{cv} + \left(\frac{F_A \cdot n_{ax}}{m \cdot g}\right) \quad (17)$$

where C_{cv} is a coefficient that depends on the type of vehicle, F_A is an axle pressure constant equal to 100N, and n_{ax} is the total number of axles of carriages.

Air resistance (F_L) has a similar form as in previous methods and is given by:

$$F_L = \frac{\rho}{2} \cdot C_L \cdot A_x \cdot v^2 \quad (18)$$

where ρ is the density of air (1.247 kg/m^3), A_x is the cross-sectional frontal area of train (m^2), and C_L is the drag coefficient calculated by summing the contributions of the carriages and locomotives:

$$C_L = \sum C_{car} + C_{loco} \quad (19)$$

where C_{car} and C_{loco} are the drag coefficients of a carriage and the front loco, respectively. C_{loco} is defined by the number of axles, shape of the locomotive and whether it is an electric or diesel powered train.

The presented models and approaches can be extended by incorporating real conditions that influence consumption and emissions, such as track resistances, driving styles, etc. The effect of regenerative braking could also be included as it contributes in energy savings in case of electric traction. Also optimal energy-efficient train driving and energy-efficient timetabling strategies can contribute in reduction of energy consumed, and thus in total emissions. A comprehensive review of approaches in energy-efficient train control and timetabling can be found in Scheepmaker et al. (2017).

4 Railway Well-to-Wheel Analyses

A Well-to-Wheel (WTW) analysis observes the whole life cycle of an energy carrier (i.e. diesel, electricity, etc.), and can be subdivided into the Well-to-Tank (WTT) stage that focuses on the energy carrier supply chain, and the Tank-to-Wheel (TTW) stage, which covers the vehicle operation (Fig. 1). Many variations of WTW analyses have been proposed in the literature for automotive and bus industry (Yazdanie et al., 2014; Li et al., 2016; Orsi et al., 2016; Correa et al., 2017; Woo et al., 2017; Dreier et al., 2018), mostly applying different modifications of the GREET (Regulated Emissions, and Energy use in Transportation) fuel-cycle model (ANL, 2016), ADVISOR (Advanced Vehicle Simulator) software (ADVISOR, 2003) and other commercial and non-commercial models. On the other hand, the number of studies analyzing railway transportation from WTW perspective

is rather scarce. Although WTW analyses are in most cases explicitly or implicitly included in LCA studies, the calculations are based mainly on aggregated data and approximate estimations.

Hoffrichter et al. (2012) evaluated energy efficiencies and CO_2 emissions for electric, diesel and hydrogen traction for railway vehicles on a WTW basis using existing estimations in the literature. They use the low heating value and high heating value of the enthalpy of oxidation of the fuel. The TTW and WTT efficiency are determined. Gaseous hydrogen (H_2) has a WTW efficiency of 25% low heating value, if produced from methane and used in a fuel cell. This efficiency is similar to diesel and electric traction in the UK, US, and California. A reduction of about 19% in CO_2 is achieved when hydrogen gas is used in a fuel cell compared to diesel traction, and a 3% reduction compared to US electricity. The paper shows that a high WTW efficiency reduces the amount of energy needed from the original source and that a reduction in overall emissions is possible. The case of diesel traction demonstrates that a high WTW efficiency does not automatically lead to lower emissions. Hydrogen as an energy carrier to provide power for railway vehicles is a suitable solution on efficiency and emission bases, if fuel cells are used. The WTW efficiency is similar to electric and diesel systems, but the CO_2 emissions are lower than for diesel traction. If electricity is largely produced from high carbon fuels, a reduction of CO_2 is possible through the utilization of hydrogen when produced from natural gas.

Esters and Marinov (2014) analyzed and compare the methods used for calculating emissions of UK rolling stock based on their type and mode of operation. The three modes under comparison were diesel, electric and bi-mode. As well as comparing these three modes of operation, a comparison between Conventional, Freight and High Speed Rail was made. Alternate fuels were considered for diesel and bi-mode locomotives and compared based on their environmental impact. The emissions of trains were studied using three methods presented in Sec. 3. Specifically, the three chosen methods were used to calculate the emissions of each train and a comparison of these methods was made. In the current UK energy climate, diesel trains emit less emissions than electric trains when factoring in mechanical and air resistances, due to domination of high carbon primary source for electricity production. Bi-mode trains have their place in the UK network but with electrification of the network currently in place, this mode of operation will become redundant in the near future. High Speed Rail, although time efficient, releases high emissions due to energy consumption increasing with the square of speed. Alternative fuels, such as biodiesel, should be a consideration for the future of rail, as emissions fall significantly with content of biodiesel in fuel blends.

Gangwar & Sharma (2014) adopted a WTW approach to quantify the emissions from diesel and electric locomotives in India. Results showed that the accumulated carbon footprint of running electric locomotives was higher, as a consequence of using coal as a primary source in electricity production. They suggest that there should be a judicious mix of both tractions to achieve a balance in environmental efficiency, sustainability and equity.

Washing and Pulugurtha (2015) used WTW analysis to combine the energy efficiencies of each component of the energy pathway into a single energy efficiency value. The focus of this paper was on WTW analysis of electric and hydrogen light rail. The inefficiencies of the hydrogen train's power plant and hydrogen production process are apparent in the hydrogen train's WTW efficiency value of 16.6–19.6%. The electric train, due to improved pathway efficiencies, uses substantially less feedstock energy with a WTW efficiency value of 25.3%. While this result is specific to Charlotte, North Carolina, the electric train efficiency is influenced by the main source of electricity production – it is 24.6% in Cleveland, Ohio (with domination of coal) and 50.3% in Portland, Oregon (with domination

of hydroelectric power).

The main limitations and issues identified concerning the available literature on WTW analysis in railway passenger transportation, alongside with those addressed in the previous section, are:

- lack of comprehensive WTW evaluation of different railway passenger vehicles, especially powered by alternative energy options, and different driving conditions;
- lack of consistent formulation and comprehensive studies of different energy carriers pathways, especially for alternative fuels, as well as different energy and electricity generation mixes.

Limitations listed first can be addressed by developing detailed vehicles models and simulation tools based on bottom-up methods which would enable identification and analysis of different technological and operational parameters, related to technology improvements, driving conditions and strategies, etc. Additionally, limitations related to WTT stage can potentially be addressed using a formal thermoeconomic analysis, which uses exergy to account for the consumption of primary resources and to allocate it over multiple products (Orsi et al., 2016), where exergy can be defined as "the amount of useful work extractable from a generic system when it is brought to equilibrium with its reference environment through a series of reversible processes in which the system can only interact with such environment" (Moran et al., 2012).

5 Railway Life Cycle Assessments

A Life Cycle Assessment (LCA) is an environmental management tool used to understand and compare how a product or a service is provided from "cradle to grave" – a term used to describe the life cycle of a product or a service from its first derivatives to its end-use (Banar and Özdemir, 2015). The main phases of each LCA are shown in Fig. 2.

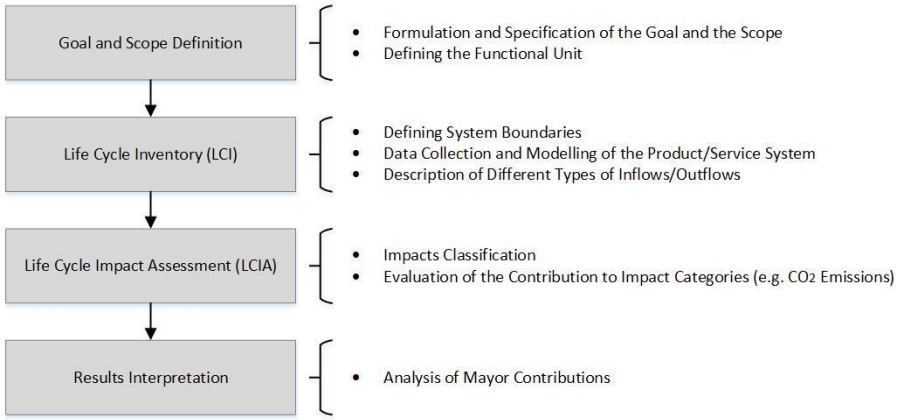


Figure 2: Main phases of a LCA

There are two different methodologies in the literature for LCA, which can also be combined into a hybrid model (Jones, 2017), depending on the goal, scope and constraints of the study:

(i) Process-based LCA – performed by mapping all processes associated with all life cycle phases of the product/service.
(ii) Economic input-output analysis-based (EIO-LCA).

A process-based methodology is performed by mapping all processes associated with all life cycle phases of the project, where inputs (e.g., electricity, steel) and outputs (e.g., air emissions, water discharges) associated with each process are included which enables the total environmental load to be calculated (Jones, 2017). It provides very detailed analysis, but it can require a vast amount of data to include upstream processes (Noori et al. 2013, 2015).

EIO-LCA combines an economic input-output (I-O) model with environmental data so the environmental load of the production of the associated commodities is determined. The I-O model identifies the interdependencies between the different economic sectors and includes the effects of the supply chain. This methodology provides an inclusive and industry-wide analysis allowing for system level comparisons, but can lack the detail of a process-based LCA because it aggregates data to industry sectors (Noori et al. 2013, 2015; Jones, 2017).

The goal and scope definition is the first stage in a LCA study. The significance of this stage is that the decisions made in this phase guide the entire study. Also, functional unit (FU) is defined in this phase. FU is defined as a reference unit for normalization of a quantified performance of a certain product (Guinee et al., 2002), and typically used FUs in railway studies are vehicle kilometer (vkm) or passenger kilometer (pkm) traveled. Several functional units can be used depending on a question that is being informed. Normalization per Vehicle Kilometer Travelled (VKT) is useful for evaluating specific corridor but this does not account passenger carrying capability.

Life cycle inventory (LCI) is one of the most effort consuming stage as it involves the collection, compilation and interpretation of the actual system data in line with the goals and scope of the study and as an input to subsequent life cycle impact assessment stage. Compiling the relevant data for extensive system boundary and collecting it scattered across various sources is usually the major challenge (Shinde et al., 2018).

The Life Cycle Impact Assessment (LCIA) identifies the environmental impacts of LCI results by associating inventory data with potential environmental impact categories (e.g. global warming, acidification, etc.). Several methods and tools are developed for assessing the environmental impacts, such as CML 2001 (University of Leiden, 2001), ReCiPe (Goedkoop et al., 2013), and others.

LCA papers for railway passenger transportation are listed in Table 1, together with the geographical information on the study (country), transport mode considered and system boundaries. Regarding the system boundaries defined, in most cases if the rail infra-structure already exists and the alternative scenarios do not entail developing a new rail network from scratch, the environmental impacts related to the infrastructure are excluded. If the study concerns construction of the new line, such as a high-speed rail line, the infrastructure is then included in the analysis.

Table 1: LCA papers by country and area of contribution

Publication	Country	Transport Mode / Area	System Boundary
Von Rozycki et al. (2003)	Germany	High-speed rail	Infrastructure construction and operation (including buildings); vehicle manufacturing, traction, and maintenance
Castella et al. (2009)	South Korea	High-speed rail	Rail car-bodies raw material production, manufacturing, use, and end-of-life
Stripple and Uppenberg (2010)	Sweden	Rail transportation (passenger and freight)	Infrastructure construction and maintenance (including tunnels, bridges, track foundation and track, stations, freight terminals, signalling system); manufacturing and operation of train carriages
Åkerman (2011)	Sweden	High-speed rail	Infrastructure construction, maintenance, and operation; vehicle manufacturing, maintenance, and use
Chang and Kendall (2011)	USA	High-speed rail	Infrastructure construction (buildings and stations excluded, as well as infrastructure operation and vehicles)
Chester and Horvath (2012)	USA	Passenger transportation (high-speed rail, emerging automobiles and aircraft technology)	Infrastructure construction, operation, maintenance, and insurance; vehicle manufacturing, operation, maintenance, and insurance
Chan et al. (2013)	Canada	Commuter rail system	Vehicles operation; fuel production/electricity generation
Banar and Özdemir (2015)	Turkey	Rail passenger transportation (high-speed rail, conventional rail)	Infrastructure (production and distribution of electrical energy, extraction and production of raw materials, construction, maintenance and operation of lines and waste disposal); rail operation (production and distribution of electrical energy, extraction and production of raw materials, production, maintenance and operation of railway vehicles); and waste disposal
Del Pero et al. (2015)	Italy	Heavy metro	Train material acquisition, manufacturing, use, and end-of-life
de Andrade and D'Agosto (2016)	Brasil	Metro line	Infrastructure construction, and operation; train manufacturing, operation, and maintenance
Dimuola et al. (2016)	Greece	Road and rail transportation (passenger and freight)	Infrastructure construction (road and rail); operation (road and rail)
Jones et al. (2017)	Portugal	High-speed rail	Track construction, operation and maintenance, and disposal; train manufacturing, operation, maintenance, and disposal
Shinde et al. (2018)	India	Suburban rail	Infrastructure construction and maintenance; vehicle manufacturing, maintenance, and operation

Stripple and Uppenberg (2010) developed environmental product declarations (EPD) for newly constructed Bothnia Railway Line in Sweden. Comprehensive life cycle model of the entire railway system was developed. Results showed the greatest contribution to the project's global warming potential (GWP) from the railway infrastructure (93.3%), while the trains operation contribution is just 6.7%, with the main GHG fossil-based CO_2, while emissions of N_2O only give minor contribution. The infrastructure construction stands for the main part of the GHG emissions, with the main source in the production of different materials, while the actual construction work is much smaller. Emissions from the infrastructure and trains operation are very small due to the use of green electric power (the electric power production mix in Sweden in year 2008 was 99.2% hydropower and 0.8% based on biomass fuel).

Akerman (2011) used LCA to research the mitigating climate change effects of a proposed Swedish high-speed rail track and found significant reduction of greenhouse gas emissions because of transportation modes shifting to HSR. The life cycle emissions reductions are found to be 550,000 tons of $CO_{2\text{-eq}}$ per annum by 2025/2030 with almost 60% of this coming from a shift from truck to rail freight and 40% from a shift from air and road travel to high-speed rail travel. However, new railway construction and maintenance may weaken that effect.

Chang and Kendall (2011) performed a process-based LCA study on a greenhouse gas emissions estimation in the construction of the California high-speed rail (CAHSR) infrastructure with specification of several infrastructure types depending on terrain. They found that 80% of the infrastructure emissions resulted from material production, and that tunneling and aerial structures which took only 15% of the route's length, resulted in 60% of the emissions.

Chan et al. (2013) investigated the GHG impact of several alternatives for the commuter rail system in Montreal, Canada. Evaluation of environmental performance and cost of current diesel powered trains against electric powered trains and hydrogen fuel cell system using steam methane reforming (SMR) and wind energy was carried out. They found that electrification, with hydroelectric power, would reduce GHG emissions by more than 98% relative to the current diesel powered trains, while using hydrogen would bring a reduction of 24% or 82% if produced via SMR or via renewable electrolysis, respectively.

Banar and Özdemir (2015) conducted a life cycle assessment and life cycle cost analysis of Turkey's railway transport systems aiming to assess the environmental and economic impact and to serve as guidance for future railways projects to reduce their life-cycle environmental impact in Turkey. The total environmental load of high-speed rail is shared by infrastructure and operations, with percentages of 58% and 42%, respectively. On the other hand, for conventional rail, infrastructure created 39% of the total environmental load, while operations had 61%.

Del Pero et al. (2015) performed a predictive LCA of a heavy metro train investigating on the recyclability/recoverability of the metro vehicles. A sensitivity analysis aimed at defining the variation of environmental impact depending on Vehicle Occupancy (VO) was also carried out. Results showed that the greatest impact results from the operation phase, as well as that there are great possibilities for improvements in this phase.

de Andrade and D'Agosto (2016) assessed the energy used and the emissions produced and avoided in the lifecycle of a new line of the metro network in Rio de Janeiro, built as a requirement for hosting the Olympic Games in 2016. Infrastructure construction, train manufacture, maintenance, infrastructure operation and train operation were considered in the 60-year lifecycle. They concluded that the increase in the renewable energy share in electricity generation and improvements in the production of cement and steel, are the key

factors in reducing emissions produced during the life cycle.

Shinde et al. (2018) performed an LCA for the Mumbai Suburban Railway with the objective of developing a comprehensive methodology for environmental evaluation of suburban railway projects in terms of energy consumption and relevant impact categories. The scope of the research comprises the construction and maintenance of railway infrastructure such as tracks, power supply installations, foot over bridges and platforms, in addition to manufacturing, maintenance and the operation phase of Electric Multiple Units (EMUs). The results show that operation phase is the main contributor (87-94%) to the total environmental impact, whereas the contribution of remaining life cycle phases is relatively insignificant (6-13%), mainly due to electricity production from non-renewable sources in India. The material and energy intensive rails entail the major contribution to construction phase (24-57%) and maintenance phase (46-71%),

Based on the existing literature on LCA in railway passenger transportation, main limitations and issues in environmental impact assessment from a life cycle perspective are identified as:

- lack of comprehensive LCA evaluations that include detailed WTW analysis and consumption phase models;
- lack of extensive comparative and sensitivity analyses that assess the effects of different scenarios (e.g. different occupancy rates), as well as technological changes, operational and policy measures;
- lack of elaborate and detailed studies that analyze emissions from the construction/production and end-of-life phases.

Main challenge in performing LCA is the incorporation of detailed emission models from the consumption phase and the WTW pathway, together with addressing the issues and challenges identified in these studies. Although there is an increasing attention on environmental issues regarding construction/production and end-of-life (EoL) phases, the impact of these activities in terms of GHG emissions is still neglected. Initiatives such as the assessment framework proposed by the European association of railway supply industry (UNIFE, 2014) which is to be used on a voluntary basis, represent a good starting point to address this issue.

6 Discussion

Based on the review of the existing research, the main challenge is answering how the available partial assessments can be brought together and, together with filling the identified gaps, allow to conduct a comprehensive LCA which will produce real-world emissions estimations.

Since the total life cycle emissions are directly influenced and dependent on the direct energy consumption and emissions, consumption phase represents the main driver of the total life cycle emissions from the rail passenger service. An effective approach could be the development of detailed direct emissions estimation models and setting them as the central and starting point in future LCA studies. Extending the existing consumption phase models by incorporating real-life conditions that influence consumption and emissions, mentioned in Sec. 3, would serve as the main input for a wider-scope WTW analysis, and subsequent LCA. Real direct measurements can be a valuable input in microscopic bottom-up models development, calibration and validation. The development of mesoscopic models which combine the preciseness of microscopic models while requiring only little more

information than the rough estimating macroscopic top-down models could help in overcoming the limitations such as high complexity and data availability. An example of such models can be found in Kirschstein and Meisel (2015) for intermodal rail/road transport.

Common approach in assessing the total WTW emissions is by multiplying the total energy consumption with the WTW emission coefficients, which usually represent adopted average values and may lead to incorrect and biased estimations. Since the real value of this coefficient is highly influenced and directly dependent on the actual energy carrier pathway, formulating and determining all the processes within the different energy carrier pathways – together with associated energy consumption and emissions – is of great importance. Elements and aspects such as primary energy source extraction, energy carrier production and distribution, electricity generation mix should explicitly be taken into account. Integrated with an effective bottom-up vehicle models, which are easy to calibrate for different technological and operational parameters and which would enable assessment of direct energy consumption and emissions, it would allow obtaining factual WTW emissions and generate important input for a subsequent LCA.

Incorporating detailed consumption and WTW models into LCA could help not only in actual emissions assessment, but also in identifying the effects of different technological changes, as well as operational and policy measures. Contrary to the common approximate top-down estimation approaches found in LCA studies, it would potentially enable more consistent estimations from the vehicles/infrastructure operation phase, especially important in case of comparing different options and measures.

Another issue identified in LCA studies is the lack of elaborate and detailed studies that analyze emissions from the construction/production and EoL phases. Although some of the train manufacturers started producing the environmental product declarations (EPDs) for their trains, this number is still relatively small. These EPDs could be valuable source of information for the LCA studies, especially regarding the materials usage, energy consumption and environmental impact from the production phase. Concerning the EoL phase, contrary to the low environmental impact of railway transport with respect to other transport modes, the amount of EoL waste generated by rolling stock in relation to the number of road vehicles is significant. The study by Delogu et al. (2017) gave an overview of EoL railway vehicles management issues and analyzed the recoverability/recyclability rate for three types of railway vehicles (electric metro, diesel commuter train and high-speed electric train). As stated in this study, the disposal of a railway passenger vehicle in terms of weight of the obtained waste corresponds to 36-42 road passenger vehicles, although there is no consideration of the comparative capacity of the vehicles (railway car in automobile equivalents) or the comparative service life of road and railway vehicles, both of which are important considerations.

7 Conclusions and Future Research Directions

This paper presented a review of existing research on life cycle emissions from railway passenger services. Studies and approaches focused on the direct emissions from the consumption phase are presented first, followed by wider-scope WTW analyses which observe energy carrier life cycle, and LCA studies which encompass infrastructure and/or vehicles life cycles and associated emissions. A comprehensive analysis of existing models enabled identifying the research gaps and addressing the main issues and challenges in assessing the overall impact in terms of GHG emissions. Additionally, possibilities in addressing the limitations and filling the identified gaps are given.

Future research will include development of a framework for life cycle emissions estimation and prediction, observing both conventional and alternative energy options for railway passenger transport. First, detailed pathways will be determined, including processes related to: primary resource recovery, extraction and transportation to the construction/production facilities; activities in construction/production; distribution of the energy carrier to the vehicles; operation and maintenance; and end-of-life activities (recycling, reuse and disposal). Environmental impacts from all processes and sub-processes will be evaluated by developing and employing bottom-up methods. Results will be validated through real-life measurements and comparison with the results of other world-wide studies. Special attention will be given to the efficiency of the system elements. Sensitivity analysis will be carried out with the aim of assessing the possibilities in improving the environmental impact of the rail passenger service, and will include technological, operational and policy measures.

Acknowledgment

This work is supported by Arriva Personenvervoer Nederland B.V. within the PhD project "Improving sustainability of regional railway services".

References

Abdelaal, M.M., Hegab, A.H., 2012. "Combustion and emission characteristics of a natural gas-fueled diesel engine with EGR", *Energy Conversion and Management*, vol. 64, pp. 301-312.

ADVISOR, 2003. *Advanced Vehicle Simulator*, available at: http://adv-vehicle-sim.sourceforge.net/

Åkerman, J., 2011. "The role of high-speed rail in mitigating climate change – The Swedish case Europabanan from a life cycle perspective", *Transportation Research Part D*, vol. 16, pp. 208-217.

ANL, 2016. *The Greenhouse gases, Regulated Emissions, and Energy use in Transportation (GREET) model*, Argonne National Laboratory, available at: https://greet.es.anl.gov/

Banar, M., Özdemir, A., 2015. "An evaluation of railway passenger transport in Turkey using life cycle assessment and life cycle cost methods", *Transportation Research Part D*, vol. 41, pp. 88-105.

Castella, P.S., Blanc, I., Ferrer, M.G., Ecabert, B., Wakeman, M., Manson, J., Emery, D., Han, S., Hong, J., Jolliet, O., 2009. "Integrating life cycle costs and environmental impacts of composite rail car-bodies for a Korean train", *The International Journal of Life Cycle Assessment*, vol. 14, pp. 429-442.

Chan, S., Miranda-Moreno, L., Patterson, Z., 2013. "Analysis of GHG Emissions for City Passenger Trains: Is Electricity an Obvious Option for Montreal Commuter Trains?", *Journal of Transportation Technologies*, vol. 3, pp. 17-29.

Chester, M., Horvath, A., 2012. "High-speed rail with emerging automobiles and aircraft can reduce environmental impacts in California's future", *Environmental Research Letters*, vol. 7, pp. 1-11.

CleanER-D, 2012. "Clean European rail-diesel, impact and performance of alternative fuel in rail applications", *June 2012*.

Correa, G., Muñoz, P., Falaguerra, T., Rodriguez, C.R., 2017. “Performance comparison of conventional, hybrid, hydrogen and electric urban buses using well to wheel analysis”, *Energy*, vol. 141, pp. 537–549.

Davis, W.J., 1926. “The tractive resistance of electric locomotives and cars”, *General Electric Review*, 29 . October

de Andrade, C.E.S., D'Agosto, M.A., 2016. “Energy use and carbon dioxide emissions assessment in the lifecycle of passenger rail systems: the case of the Rio de Janeiro Metro”, *Journal of Cleaner Production*, vol. 126, pp. 526-536.

Del Pero, F., Delogu, M., Pierini, M., Bonaffini, D., 2015. “Life Cycle Assessment of a heavy metro train”, *Journal of Cleaner Production*, vol. 87, pp. 787-799.

Delogu, M., Del Pero, F., Berzi L., Pierini, M., Bonaffini, D., 2017. “End-of-Life in the railway sector: Analysis of recyclability and recoverability for different vehicle case studies”, *Waste Management*, vol. 60, pp. 439-450.

Dimoula, V., Kehagia, F., Tsakalidis, A., 2016. “A Holistic Approach for Estimating Carbon Emissions of Road and Rail Transport Systems”, *Aerosol and Air Quality Research*, vol. 16, pp. 61–68.

Dreier, D., Silveira, S., Khatiwada, D., Fonseca, K.V.O., Nieweglowski, R., Schepanski, R., 2017. “Well-to-Wheel analysis of fossil energy use and greenhouse gas emissions for conventional, hybrid-electric and plug-in hybrid-electric city buses in the BRT system in Curitiba, Brazil”, *Transportation Research Part D*, vol. 58, pp. 122–138.

EC, 2014. *Well-to-Wheels Analysis of Future Automotive Fuels and Powertrains in the European Context*, European Commission JRC Technical Reports, available at: https://iet.jrc.ec.europa.eu/about-jec/sites/iet.jrc.ec.europa.eu.about-jec/files/documents/report_2014/wtt_report_v4a.pdf, accessed: 15.09.2018.

EEA, 2017. *Greenhouse gas emissions from transport*, Report of the European Environment Agency.

Esters, T., Marinov, M., 2014. “An analysis of the methods used to calculate the emissions of rolling stock in the UK,” *Transportation Research Part D*, vol. 33, pp. 1-16.

EU, 2017. “Statistical Pocketbook 2017: EU Transport in Figures”, *European Union*.

Gangwar, M., Sharma, S.M., 2014. “Evaluating choice of traction option for a sustainable indian railways,” *Transportation Research Part D*, vol. 33, pp. 135–145.

Garcia, A.G., 2010. “Energy Consumption and Emissions of High-Speed Trains”, *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2159, pp. 27–35.

Goedkoop, M., Heijungs, R., Huijbregts, M., De Schryver, A., Struijs, J., van Zelm, R., 2013. *ReCiPe 2008 A life cycle impact assessment method which comprises harmonized category indicators at the midpoint and the endpoint level*, Ministerie van VROM, Den Haag.

Guinee, J.B., Gorree, M., Heijungs, R., Huppes, G., Kleijn, R., de Koning, A., van Oers, L., Sleeswijk, A.W., Sangwon, S., Udo de Haes, H.A., de Bruijn, J.A., van Duin, R., Huijbregts, M.A.J., 2002. *Handbook on Life Cycle Assessment: Operational Guide to the ISO Standards*, Kluwer Academic Publishers, Dordrecht, The Netherlands.

Hoffrichter, A., Miller, A.R., Hillmansen, S., Roberts, C., 2012. “Well-to-wheel analysis for electric, diesel and hydrogen traction for railways,” *Transportation Research Part D*, vol. 17, no. 1, pp. 28-34.

IPCC, 2007. *Climate Change 2007: The Physical Science Basis*, available at: https://www.ipcc.ch/report/ar4/wg1/, accessed: 20.09.2018.

IPCC, 2014. *Climate Change 2014: Mitigation of Climate Change*, available at: https://www.ipcc.ch/report/ar5/wg3/, accessed: 25.09.2018.

Johnson, G.R., Jayaratne, E.R., Lau, J., Thomas, V., Juwono, A.M., Kitchen, B., Morawska L., 2013. “Remote measurement of diesel locomotive emission factors and particle size distributions”, *Atmospheric Environment*, vol. 81, pp. 148–157.

Jones, H., Moura, F., Domingos, T., 2017. “Life cycle assessment of high-speed rail: a case study in Portugal”, *The International Journal of Life Cycle Assessment*, vol. 22, pp. 410-422.

Kirschstein, T., Meisel, F., 2015. “GHG-emission models for assessing the eco-friendliness of road and rail freight transports”, *Transportation Research Part B*, vol 73, pp. 13-33.

Lapuerta, M., Armas, O., Rodriguez-Fernandez, J., 2008. “Effect of biodiesel fuels on diesel engine emissions”, *Progress in Energy and Combustion Science*, vol. 34, pp. 198–223.

Li, M., Zhang, X., Li, G., 2016. “A comparative assessment of battery and fuel cell electric vehicles using a well-to-wheel analysis,” *Energy*, vol. 94, pp. 693–704.

Lindgreen, E.B.G., Sorenson, S.C., 2005. *Driving resistance from railroad trains*, DTU Orbit, Annual Report.

Moran, M.J., Shapiro, H.N., Boettner, D., Bailey, M., 2012. *Fundamentals of engineering thermodynamics*, John Wiley & Sons.

Network Rail, n.d. “Comparing environmental impact of conventional and high speed rail”, *Planning and Regulation, New Lines Programme*.

Noori, M., Kucukvar, M., Tatari, O., 2013. “A macro-level decision analysis of wind power as a solution for sustainable energy in the USA”, *International Journal of Sustainable Energy*, vol. 2013, pp. 1–16.

Noori, M., Kucukvar, M., Tatari, O., 2015. “Economic input-output based sustainability analysis of onshore and offshore wind energy systems”, *International Journal of Green Energy*, vol. 12, pp. 939–948.

Orsi, F., Muratori, M., Rocco, M., Colombo, E., Rizzoni, G., 2016. “A multi-dimensional well-to-wheels analysis of passenger vehicles in different regions: Primary energy consumption, CO2emissions, and economic cost”, *Applied Energy*, vol. 169, pp. 197–209.

Papagiannakis, R.G., Hountalas, D.T., 2003. “Experimental investigation concerning the effect of natural gas percentage on performance and emissions of a DI dual fuel diesel engine”, *Applied Thermal Engineering*, vol. 23, pp. 353–365.

Papagiannakis, R.G., Kotsiopoulos, P.N., Zannis, T.C., Yfantis, E.A., Hountalas, D.T., Rakopoulos, C.D., 2010a. “Theoretical study of the effects of engine parameters on performance and emissions of a pilot ignited natural gas diesel engine”, *Energy*, vol. 35, pp. 1129–1138.

Papagiannakis, R.G., Rakopoulos, C.D., Hountalas, D.T., Rakopoulos, D.C., 2010b. “Emission characteristics of high speed, dual fuel, compression ignition engine operating in a wide range of natural gas/diesel fuel proportions”, *Fuel*, vol. 89 (7), pp. 1397-1406.

Park, D., Yoon, Y., Kwon, S.B., Jeong, W., Cho, Y., Lee, K., 2012. “The effects of operating conditions on particulate matter exhaust from diesel locomotive engines”, *Science of the Total Environment*, vol. 419, pp. 76-80.

Poompipatpong, C., Cheenkachorn, K., 2011. “A modified diesel engine for natural gas operation: performance and emission tests”, *Energy*, vol. 36 (12), pp. 6862-6866.

RSSB, 2007. “Quantification of Benefit of Train Mass Reduction”, *T712 - August 2010*.

RSSB, 2010. “Investigation into the use of bio-diesel fuel on Britain's railways”, *T697 - August 2010*.

Scheepmaker, G.M., Goverde, R.M.P., Kroon, L.G., 2017. “Review of energy-efficient train control and timetabling”, *European Journal of Operational Research*, vol. 257, no. 2, pp. 355–376.

Shinde, A.M., Dikshit, A.K., Singh, R.K., Campana, P.E., 2018. “Life cycle analysis based comprehensive environmental performance evaluation of Mumbai Suburban Railway, India”, *Journal of Cleaner Production*, vol. 188, pp. 989-1003.

Stripple, H., Uppenberg, S., 2010. *Life Cycle Assessment of Railways and Road Transports*, Swedish Environmental Research Institute, Goteborg, Sweden.

SYSTRA, 2011. “Factors affecting carbon impacts of HSR”, *Carbon Impacts HS2*, Vol. 3.1(1), pp. 4–11.

UIC, 2003. *Evaluation of energy efficiency technologies for rolling stock and train operation of railways*, International Union of Railways.

UIC, 2012. *Moving towards sustainable mobility*, International Union of Railways, Paris, France.

UN, 1987. *Report of the World Commission on Environment and Development: Our Common Future*, Gro Harlem Brundtland, Oslo, 20 March 1987.

UN, 2015. *Paris Agreement*, Paris, France.

UNIFE, 2014. *A 2030 Framework for Climate and Energy Policies*, available at: http://www.unife.org/component/attachments/?task=download&id=104, accessed on: 20.09.2018.

University of Leiden, 2001. *Centre for Environmental Studies: CML 2001 Characterization Method*, available at: http://www.cml.leiden.edu/.

Van Oort, N., Van der Bijl, R., Verhoof, F., 2017. “The wider benefits of high quality public transport for cities”, *European Transport Conference*, Barcelona, Spain, October 2017.

Van Wee, B., Janse, P., Van den Brink, R., 2005. “Comparing energy use and environmental performance of land transport modes”, *Transport Reviews*, vol. 25 (1), pp. 3-24.

Von Rozycki, C., Koeser, H., Schwarz, H., 2003. “Ecology profile of the German high-speed rail passenger transport system, ICE”, *The International Journal of Life Cycle Assessment*, vol. 8 (2), pp. 83-91.

Washing, E.M., Pulugurtha, S.S., 2015. “Well-to-Wheel Analysis of Electric and Hydrogen Light Rail”, *Journal of Public Transportation*, vol. 18 (2), pp. 74-88.

Woo, J.R., Choi, H., Ahn, J., 2017. “Well-to-wheel analysis of greenhouse gas emissions for electric vehicles based on electricity generation mix: A global perspective”, *Transportation Research Part D*, vol. 51, pp. 340–350.

Xue, J., Grift, T.E., Hansen, A.C., 2011. “Effect of biodiesel on engine performances and emissions”, *Renewable and Sustainable Energy Reviews*, vol. 15 (2), pp. 1098–1116.

Yazdanie, M., Noembrini, F., Dossetto, L., Boulouchos, K., 2014. “A comparative analysis of well-to-wheel primary energy demand and greenhouse gas emissions for the operation of alternative and conventional vehicles in Switzerland, considering various energy carrier production pathways”, *Journal of Power Sources*, vol. 249, pp. 333–348.

A Mixed Integer Linear Programming Approach to a Rolling Stock Rostering Problem with Splitting and Combining

Satoshi Kato [a], Naoto Fukumura [b], Susumu Morito [c]
Koichi Goto [b], Narumi Nakamura [b]
[a] Transport Operation Systems Laboratory
Signalling and Transport Information Technology Division
Railway Technical Research Institute
2-8-38 Hikari-cho, Kokubunji-shi, Tokyo 185-8540, Japan
E-mail: kato.satoshi.58@rtri.or.jp, Phone: +81 (0) 42 573 7311
[b] JR Souken Information Systems
[c] Department of Industrial and Management Systems Engineering, Waseda University

Abstract
Railway operators must schedule resources such as rolling stock and crew in order to operate trains as defined by a timetable. This paper considers scheduling of rolling stock, which is usually done by creating a roster. A roster is a series of trains to be performed by the particular rolling stock. The number of train-sets required to operate a given group of trains is essentially determined by the roster and generation of an efficient roster is essential. Important considerations of the roster generation include maintenance such as pre-departure inspection. On some lines in Japan, splitting and combining are often used to adjust transportation capacity flexibly. Under this type of operation, splitting and combining become necessary. These shunting operations require time and manpower, so it is necessary to reduce the amount of splitting and combining. This paper presents a mixed integer linear programming model so that the amount of splitting and combining is reduced together with the roster length and the distance of empty runs. Results of computational studies will be presented based on real instances of several lines in Japan, indicating the computational effectiveness of the methodology and with respect to the reasonableness of the resultant rosters.

Keywords
Rolling stock rostering, Splitting and combining, Maintenance, Mixed integer linear programming, Travelling salesman problem

1 Introduction

Railway operators must schedule resources such as rolling stock and crew in order to operate trains as defined by a timetable. This paper considers scheduling of rolling stock, which is usually done by creating a roster. A roster is a series of trains to be performed by particular rolling stock, and cyclic execution of the roster theoretically determines which train-sets are assigned to which train. The number of train-sets required to operate a given group of trains is essentially determined by the roster and generation of an efficient roster is essential. Important considerations of the roster generation include maintenance such as pre-departure inspections.

On some lines in Japan, splitting and combining are often used to adjust transportation

capacity flexibly. For example, two train-sets are combined together (say, two 3-car train-sets are combined together, thus effectively forming a 6-car train-set) during morning and evening rush hours, but during the day time, only one train-set is assigned to each train. Under this type of operation, shunting operations of splitting and combining would become necessary. These shunting operations require time and manpower, so it is necessary to reduce the amount of splitting and combining.

Many studies exist on efficient planning and management of rolling stock and a variety of models and algorithms for optimization have been developed. Abbink et al. (2004) give an integer programming model to allocate train types so that shortage of capacity during rush hours is minimized. One fundamental study in the field is Alfieri et al. (2006), presenting a multi-commodity flow model of rolling stock circulation for determining the appropriate number of train units of different types together with their efficient circulation on a single Dutch line. These studies, however, do not consider maintenance.

Giacco et al. (2014) formulate a rolling stock rostering problem with maintenance considerations as a generalized travelling salesman problem (TSP) where a roster is represented as a tour of the associated network. Borndörfer et al. (2016) and Reuther and Schlechte (2018) dealing with the same problem give a mixed integer programming model based on a hypergraph. Morooka et al. (2017) present computational experiences with the Giacco's model and its variants using real instances of several lines in Japan. Nishi et al. (2017) propose a column generation and Lagrangian heuristics for rostering with maintenance considerations.

Regarding research with explicit considerations for splitting and combining, Fioole et al. (2006) present an integer programming model of rolling stock circulation with objective criteria such as operational costs, service quality, and reliability including reduction of shunting movements. Peeters and Kroon (2008) develop a branch-and-price algorithm for a similar problem. These studies, however, do not consider maintenance. Tsunoda et al. (2015) give a multi-commodity network flow model to estimate the required number of train units to meet future traffic demands under a splitting and combining policy, but do not consider maintenance or reduction of shunting operations.

This paper presents an optimization-based methodology to construct a roster with maintenance considerations under the existence of splitting and combining. More specifically, a mixed integer linear programming (MILP) model is proposed based on a travelling salesman problem with multiple arcs between nodes so that the amount of splitting and combining is reduced together with the roster length and the distance of empty runs. Results of computational studies will be presented based on real instances of several lines in Japan with roughly 100 to 200 trains, indicating the effectiveness of the methodology computationally and with respect to the reasonableness of the resultant rosters.

2 Problem Definition

2.1 Rolling Stock Rostering Problem

A rolling stock schedule is produced by covering all trains shown on a train timetable. Since the number of rolling stock required to meet the train timetable requirements is determined by a rolling stock schedule, it is necessary to produce an efficient schedule. A rolling stock schedule must satisfy the minimum time interval between the arrival of a train and its subsequent departure, together with maintenance requirements as described

shortly.

A rolling stock schedule is achieved as a cyclic schedule called a roster. A roster is a finite series of daily schedules performed by a train-set, called duties, which are performed by a particular train-set in a cyclic fashion. The number of train-sets needed to implement a given roster coincides with the number of duties in the roster, which is sometimes called the roster length. Naturally, generating a roster with the shortest roster length is desirable. Rosters are normally created for each type of train-set.

Important considerations of the roster generation are a type of maintenance that must be performed within a specified interval. Four types of maintenance occur in Japan; namely, pre-departure inspection, regular inspection, bogie inspection, and general inspection. This paper focuses on pre-departure inspections, whose intervals are the shortest. This is because other inspection types often take at least one day, while pre-departure inspections only need a few hours and a train-set that has been inspected or is to be inspected is often assigned to trains within the same calendar day; thus, when to perform inspections must be determined. Throughout the rest of this paper, pre-departure inspections are simply called "inspections".

The scenario considered in this paper often occurs in non-metropolitan cities, and deals with the case where only k-car train-sets are available, where typical values of k are 2 to 4. During morning and evening rush hours, some trains are operated by combining two k-car train-sets, thus effectively yielding a $2\times k$-car train-set. In contrast, except for these rush hours, services are operated by single k-car train-sets. This type of operation gives an effective way to utilize the limited number of rolling stock, and is particularly suited for lines in which rush hour demands differ substantially from those of other time periods.

In order to achieve the above types of operations, splitting and combining of train-sets would become necessary. Here, splitting means to split a combined train-set into two separate train-sets, and combining means to combine two separate train-sets together. Splitting and combining allow flexible adjustments of traffic capacity, but additional time and work of operators will be required, so reducing the amount of splitting and combining is desirable.

Figure 1 shows an example of a rolling stock schedule. In this example, there are four stations, Stations A through D, and nine trains (in service), Trains 1 through 9. Station C is adjacent to a rolling stock depot, where the inspection of a train-set can be performed. A circle means the start of a duty and a triangle means its end. The double lines such as Trains 2, 6, and 9 indicate what we call "double-unit trains" that are operated by two train-sets combined together. On the other hand, single lines such as Trains 1, 3, 4, 5, 7, and 8 indicate what we call "single-unit trains" that are operated by a single train-set.

For those double-unit trains shown by double lines, the left lines indicate the front side of the train-sets, and the right lines indicate the rear side of the train-sets. The two train-sets operating Train 2 are assigned to two distinct trains, Trains 5 and 7, upon arrival at Station A, thus indicating the existence of splitting at Station A. In contrast, combining at Station D would be performed before the departure of Train 6. For this rolling stock schedule, one splitting as well as one combining would be required. Maintenance is performed at the depot next to Station C upon the arrival of Train 4, after which the train-set is deadheaded to Station D.

Figure 2 shows a roster associated with the rolling stock schedule given in Figure 1. There are six duties, Duties 1 through 6, which are assigned to train-sets in this order. After performing Duty 6, a train-set is assigned to Duty 1 again, and this cycle will be repeated. Note that the ending station of a duty coincides with the starting station of the

subsequent duty so that the train-set can continue the schedule cyclically. The number of duties, namely, the roster length, corresponds to the minimum number of train-sets required to achieve the schedule. Duty 3 includes an inspection and this roster contains one inspection every six days, making the schedule feasible if the upper limit of the inspection cycle is six days. The letters F and B shown next to Trains 2, 6, and 9 indicate Front and Back, meaning the front and back halves of the train-sets, respectively. For example, Train 2 is operated by combining the train-sets of Duty 2 and Duty 5.

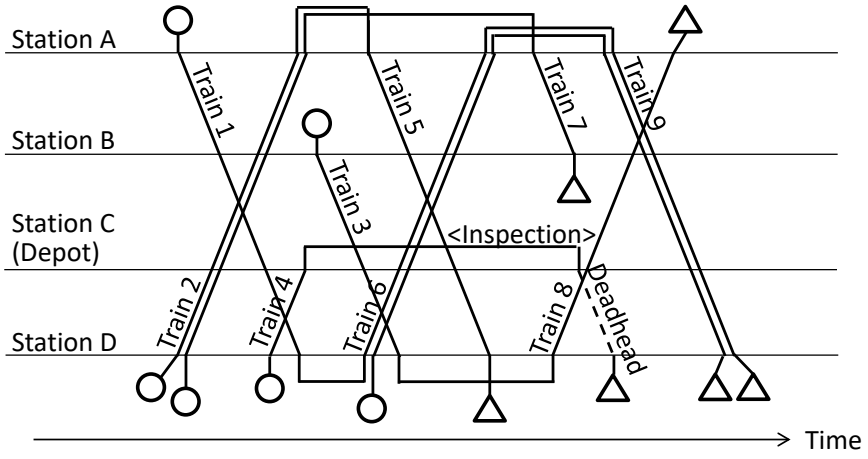


Figure 1. Sample rolling stock schedule

Duty	
1	A ○ Train 1 D Train 6(F) A Train 9(B) △ D
2	D ○ Train 2(F) A Train 5 △ D
3	D ○ Train 4 C <Inspection> Deadhead – – △ D
4	D ○ Train 6(B) A Train 9(F) △ D
5	D ○ Train 2(B) A Train 7 △ B
6	B ○ Train 3 D Train 8 △ A

Figure 2. Sample rolling stock roster

2.2 Splitting and Combining

This paper considers rolling stock rostering with splitting and combining, and seeks the generation of practical rolling stock schedules. For this purpose, it is necessary to identify where splitting and combining occur and to count the exact number of splittings and combinings. Judging when splitting and combining occur, however, is complicated because doing so depends not only on the required number of train-sets before and after train connections, but also on their train positions.

Splitting will always be necessary to connect from a double-unit train to a single-unit train. Similarly, combining will be necessary to connect from a single-unit train to a double-unit train. The difficulties occur when a double-unit train is connected to another double-unit train, where splitting and combining may or may not occur. When the order of the two train-sets of a double-unit train is the same for the subsequent double-unit train, splitting or combining are not required. On the other hand, there may be a case requiring splitting and combining because the order of the two train-sets of a double-unit train is reversed for the subsequent double-unit train. It is also possible that two train-sets of a double-unit train are connected to two distinct double-unit trains, which require splitting and combining. Therefore, it is necessary to look at train positions, since the existence or non-existence of splitting and combining depends on the train positions.

Figure 3 shows several types of train connections indicating complexities due to splitting and combining. In Figure 3a, Train 2 connects to Train 1 after changing the direction of movement, thus requiring no splitting or combining. In Figure 3b, though, train positions are reversed between the two trains, thus requiring one splitting and one combining. Considering connections between two incoming trains and two outgoing trains, Figure 3c shows a case where Train 2 connects to Train 1 and Train 4 to Train 3 keeping the same train positions, thus requiring no splitting or combining. However, connections as in Figure 3d are possible where splitting will be required after the arrivals of Trains 2 and 4, and also combining before the departures of Trains 1 and 3. Avoiding such connections as in Figure 3d which require many splittings and combinings is desirable.

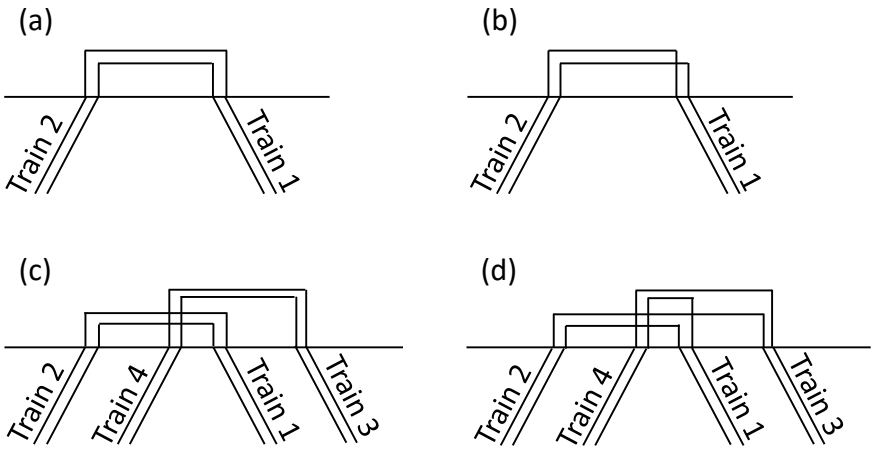


Figure 3. Complication of splitting and combining

We now state our rostering optimization problem with splitting and combining.

1. Given a number of an identical type of train-sets, a single roster would be constructed.
2. Given a set of trains in a given timetable to which train-sets are assigned, the number of train-sets to be assigned to each train is specified as either 1 or 2 in advance. No train requires three or more train-sets. We call those trains operated by a single train-set "single-unit trains", while those trains operated by two train-sets, are called "double-unit trains".
3. Maintenance requirements call for a single type of pre-departure inspection to be performed within minimum and maximum intervals measured in days.
4. The locations and time period during which inspections could be performed are known in advance. Generally, there are multiple locations for maintenance.
5. Empty runs could be inserted as needed.
6. Performance measures include the roster length, the total distance of empty runs, and the amount of splitting and combining, and their weighted sum is to be minimized.

3 Mixed Integer Linear Programming Model for a Rolling Stock Rostering Problem with Splitting and Combining

3.1 Network Model

Our MILP model is based on the roster optimization model of Giacco et al. (2014), where a network is considered in which a node corresponds to each train, and an arc to the connection between trains. We now describe how arcs are drawn in the network. For each connection from node i (its associated train) to node j (its associated train), we check if an arc can be drawn by grouping them into four different types, as follows.

(i) No empty run and no inspection

(ii) Inspection and no empty run

(iii) Empty run and no inspection

(iv) Empty run and inspection

In the following, the details of arc settings for each type will be described. Notations used for the network model are defined in Table 1.

Table 1: Notations for the network model

Notation	Definition
arr_time_i	Arrival time at the destination of the train corresponding to node i
dep_time_i	Departure time at the origin of the train corresponding to node i
arr_sta_i	Destination station of the train corresponding to node i
dep_sta_i	Origin station of the train corresponding to node i
emp_time_{ij}	Time of empty run from destination station of the train corresponding to node i to origin station of the train corresponding to node j
ins_min_time	Earliest possible start time of inspection
ins_max_time	Latest possible completion time of inspection
ins_time	Time required for inspection
$interval\_time$	Minimum time interval between two trains

Type 1: No Empty Run and No Inspection

An arc is drawn if arr_sta_i is equal to dep_sta_j. If $arr\_time_i + interval\_time \leq dep\_time_j$, the arc is set as a same-day arc. That is, the date remains the same after passing through the arc. If $arr\_time_i + interval\_time > dep\_time_j$, the arc is set as a next-day arc. That is, the date changes by one after passing through the arc.

Type 2: Inspection and No Empty Run

An arc is drawn if arr_sta_i is equal to dep_sta_j, and also if arr_sta_i is a station capable of performing an inspection.

1. Case where an inspection is performed the same day
   An arc is drawn as a same-day arc if the following condition is satisfied:

$$\max\{arr\_time_i, ins\_min\_time\} + ins\_time \leq \min\{dep\_time_j, ins\_max\_time\} \quad (1)$$

   Judge if an inspection is possible within the same day.
2. Case where an inspection is performed the next day
   An arc is drawn as a next-day arc if the following condition is satisfied:

$$\max\{arr\_time_i, ins\_min\_time\} + ins\_time > \min\{dep\_time_j, ins\_max\_time\} \wedge ins\_min\_time + ins\_time \leq dep\_time_j \quad (2)$$

   Here an inspection can be started at ins_min_time the next day.

Type 3: Empty Run and No Inspection

An arc is drawn when arr_sta_i is different from dep_sta_j. The arc is set as a same-day arc if the following condition is satisfied. Otherwise, the arc is set as a next-day arc.

$$arr\_time_i + 2 * interval\_time + emp\_time_{ij} \leq dep\_time_j \quad (3)$$

Type 4: Empty Run and Inspection

An arc is drawn when arr_sta_i and dep_sta_j are different, and also if either arr_sta_i or dep_sta_j is a station capable of performing an inspection (we assume that this station is adjacent to the rolling stock depot).

1. Case where an inspection is performed at arr_sta_i the same day
   If the following condition is satisfied, the arc is set as a same-day arc.

$$\max\{arr\_time_i, ins\_min\_time\} + ins\_time \leq \min\{dep\_time_j - emp\_time_{ij}, ins\_max\_time\} \quad (4)$$

   Judge if an inspection is possible within the same day.

2. Case where an inspection is performed at arr_sta_i the next day
   If the following condition is satisfied, the arc is set as a next-day arc.

$$\max\{arr\_time_i, ins\_min\_time\} + ins\_time > \min\{dep\_time_j - emp\_time_{ij}, ins\_max\_time\} \wedge ins\_time\_min + ins\_time \leq dep\_time_j - emp\_time_{ij} \quad (5)$$

   Here an inspection can be started at ins_time_min the next day.

3. Case where an inspection is performed at dep_sta_j the same day
   If the following condition is satisfied, the arc is set as a same-day arc.

$$\max\{arr\_time_i + emp\_time_{ij}, ins\_min\_time\} + ins\_time \leq \min\{dep\_time_j, ins\_max\_time\} \quad (6)$$

   Judge if an inspection is possible within the same day.

4. Case where an inspection is performed at dep_sta_j the next day
   If the following condition is satisfied, the arc is set as a next-day arc.

$$\max\{arr\_time_i + emp\_time_{ij}, ins\_min\_time\} + ins\_time > \min\{dep\_time_j, ins\_max\_time\} \wedge ins\_time\_min + ins\_time \leq dep\_time_j \quad (7)$$

   Here empty run is inserted on the same day the train arrives at arr_sta_i and then an inspection can be started at ins_time_min the next day.

3.2 Modelling as a Travelling Salesman Problem

Since several arcs in type 1 to 4 are set between nodes, the network generally includes multiple arcs between nodes. The model of Giacco et al. tries to find an optimal Hamiltonian path on the network in such a way that maintenance requirements are satisfied. Note that the problem becomes a generalized TSP on the (directed) network with multiple arcs. With regard to performance measures, the roster length could be measured by assigning a weight of 1 to the arc when the transition between nodes corresponds to date change; 0, otherwise. Similarly, the empty distance could be measured by assigning an empty distance to the arc when the transition between nodes includes an empty run.

Figure 4 shows a network for the services of five trains and the nodes correspond to these trains. Four possible types of arcs exist between a pair of nodes, and only feasible arcs are drawn.

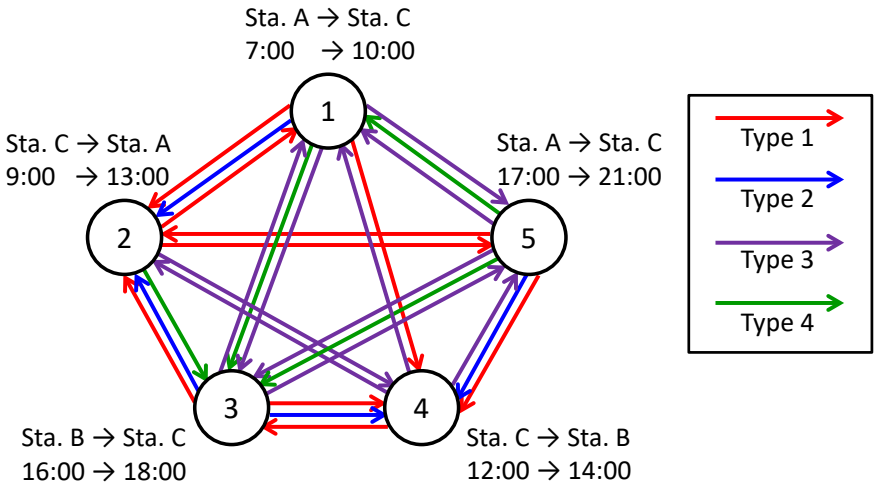


Figure 4: An example of the network model

3.3 Model Extensions for Splitting and Combining

Basic Idea

In our problem, there are two different types of trains, namely, single-unit trains and double-unit trains. This is modeled within the framework of the Giacco model by "dualizing" nodes for the double-unit trains. That is, two distinct nodes are prepared for those double-unit trains.

Whether splitting and/or combining operations are needed is partially judged by examining two consecutive nodes on the selected tour. If two adjacent nodes on the tour are both single-unit trains, then there is no shunting operation between them, but if the adjacent nodes call for a different number of train-sets, at least one shunting operation will definitely be needed between them. However, when the transition goes from one double-unit train to another double-unit train, it is possible that splitting and combing operations would be needed despite the fact that the same composition of two train-sets of the previous train may remain the same for the next train without shunting operations.

Train Position

In order to consider splitting and combining, we now define what we call "train position", which is sometimes simply called "position". The train position of a single-unit train is defined to be 0, with the train position of a double-unit train being 1 and 2. Here, the train position of a double-unit train represents an absolute geographic position. For example, the east side represents 1, and the west side 2, if the line extends in the east-west direction. Then, the train position of east-bound double-unit train 1M will be 1 for the front half of the train-sets, and 2 for the rear half. On the other hand, the train position of west-bound double-unit train 2M will be 2 for the front half of the train-sets, and 1 for the rear half. Assigning train positions allows us to judge the existence of splitting and combining during connections of double-unit trains. In the above example, if train 1M connects to

train 2M with the same train positions, no splitting and combining would occur as the connection keeps the same train composition. However, if train positions change after the connection from 1M to 2M in such a way that the positions are reversed (i.e., 1→2, 2→1), splitting is required first, then reversing the order of two train-sets, and finally combining the two train-sets together, thus requiring one splitting and one combining.

In our modification of the original Giacco model to consider splitting and combining, the network model is revised in such a way that each double-unit train is represented by two nodes, as mentioned above. In particular, for each double-unit train, we prepare one node with train position 1, and another node with train position 2. On the other hand, each single-unit train is represented by a single node with train position 0, just like the original Giacco model.

3.4 Counting the Amount of Splitting and Combining Based on Train Positions of Connection

Following the basic approach described above, we explicitly count in our model the amount of splitting and combining by judging the existence or non-existence of splitting and combining based on the information of train positions of trains before and after connections.

To do so, we separate the case of splitting from the case of combining. From a network viewpoint, this corresponds to separating considerations of the predecessor node of an arc from considerations of the successor node of the arc.

Splitting

(A) Connection from train position 0
This corresponds to a connection from a single-unit train, and therefore there is no splitting.

(B) Connection from train position 1
Here situations differ depending on the train positions after the connection.

(B1) Connection to train position 0
This case corresponds to a connection from a double-unit train to a single-unit train, so splitting occurs. We thus increment the number of splittings by one.

(B2) Connection to train position 1
This is the connection from position 1 to position 1, so there is no splitting, provided that the train-set of position 2 (before connection) connects to the same train as the train of position 1 (after connection). However, if the train-set of position 2 connects to a different train from the train of position 1 (after connection), splitting will be required. Therefore, the number of splittings will be either 0 or 1 depending on the status of position 2.

(B3) Connection to train position 2
The trains before and after connection are both double-unit trains in this case. As described in Section 3.3, whenever a train position changes, splitting is always required, so we increment the number of splittings by one.

(C) Connection from train position 2
For each train with position 2, there is always the same train with position 1. When splitting occurs, it will be counted in case B, above, for position 1. To avoid double counting, we assume that no splitting occurs for a connection from train position 2.

Combining

(D) Connection to train position 0
This corresponds to a connection to a single-unit train, so there is no combining.

(E) Connection to train position 1
Here situations differ depending on the train positions before the connection.

(E1) Connection from train position 0
This case corresponds to a connection from a single-unit train to a double-unit train, so combining occurs. We increment the number of combinings by one.

(E2) Connection from train position 1
This is the connection from position 1 to position 1, so there is no combining, provided that the train-set of train position 2 (after connection) connects from the same train as the train of position 1 (before connection). However, if the train-set of position 2 connects from a different train from the train of position 1 (before connection), combining will be required. Therefore, the number of combinings will be either 0 or 1 depending on the status of position 2.

(E3) Connection from train position 2
The trains before and after connection are both double-unit trains in this case. As described in Section 3.3, whenever a train position changes, combining is always required, so we increment the number of combinings by one.

(F) Connection to train position 2
For each train with position 2, there is always the same train with position 1. When combining occurs, it will be counted in case E above for position 1. To avoid double counting, we assume that no combining occurs for a connection from train position 2.

Application to Our Model

What is described above will be combined with the network model of Section 3.1 to derive the modified model, as will be described in detail below:

1. The case of B1, B3, E1, and E3
   Add cost to the corresponding arc.
2. The case of B2 and E2
   Existence or non-existence of splitting and combining depends on position 2, so the amount of splitting and combining is counted based on the position 2 after the connection. In order to count splitting and combining, we adopt several logical conditions. The details of this method will be described in the next section.

Following the basic approach described above, we explicitly count in our model the amount of splitting and combining by judging the existence or non-existence of splitting and combining based on the information of positions of trains before and after connections.

3.5 MILP Formulation

We now formulate the MILP problem. Notations for the MILP formulation are described in Table 2. Constraints can be classified into the four categories: assignment constraints, subtour elimination constraints, inspection constraints, forcing constraints for splitting and combining, and other constraints.

Table 2. Notation for MILP formulation

Notation	Definition
V	Set of nodes (Set of trains i, Index ranges over $0,1,\dots,\lvert V\rvert-1$)
V^1	Set of nodes with train position 1
A	Set of arcs
A^1	Set of arcs with no empty run and no inspection
A^2	Set of arcs with inspection and no empty run
A^3	Set of arcs with empty run and no inspection
A^4	Set of arcs with empty run and inspection
K	Set of arc types
c_{ij}^k	1 if the date changes when the arc from node i to node j of type k is selected, and 0 otherwise
d_{ij}^k	Distance of an empty run between nodes i,j of type k (will be positive when k is 3 or 4, and 0 when k is 1 or 2)
e_{ij}^k	Additional cost for splitting and combining between nodes i,j of type k
l	Lower limit of the inspection interval (in days)
m	Upper limit of the inspection interval (in days)
p_i	Node with train position 2 which shares the same train as node i with train position 1
x_{ij}^k	1 if the arc between nodes i,j of type k is selected, and 0 otherwise
y_{ij}	Order of the arc between nodes i,j on the selected tour
z_{ij}^k	Number of days since previous inspection of the arc between nodes i,j of type k on the selected tour
r_i	1 if connection from node i with train position 1 requires splitting, 0 otherwise
s_i	1 if connection to node i with train position 1 requires combining, 0 otherwise

Assignment Constraints

$$\sum_{k\in K}\sum_{j:(i,j,k)\in A} x_{ij}^k = 1,\ \ \forall i \in V \tag{8}$$

$$\sum_{k\in K}\sum_{i:(i,j,k)\in A} x_{ij}^k = 1,\ \ \forall j \in V \tag{9}$$

Equation (8) ensures that only one arc which emanates from node i is selected, and equation (9) ensures that only one arc which enters node j is selected.

Subtour Elimination Constraints

$$\sum_{j\in V} y_{ij} = \sum_{h\in V} y_{hi} + 1,\ \ \forall i \in V/\{0\} \tag{10}$$

$$y_{ij} \le |V| \sum_{k\in K} x_{ij}^k,\ \ \forall i \in V, \forall j \in V \tag{11}$$

$$\sum_{j \in V} y_{0j} = 1 \tag{12}$$

Subtours of the TSP would be eliminated by equations (10) to (12).

Inspection Constraints

$$\sum_{k \in K} \sum_{j:(i,j,k) \in A} z_{ij}^k = \sum_{k \in K} \sum_{j:(i,j,k) \in A^1 \cup A^3} \left(c_{ij}^k x_{ij}^k + z_{ij}^k\right) + \sum_{k \in K} \sum_{j:(i,j,k) \in A^2 \cup A^4} c_{ij}^k x_{ij}^k, \forall i \in V \tag{13}$$

$$l x_{ij}^k \leq z_{ij}^k, \qquad \forall (i,j,k) \in A^2 \cup A^4 \tag{14}$$

$$m x_{ij}^k \geq z_{ij}^k, \qquad \forall (i,j,k) \in A \tag{15}$$

The number of days since a previous inspection on a tour is calculated by equation (13). The first term corresponds to the case without inspection, and the second term the case with inspection. Lower and upper limits of the inspection interval are ensured by equations (14) and (15), respectively. With equations (13) to (15), inspections will be performed within the specified inspection intervals.

Forcing Constraints for Splitting and Combining

We now create logical conditions which say that a splitting would occur if for each double-unit train, position 1 is connected, and also if position 2 is not connected to the identical train, splitting will occur.

$$\sum_{k \in K} x_{ij}^k = 1 \ \wedge \ \sum_{k \in K} x_{p_i p_j}^k = 0 \longrightarrow r_i = 1, \qquad \forall i \in V^1, \forall j \in V^1 \tag{16}$$

The above logical conditions will be transformed into the following linear constraints:

$$\sum_{k \in K} x_{ij}^k - \sum_{k \in K} x_{p_i p_j}^k \leq r_i, \qquad \forall i \in V^1, \forall j \in V^1 \tag{17}$$

The case of combining would be similar to that of splitting. Logical conditions which say that combining would occur if for each double-unit train, position 1 is connected, and also if position 2 is not connected to the identical train, combining will occur.

$$\sum_{k \in K} x_{ji}^k = 1 \ \wedge \ \sum_{k \in K} x_{p_j p_i}^k = 0 \longrightarrow s_i = 1, \qquad \forall i \in V^1, \forall j \in V^1 \tag{18}$$

The above logical conditions will be transformed into the following linear constraints:

$$\sum_{k \in K} x_{ji}^k - \sum_{k \in K} x_{p_j p_i}^k \leq s_i, \qquad \forall i \in V^1, \forall j \in V^1 \tag{19}$$

Other Constraints

Other constrains are as follows.

$$x_{ij}^k \in \{0,1\}, \qquad \forall (i,j,k) \in A \tag{20}$$

$$y_{ij} \geq 0, \text{integer}, \qquad \forall i \in V, \forall j \in V \tag{21}$$

$$z_{ij}^k \geq 0, \qquad \forall (i,j,k) \in A \tag{22}$$

Objective Function

Under constraints (8) to (15), (17), and (19) to (22), the following objective function is minimized:

$$\alpha \sum_{(i,j,k)\in A} c_{ij}^{k} x_{ij}^{k} + \beta \sum_{(i,j,k)\in A} d_{ij}^{k} x_{ij}^{k} + \gamma \left\{ \sum_{(i,j,k)\in A} e_{ij}^{k} x_{ij}^{k} + \sum_{i\in V^{1}} r_i + s_i \right\} \tag{23}$$

The first term in equation (23) indicates the roster length (the number of days required to complete the roster, which is equivalent to the number of train-sets required to perform the roster), and the second term shows the total distance of empty runs. The third and fourth terms mean the amount of splitting and combining. The third one indicates the sum of additional costs for splitting and combining in the case of B1, B3, E1, and E3 described in Section 3.4. The fourth one is the sum of splitting and combining, which is calculated by the logical conditions (16) and (18) in the cases of B2 and E2 described in Section 3.4. Here, α, β, and γ are weight parameters.

4 Case Study

4.1 Lines and Settings

The proposed methodology is evaluated based on real instances of several lines in Japan with roughly 100 to 200 train services. Table 3 shows the details of each railway line, namely, the number of trains (sum of the single-unit trains and the double-unit trains), the number of the double-unit trains, the numbers of nodes and arcs in the network, together with the line length. The four lines are designated as A, B, C, D and for Lines B and C, problem instances with the reduced numbers of nodes by fixing some more or less obvious connections are added, namely Instances B-133 and B-147 for the original Instance B-161, and Instance C-177 for the original Instance C-256. In total, seven instances are tested. The numbers of trains range roughly from 100 to 200 which are typical in Japanese railways, and splitting and combining are performed in all these lines. Even though Line D involves fewer trains than Line C, the number of nodes of Instance D-258 is slightly more than those of Instance C-256 because Line D has more double-unit trains.

In our experiments, the proposed approach was evaluated based on the roster length, the total distance of empty runs, the amount of splitting and combining, and the computational time. We also analysed how solutions change when parameter weights are adjusted. The proposed algorithm was tested on a PC with Windows 10 Professional (64 bit), Core i7-8700K, and 64 GB RAM. In addition, Gurobi Optimizer 8.1.0 was used to solve the MILP model.

Table 3: Details of actual railway lines

Instance	No. of trains	No. of double-unit trains	No. of nodes	No. of arcs	Line length (km)
A-121	89	32	121	20,173	148.6
B-133	114	19	133	24,743	112.8
B-147	128	19	147	30,025	112.8
B-161	142	19	161	35,021	112.8
C-187	153	34	187	49,622	206.1
C-256	217	39	256	88,370	206.1
D-258	178	80	258	83,434	197.5

4.2 Results of Computational Experiments

Table 4 summarizes results of our computational experiments. Weight parameters were set at $\alpha = 1$, $\beta = 0.001$, and $\gamma = 0.01$ in these experiments. Computations were terminated after the maximum CPU time of 10,800 seconds, and the best results obtained at termination were shown if the maximum CPU time was reached before optimality.

Table 4 indicates that the optimal solutions were obtained except for Instance D-258. The total distances of the empty runs also appear to be reasonably small compared to line lengths. The number of splittings and combinings were either 8 or 10, which are small if we consider the numbers of the double-unit trains. Generally, good practical schedules were judged to be obtained for all instances. Instance D-258, however, could not be solved to optimality after the time limit of 10,800 seconds, even though the solution quality seemed to be good enough. Considering the fact that Instance D-258 required substantially more CPU time than Instance C-256 whose number of nodes and arcs are comparable to those of Instance D-258, it appears that the more double-unit trains we have, the more difficult is solving the instance, provided that the number of nodes is approximately the same.

Tables 5 and 6 show the effects of changing weight parameters β and γ (by fixing $\alpha = 1$) for bigger Instances C-256 and D-258, respectively. Since the roster length is constant regardless of whether the parameter is β or γ, the roster length is considered to be minimized in each instance. Under the weight parameter $\gamma = 0$, the amount of splitting or combining jumps up unreasonably high, which implies that in order to obtain practically reasonable results, counting and reducing the amount of splitting and combining must be included in the model mechanisms. We found that raising the value of γ even just a little reduced the amount of splitting and combining, and making β relatively larger is more effective. It should also be noted that optimal solutions are obtained quickly when $\gamma = 0$, but as γ is increased above zero, optimality could not be reached within the set CPU time limit in many instances. This could be attributed to the fact that logical conditions were introduced into the model to count the amount of splitting and combining, which makes the problem difficult to solve. Considering the fact that practically good upper bounds are obtained, CPU times could be reduced by improving the LP lower bounds with the generation of effective cuts.

The results in Tables 5 and 6 indicate that reduction of splitting and combining would be essential to obtain a reasonable roster. Otherwise, the resultant rosters would be unrealistic and may not be practically acceptable due to too many splittings and combinings. Computational feasibility is confirmed for the range of instances tested even

though CPU time increases when we consider reduction of splitting and combining operations.

Table 4. Results of computational experiments

Instance	Roster length	Empty run (km)	No. of splittings and combinings	CPU time (sec.)
A-121	17	39.5	8	67
B-133	13	4.8	8	49
B-147	13	4.8	8	102
B-161	13	4.8	8	94
C-187	22	75.4	10	1,845
C-256	22	75.4	10	3,254
D-258	29	131.6	8	*10,800

* indicates termination before optimality due to maximum time limit.

Table 5. Influence of change of the weight parameters (Instance C-256)

β	γ	Roster length	Empty run (km)	No. of splittings and combinings	CPU time (sec.)
0	0	22	3257.2	76	819
0	0.01	22	1506.8	8	*10,800
0.001	0	22	75.4	68	804
0.001	0.00001	22	75.4	10	*10,800
0.001	0.0001	22	75.4	10	*10,800
0.001	0.01	22	75.4	10	3,254
0.001	0.1	22	252.4	6	*10,800

* indicates termination before optimality due to maximum time limit.

Table 6. Influence of change of weight parameter (Instance D-258)

β	γ	Roster length	Empty run (km)	No. of splittings and combinings	CPU time (sec.)
0	0	29	5335.6	140	1,082
0	0.01	29	5368.8	2	*10,800
0.001	0	29	131.6	134	778
0.001	0.00001	29	131.6	10	10,675
0.001	0.0001	29	131.6	8	*10,800
0.001	0.01	29	131.6	8	*10,800
0.001	0.1	29	192.8	4	*10,800

* indicates termination before optimality due to maximum time limit.

5 Conclusions

In this paper, we focused on railway rolling stock rostering problems with maintenance considerations. Splitting and combining are used to adjust transportation capacity flexibly in Japanese railways. On the other hand, it is desirable that the amount of splitting and combining be minimized because these shunting operations require time and manpower. This paper proposes an MILP model based on a TSP with multiple arcs between nodes so that the amount of splitting and combining is reduced. Numerical experiments based on actual lines in Japan show that the proposed model incorporating the mechanisms to count

and reduce the number of shunting operations can generate practically good rosters with a reduced number of splittings and combinings. Computational feasibility is confirmed for the range of instances tested even though CPU time increases when we consider reduction of splitting and combining.

Possible future work includes:

- reduction of CPU time
- extensions of the model to the cases where three or more train-sets are assigned to some trains

Reduction of CPU time could be achieved by improving LP lower bounds with the generation of effective cuts. Considerations of trains to which three or more train-sets are assigned, may be included into the model by expanding the idea of train positions, but logical conditions would be expected to become very complicated which may increase CPU burden, so alternative approaches may also need to be considered.

References

Abbink, E., Van den Berg, B., Kroon, L., Salomon, M., 2004. "Allocation of railway rolling stock for passenger trains", *Transportation Science*, vol. 38, pp. 33-41.

Alfieri, A., Groot, R., Kroon, L., Schrijver, A., 2006. "Efficient circulation of railway rolling stock", *Transportation Science*, vol. 40, no. 3, pp. 378-391.

Borndörfer, R., Reuther, M., Schlechte, T., Waas, K., Weider, S., 2016. "Integrated optimization of rolling stock rotations for intercity railways", *Transportation Science*, vol. 50, pp. 863-877.

Giacco, G, L., D'Ariano, A., Pacciarelli, D., 2014. "Rolling stock rostering optimization under maintenance constraints", *Journal of Intelligent Transportation Systems: Technology, Planning, and Operations*, vol. 18, pp. 95-105.

Fioole, P.-J., Kroon, L., Maroti, G., Schrijver, A., 2006. "A rolling stock circulation model for combining and splitting of passenger trains", *European Journal of Operational Research*, vol. 174, pp. 1281-1297.

Morooka, Y., Fukumura, N., Takayuki, S., Imaizumi, J., Morito, S., 2017. "Rolling stock optimization based on the model of Giacco et al.: computational evaluation and model extensions", In: *Proceedings of 7th International Conference on Railway Operations Modelling and Analysis (RailLille2017)*, Lille, France.

Nishi, T., Ohno, A., Inuiguchi, M., Takahashi, S., Ueda, K., 2017. "A combined column generation and heuristics for railway short-term rolling stock planning with regular inspection constraints", *Computers and Operations Research*, vol. 81, pp. 14–25.

Peeters, M., Kroon, L., 2008. "Circulation of railway rolling stock: a branch-and-price approach", *Computers and Operations Research*, vol. 35, pp. 538-556.

Reuther, M., Schlechte, T., 2018. "Optimization of rolling stock rotations", In: Borndörfer, R., et al. (eds), *Handbook of Optimization in the Railway Industry*, International Series in Operations Research & Management Science 268, Springer, Switzerland.

Tsunoda, M., Imaizumi, J., Morito, S., 2015. "A model for estimating the required number of train units under split-and-merge policy for decision making in railways –a mathematical formulation by integer multi-commodity network flow-", In: *Proceedings of 6th International Conference on Railway Operations Modelling and Analysis (RailTokyo2015)*, Narashino, Japan.

A Traveller Perspective on Railway Punctuality: Passenger Loads and Punctuality for Regional Trains in Sweden

Ida Kristoffersson [a,1], Roger Pyddoke[a]
[a] VTI Swedish National Road and Transport Research Institute
Box 55685, SE-102 15 Stockholm, Sweden
[1] E-mail: ida.kristoffersson@vti.se, Phone: +46 (0) 8 518 388 11

Abstract
This paper examines the extent to which delayed trains are also trains with more passengers. The paper uses unique passenger load data about regional trains in Sweden and combines this with Swedish delay statistics for the same train numbers and from the same time periods. Results show that trains with high passenger numbers are not delayed to a greater extent compared to trains with fewer passengers. Train punctuality is thus a good indicator of traveller punctuality in this case. These results also suggest that long boarding and alighting times due to high passenger numbers are not a main cause of delays, possible causes of delays are instead external factors such as track maintenance or dense train movements. Therefore, this result suggests that policy makers should look further into the latter causes. Furthermore, the paper also compares the share of travellers and trains that are more than half an hour late, i.e. that are significantly late. These differences are also small but larger than for the less delayed trains. For one of the railway lines, trains with high passenger loads are more than proportionally hit by long delays. Such cases suggest that train control priorities could be re-examined with more focus on improving the service for railway travellers.

Keywords
Train punctuality, Passenger load, Regional train, Train traveller demand, Delays

1 Introduction

Punctuality is a key issue for the railway in order to be an attractive mode choice for travellers. After safety, punctuality is the most important performance indicator for most railway infrastructure managers including the Swedish Transport Administration (Swedish Transport Administration, 2018). Furthermore, passengers rate punctuality as the key success factor for railway travel, and requests for trains to be more punctual are more common than requests for trains to be more frequent (Transport Focus, 2015).

A substantial amount of research has been conducted to understand which factors influence railway punctuality. These studies use different modelling techniques: analytical models (Bergström and Krüger, 2013), multiple regression models (Olsson and Haugland, 2004; Palmqvist et al., 2017) and machine learning (Marković et al., 2015). There are also a number of studies in the literature focusing on construction of time tables that are more robust to disturbances (Andersson et al., 2015; Cerreto et al., 2016; Liebchen et al., 2010; Solinen et al., 2017).

However, all the above-mentioned studies focus on *train* punctuality rather than on *traveller* punctuality. Traveller punctuality is an important area, as there is reason to believe that passenger loads can vary considerably both between trains on the same railway segment, and for the same train on different railway segments. Such research has, however been difficult to conduct in this area in Sweden (and in many other countries) due to the reluctance of railway operators to give access to passenger load data. The issue is of the magnitude that new methods to calculate number of train travellers from mobile phone data is under development (Sørensen et al., 2018). It is only recently that it has become possible to get access to historic passenger load data for certain railway lines in Sweden.

Average punctuality of all trains in Sweden has been around 90% in the latest years (Trafikanalys, 2018), i.e. 90% of trains arrive less than six minutes after schedule to the destination. This could be seen as a fairly high figure with a vast majority of the trains being punctual, but traveller expectations on punctuality are very high (Transport Focus, 2015). There is no reliable evidence that delays are larger in Sweden than in other European countries, but comparisons with other European countries suggest that Sweden has lower levels of punctuality than comparable countries (BCG, 2017). It is also possible that travellers experience more delays, since trains might be delayed to a greater extent at times when many people travel. This could be due to disturbances spreading, and thus more trains are affected by disturbances, at times when there is a high capacity utilization in the railway network, or because of delays at stations due to long boarding and alighting times when passenger numbers are high. It is also possible that travellers' perception of punctuality has more to do with a psychological phenomenon that people are more likely to remember outstanding events (Phelps and Sharot, 2008), i.e. train journeys with long delays compared to journeys when everything goes according to plan. A third possibility is that travellers concern about punctuality is driven by a very strong negative valuation of a small risk of facing a very long delay (Börjesson and Eliasson, 2011). Whether trains are more delayed when many people travel is still an open question, since there is a lack of studies combining passenger load data with punctuality data. One study of commuters in the greater Stockholm area in Sweden investigates delays during peak hours and find that, to be 95% certain to arrive on time, a buffer time corresponding to up to 37% of the travel time needs to be added (Föreningen TIM-pendlare, 2015).

There is also no clear consensus on which measures are the most cost-efficient measures for reducing delays in Sweden. With punctual trains defined as trains arriving less than six minutes after schedule to the destination, there has been a focus on cutting short delays.

Long delays are however extra disturbing to passengers, since it may imply that passengers miss connections or do not arrive in time for a meeting. Nelldal (2016) analyse the causes of long delays but includes no information about how many passengers were affected by these long delays.

In this paper, rare Swedish passenger load data (Jansson et al., 2017) is combined with train punctuality data in order to extend previous work on train punctuality with analyses of two aspects of traveller punctuality:

1) The share of *punctual* travellers (arriving 5 minutes and 59s after schedule or earlier to the destination) compared to the share of *punctual* trains.
2) The share of *significantly late* travellers (arriving 30 minutes after schedule or later to the destination) compared to the share of *significantly late* trains.

2 Context

2.1 The Study Area

The study area in this paper is the so called Mälarbanan between Stockholm and Hallsberg[1], see the green line in Figure 1. This is an electrified track about 200 km long which mainly constitutes of double track, e.g. the Stockholm-Västerås part of Mälarbanan constitutes solely of double track. Train traffic on Mälarbanan is mainly regional trains running from Hallsberg to/from Stockholm and from Västerås to/from Stockholm. On the part of Mälarbanan located in Stockholm County, the regional trains also share tracks with commuter trains.

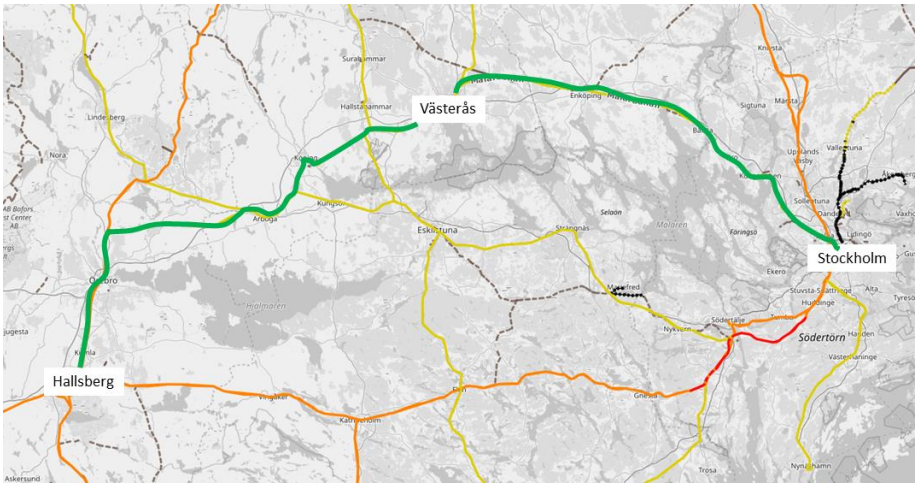


Figure 1: The studied track Mälarbanan north of the lake Mälaren in Sweden.

[1] To be exact, Mälarbanan does not go all the way to Hallsberg, rather it merges with another railway track in Hovsta about 35 km north of Hallsberg.

2.2 Efforts to Increase Punctuality of Swedish Trains

In Sweden, the infrastructure manager and train operators work together to increase punctuality in a cooperation called TTT (Together for Trains on Time). Data on disturbances, reasons for disturbances and time deviations at stations are collected and analysed within TTT. This data shows that regional trains in Sweden are in general more punctual than long-distance trains, but not as punctual as commuter trains (all measured with 5min and 59s allowed time deviation at the destination), see Table 1. Note that the difference between high-speed long-distance trains and other long-distance trains is quite large.

Table 1: Comparison of punctuality across different train types for the year 2015.

2015	**Highspeed long-distance**	**Other long-distance**	**Regional**	**Commuter train**
Total number of trains	41145	29482	376054	406056
Number of trains <6 min late	30303	24902	337996	383270
Share of punctual trains	73.6%	84.5%	89.9%	94.4%

Punctuality is higher on Saturdays and Sundays. For regional trains during 2015, the average share of punctual trains was 92.7% on weekends, compared to 89.2% on weekdays.

A train leaving a station three minutes or more after schedule is registered as a disturbed train and a reason for the disturbance needs to be given. Total number of disturbance minutes for each train is registered in the TTT database, called LUPP. The disturbance minutes varies depending on time of day. Figure 2 shows the variation of disturbance minutes depending on time of day for different types of trains for the year 2015. For all train types it is clear that disturbances are not evenly spread over the day, rather two clear peaks can be seen, one in the morning and one in the late afternoon. For regional trains on Mälarbanan (studied in this paper) there is a more distinct morning peak compared to other train types, suggesting a high share of commute travelling. Disturbance minutes per hour is highest around eight in the morning and around six in the afternoon. This observation combined with the observation from Jansson et al. (2017) that trains have the largest passenger loads in these time periods suggests the hypothesis that passengers could be more than proportionally hit by delays compared to trains.

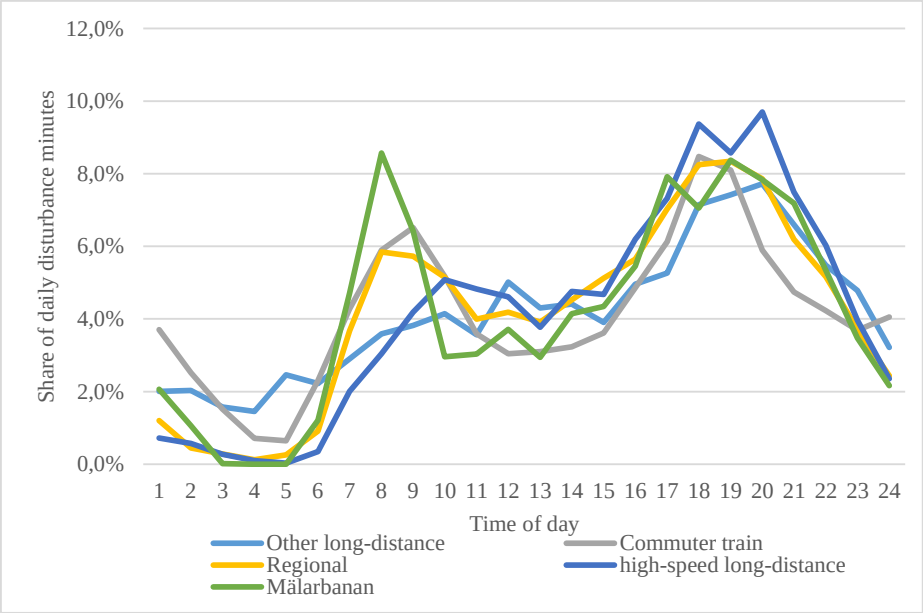


Figure 2: Share of daily disturbance minutes depending on time of day for regional trains compared to other train types for the year 2015. Data from the LUPP-database.

In the report from TTT for the year 2017 (JBS, 2018), the relative importance of different causes for delays are analysed. The results suggest that management of the railways and train operations are two large categories of causes for delays. So far not much is reported about the distribution of delays in time and geography.

3 Method and Data

3.1 Definitions

Two definitions are used in this paper:

1. A train or traveller is considered **punctual** if arriving to the destination of the train 5 minutes and 59s after schedule or earlier.

2. A train or traveller is considered **significantly late** if arriving to the destination of the train 30 minutes after schedule or later.

Both punctual and significantly late trains and travellers are thus binary measures that adopt the values true or false.

3.2 Passenger Load Data

In this paper, two main data sources are combined to analyse traveller punctuality. On the one hand, passenger load data from weekdays in September and October 2015 for regional trains on Mälarbanan is used (Jansson et al., 2017). This data includes occupancy and number of seats per train number and railway segment (between larger stations). Thus, from this data number of passengers on the train can be calculated. An example of the passenger load data is shown in Table 2.

Table 2: Example of passenger load data from Jansson et al. (2017). Occupancy and number of seats for trains from Västerås to Stockholm C weekdays in September 2015.

Train number	#Seats	Occupancy (%)			
		Västerås-Enköping	Enköping - Bålsta	Bålsta - Sundbyberg	Sundbyberg - Stockholm
707	239	32	57	61	59
783	293	47	69	73	69
781	395	41	50	50	50
787	387	61	72	74	70
791	243	27	30	29	28
795	300	39	44	42	38
797	277	39	41	37	33
799	243	35	35	32	29
767	244	11	11	10	10

Jansson et al. (2017) show that average passenger loads are highest in the morning and afternoon peak (but still below 80 percent of seating capacity at the highest) and that very low occupancy numbers often are the case for trains running in the late evening, see Figure 3 and Figure 4. For the line Västerås – Stockholm C there is a clear morning peak for occupancy, see Figure 3. In the other direction, Stockholm C – Västerås, the highest occupancy occurs in the afternoon peak, see Figure 4. The pattern in the two figures indicate a large share of commute travel to Stockholm.

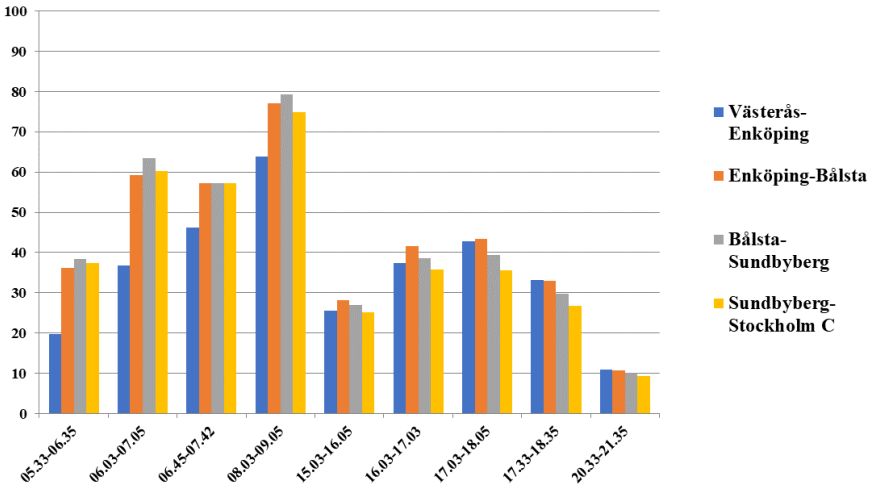


Figure 3: Occupancy (in per cent) for regional trains on weekdays from Västerås to Stockholm C, average values for September 2015, October 2015 and April 2016 (Jansson et al., 2017).

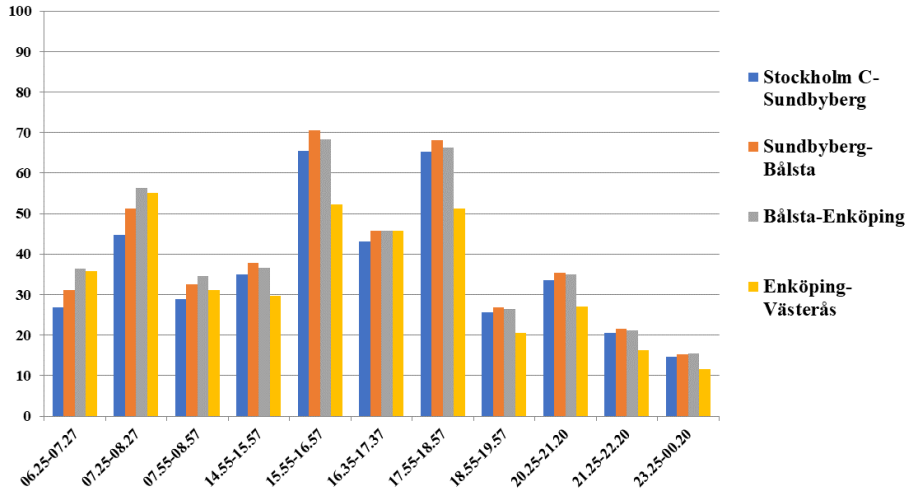


Figure 4: Occupancy for regional trains on weekdays from Stockholm C – Västerås, average values for September 2015, October 2015 and April 2016 (Jansson et al., 2017).

3.3 Punctuality Data

For the same time period (weekdays in September and October 2015) and the same train numbers, train delay data have been extracted from the Swedish delay database LUPP, in which deviations from the time table are recorded for all stations along the line.

Table 3 shows an example of train delay data from LUPP. Note that in LUPP time deviations are also given for stations in-between the major stations, but these have not been extracted here. Note also that time deviation from schedule can be negative. Small negative time deviations are quite common in the data, but few observations are more than 7 minutes early. Trains and travellers arriving early to the destination are in this paper considered to be punctual. The train in the example of Table 3 is not punctual, since it arrives to the destination Stockholm C 34 minutes late - it is even *significantly late* according to the definition in 3.1, since it is more than 30 minutes late.

Table 3: Example of train punctuality data for a train from Västerås to Stockholm C extracted from the LUPP database.

Date	Place	Train number	Arrival/ Departure	Planned time	Time deviation (min)
2015-09-01	Västerås	781	Departure	06:45	0
2015-09-01	Enköping	781	Arrival	06:58	1
2015-09-01	Enköping	781	Departure	06:59	1
2015-09-01	Bålsta	781	Departure	07:09	0
2015-09-01	Sundbyberg	781	Departure	07:31	34
2015-09-01	Stockholm C	781	Arrival	07:38	34

There are in total 1343 observations of train journeys in the format of Table 3 for which time deviation at the destination is given. Each of these observations are matched to passenger load data for the corresponding train number and measurement month. Most observations are from September 2015, with 436 observations from October 2015. After combining punctuality and passenger load data, the data contains around 220 000 traveller journeys, which implies an average of 164 passengers per train.

3.4 Result Quantities

Table 4 lists the result quantities and shows how they are calculated. Note that result quantities are only calculated for the destination of the train using passenger loads from the last section of the journey and time deviation at the destination. Passenger load data and time deviation for earlier sections of the journey are not used in calculations and are only included here for completeness.

Table 4: Result quantities and how they are calculated.

Result Quantity	Calculation	Unit
Share of punctual trains	Number of punctual trains / total number of trains for the geographic relation under study	%
Share of punctual travellers	Number of punctual travellers / total number of travellers for the geographic relation under study	%
Share of significantly late trains	Number of significantly late trains / total number of trains for the geographic relation under study	%
Share of significantly late travellers	Number of significantly late travellers / total number of travellers for the geographic relation under study	%

4 Results

The results of the analysis from a traveller perspective is presented in this chapter. First, we note that train punctuality for the regional trains on Mälarbanan studied in this paper is on average 90.5% when looking at the two different railway lines in both directions. This figure is somewhat higher, but similar to the punctuality of 89.9% for all regional trains in Sweden during 2015 presented in Table 1. Thus, it appears that the studied observations are representative of regional trains in Sweden, or at least do not deviate too much when it comes to punctuality.

4.1 Traveller Punctuality Compared to Train Punctuality

Figure 5 shows that, for regional lines running on Mälarbanan, traveller punctuality is somewhat lower than train punctuality. Average traveller punctuality is 88.8% compared to 90.5% for train punctuality. This suggests that trains are not much more delayed at times when many passengers travel. As indicated in Table 1, trains travelling longer are more often delayed. The same pattern can be found in the results of this paper, where the share of punctual trains is lower for Hallsberg-Stockholm compared to Västerås-Stockholm (which is a shorter distance as can be seen in Figure 1).

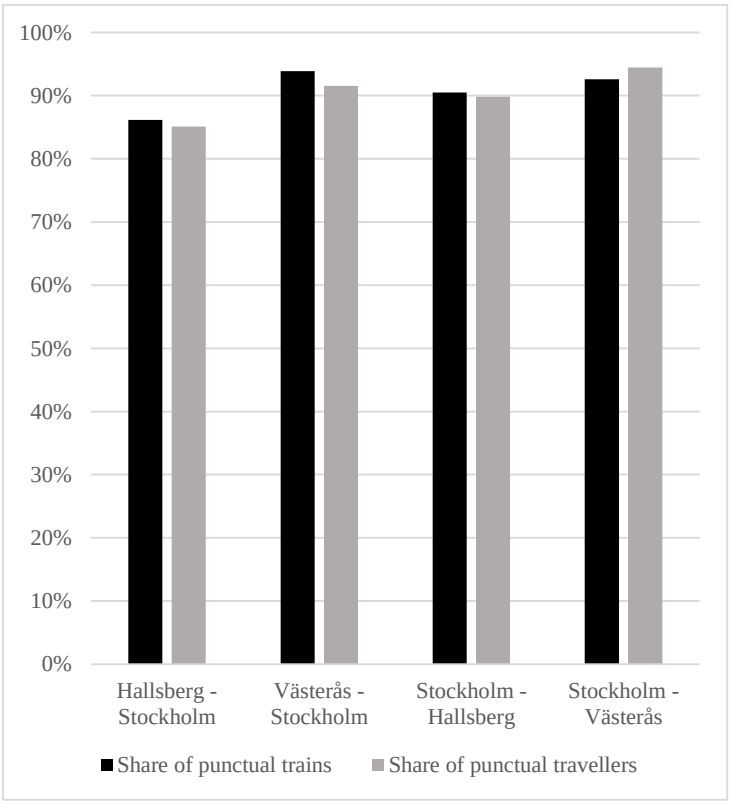


Figure 5: Share of punctual travellers compared to the share of punctual trains for regional train lines on Mälarbanan.

4.2 Significantly Late Travellers Compared to Significantly Late Trains

Figure 6 shows a comparison of significantly late travellers and trains. Here, the figures differ somewhat more between travellers and trains compared to the measurements of punctuality above. The share of significantly late travellers is substantially larger than the share of significantly late trains for the relation Västerås-Stockholm (2.9% compared to 1.5%), but on the other hand, the opposite is true for the relation Stockholm-Västerås, which has a higher share of significantly late trains than travellers (2.0% compared to 1.2%). Looking at an overall average this evens out the differences between travellers and trains and the average share of significantly late is 1.8% for travellers compared to 1.7% for trains.

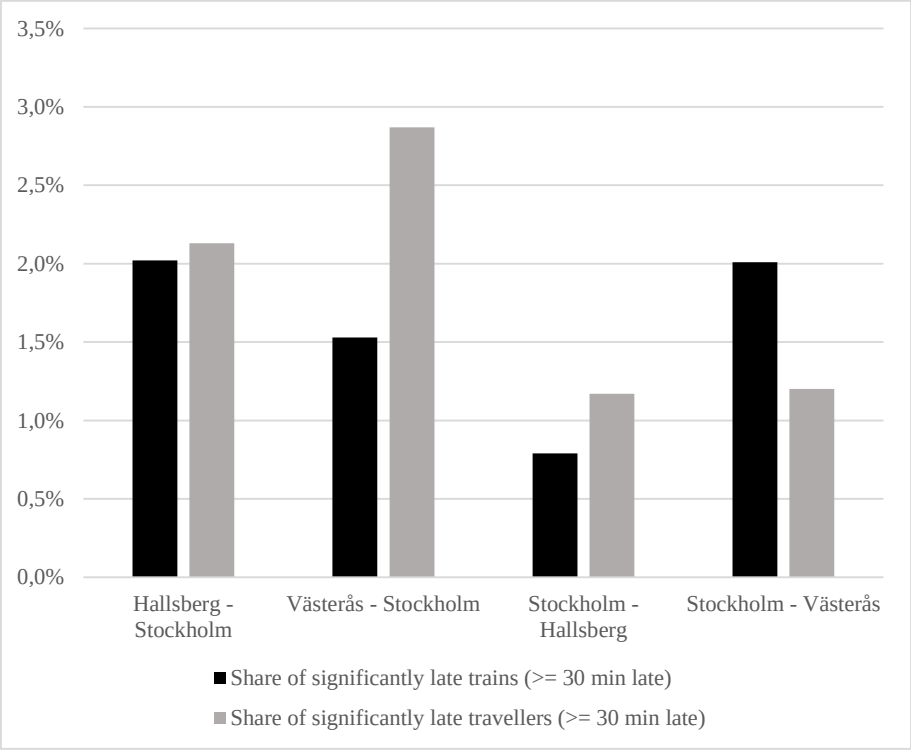


Figure 6: Share of significantly late travellers compared to share of significantly late trains for regional lines on Mälarbanan.

A deeper analysis of the data reveals that, for the relation Västerås to Stockholm, the significantly late trains are trains running in the morning peak, i.e. when occupancy rates are the highest and most people travel (see also Figure 3). Therefore, a larger share of travellers is affected. This is in line with the hypothesis made in the introduction. However, when analysing the opposite direction – Stockholm to Västerås – it is not primarily the trains running in the afternoon peak that are significantly late. On the contrary, the significantly late trains mainly depart from Stockholm C in the early morning peak. Figure 4 shows that occupancy rates for the early morning peak are rather low in this direction of travel (around 30%). Therefore, the share of significantly late trains is higher than the share of significantly late travellers in this case. One possible interpretation of these results could be that congestion/delays at the large station Stockholm C in the morning peak affects trains in both directions even though occupancy rates are much higher in the inbound direction.

There are more useful interpretations to be made from the analysis of significantly late travellers. First, it is notable that the share of significantly late travellers is not negligible. Out of 220 000 travellers in the data, around 4 000 (1.8%) are significantly late when arriving at the train's destination. Second, Figure 6 shows that the variation between relations is larger for significantly late travellers compared to traveller punctuality.

5 Discussion

If delays are mainly driven by boarding and alighting times policy implications will be different than if it is external factors or dense train movements that drive delays. Punctuality is however rarely associated with passenger numbers and consequently the impact it has on passengers has not been measured.

This study indicates that for regional trains in the Hallsberg-Västerås-Stockholm corridor in Sweden, trains do not seem to be much more delayed at times when many passengers travel. Train punctuality is thus a good measure of traveller punctuality in this case (not accounting for travellers that travel further with connecting trains). This result suggests that high passenger numbers and subsequent longer boarding and alighting times are not the main cause of delays for the context under study. Rather indications from the JBS (2018) study are that external factors, such as infrastructure or signal system failures; or knock-on delays due to dense train movements (with not as densely loaded trains in this case) could be possible causes of delays.

If it is external factors and/or dense train movements that is the main cause of delays, then reducing number of trains in the network (maybe combined with lengthening of trains and platforms) could be beneficial. Improved timetable planning and improved processes for operational traffic (decision support, improved information about rolling stock, improved interaction between infrastructure manager and railway undertaking etc.) are also effective measures to increase punctuality. Long railway lines will have increased probability of being disturbed if external factors occur evenly spread along the railway network. Data shows that punctuality is lower for long-distance trains, especially high-speed long-distance trains. In a railway network, such as the Swedish network, with many single tracks and a mix of trains with different speeds, it is quite common that high-speed trains get stuck behind a slower train. Changing operation rules to prioritize the faster train could be beneficial in this situation.

Furthermore, results of this paper suggest that it is important to monitor the share of significantly late travellers. The share of significantly late travellers and trains differ more compared to traveller and train punctuality. It is therefore important to bear in mind that even if only a small share of trains is significantly late, the share of travellers might be substantially larger. It is also important to note that the share of significantly late travellers is not negligible, at least not for regional trains in Sweden. A delay of more than half an hour imposes great trouble for travellers and reduces traveller trust in the railway, which will have a negative impact on the railway market share. This suggests that significantly late travellers/trains should be monitored and more in focus of future measures to decrease railway delays.

This paper examines the hypothesis that trains in peak carry more passengers and are more hit by delays and that therefore passengers on average experience more delays than trains. We find only small differences between the shares of trains and travellers that are delayed. We find some indications of uneven distribution of delays, both with respect to train types and time of day. Our assessment is that there is a need for further analysis of both the causes for delays and the time and geographical patterns of delays to examine the evenness of these distributions. Are the likelihoods of being hit by delay even per train kilometre and track kilometre?

Acknowledgements

This work was founded by the Swedish Transport Administration under Grant number TRV2018/102432. We would like to thank Mats Gummesson at the Swedish Transport Administration for providing us with data on time deviations and disturbance minutes from the LUPP database. We would also like to thank Magnus Wahlborg at the Swedish Transport Administration for comments on an earlier version of this paper.

References

Andersson, E.V., Peterson, A., Krasemann, J.T., 2015. Reduced railway traffic delays using a MILP approach to increase Robustness in Critical Points. J. Rail Transp. Plan. Manag. 5, 110–127.

BCG, 2017. The 2017 European Railway Performance Index.

Bergström, A., Krüger, N.A., 2013. Modeling passenger train delay distributions: evidence and implications. CTS Work. Pap. 20133.

Börjesson, M., Eliasson, J., 2011. On the use of "average delay" as a measure of train reliability. Transp. Res. Part Policy Pract. 45, 171–184.

Cerreto, F., Nielsen, O.A., Harrod, S., Nielsen, B.F., 2016. Causal Analysis of Railway Running Delays, in: 11th World Congress on Railway Research (WCRR 2016), Milan, Italy.

Föreningen TIM-pendlare, 2015. Verkliga förseningar för tågpendlare Jämförelse med officiell statistik.

Jansson, K., Pyddoke, R., Paulin, C., 2017. Variationer i beläggning i tid och rum för tre tåglinjer norr om Mälaren (No. K2 Research 2017:6).

JBS, 2018. Tillsammans för tåg i tid - Resultatrapport 2018 (No. 2018:109).

Liebchen, C., Schachtebeck, M., Schöbel, A., Stiller, S., Prigge, A., 2010. Computing delay resistant railway timetables. Comput. Oper. Res. 37, 857–868.

Marković, N., Milinković, S., Tikhonov, K.S., Schonfeld, P., 2015. Analyzing passenger train arrival delays with support vector regression. Transp. Res. Part C Emerg. Technol. 56, 251–262.

Nelldal, B.-L., 2016. Stora trafikavbrott och förseningar vid Sveriges järnvägar och dess effekter (No. TRITA-TSC-RR 16-003). KTH Royal Institute of Technology, Stockholm.

Olsson, N.O., Haugland, H., 2004. Influencing factors on train punctuality—results from some Norwegian studies. Transp. Policy 11, 387–397.

Palmqvist, C.-W., Olsson, N., Hiselius, L., 2017. Some influencing factors for passenger train punctuality in Sweden.

Phelps, E.A., Sharot, T., 2008. How (and why) emotion enhances the subjective sense of recollection. Curr. Dir. Psychol. Sci. 17, 147–152.

Solinen, E., Nicholson, G., Peterson, A., 2017. A microscopic evaluation of railway timetable robustness and critical points. J. Rail Transp. Plan. Manag. 7, 207–223.

Sørensen, A.Ø., Bjelland, J., Bull-Berg, H., Landmark, A.D., Akhtar, M.M., Olsson, N.O., 2018. Use of mobile phone data for analysis of number of train travellers. J. Rail Transp. Plan. Manag.

Swedish Transport Administration, 2018. Punktlighet i tågtrafiken (Punctuality in railway transport).

Trafikanalys, 2018. Punktlighet på järnväg 2017 (Railway punctuality 2017).

Transport Focus, 2015. Train punctuality: the passenger perspective.

Centralizing and Migrating Operational Infrastructure Databases

Alexander Kuckelberg [a,1], Bianca Multykin [a,2]
[a] VIA Consulting & Development GmbH
Römerstr. 50, 52064 Aachen, Germany
[1] E-mail: a.kuckelberg@via-con.de, Phone: +49 (0) 241 463662 16
[2] E-mail: b.multykin@via-con.de, Phone: +49 (0) 241 463662 39

Abstract

Since years and decades, IT systems are used to plan, to monitor and to control train operations and railway traffic on network regions. Especially in technically advanced railway networks, the usage of computer based systems for dispatching and controlling traffic started quite early, e.g. in the 90s. This implies the necessity to update and renew outdated structures nowadays.

As IT system performance, database sizes and functionalities grew within the last decades, a wide range of existing system limitations are not valid anymore and can be overcome by new systems, processes and hardware. Larger data sets and therefore the migration and aggregation of valid data sets within one new, larger data set are possible now.

However, for operational systems it is highly advisable to follow an evolving strategy for the migration of distributed structures instead of a revolutionary approach to ensure the operability of working systems and ongoing operations. Such a strategy requires the migration of existing data sets and processes whereby the question arises how to migrate e.g. formerly overlapping infrastructure areas, how to aggregate semantically identical data sets with distinct technical keys etc.

This paper introduces these challenges from various points of view and presents approaches chosen by the authors to establish such a migration process of existing and running operational infrastructure databases. It focuses on technical aspects to migrate and aggregate infrastructure data but also outlines challenges with respect to migration of workflows and processes towards centralized services.

Keywords

Infrastructure Topology, Databases, Data Consolidation, Legacy Systems

1 Introduction

Current operational IT systems used for train control and dispatching often realize a microscopic infrastructure model as a base data model. The migration and aggregation of legacy systems into new and larger systems arises several challenges, e.g. consolidation of different, probably overlapping infrastructure data sets, migration problems with respect to unsynchronized data maintenance and guarantee of consistency for resulting consolidated data.

Within this chapter, two elementary aspects of the problem are introduced: The microscopic infrastructure data model used by legacy systems and the problem of

overlapping responsibilities and unsynchronized microscopic data sets.

After introducing the basic data structures and clarifying the problems implied by distributed IT systems, the resulting challenges are described within Chapter 2 for some selected aspects in detail. Chapter 3 outlines approaches which have been selected (and implemented) to solve the challenges and finally Chapter 4 concludes.

1.1 Microscopic Infrastructure Model

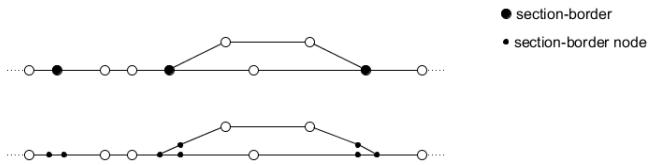


Figure 1: Mesoscopic (above) and microscopic (below) view on railway topology.

In railway operations research approaches as well as in operational systems the modelling of railway infrastructure is an essential step towards a functional data model that can be processed automatically by IT systems. The granularity of such models determines the capability of systems set upon these models. For operational systems like train control and dispatching, microscopic infrastructure models turned out to fit functional requirements in a good manner. The systems considered by this work implement microscopic infrastructure models as follows:

- The real infrastructure is modelled by a graph model consisting of nodes and edges, where nodes represent infrastructure elements (signals, stopping positions, switches and crossings, axle counters etc.) and edges represent tracks connecting the infrastructure elements.
- Track sections – tracks without branches – are represented by a sequence of inner nodes (with exactly two neighbours), bordered by two outer nodes.
- Outer nodes and therefore section-border nodes are track end nodes, buffer nodes, branches of switches and crossings, transition nodes towards new logical/operational node affiliation, etc. (Figure 1). Neighboured sections are connected by an edge.
- All nodes are logically clustered into operational control points (OCP), and sections end when entering a new operational control point. Consequently, these sections end with an OCP bordering node and are linked to the corresponding OCP bordering node starting a section of the other OCP.
- All nodes have a mileage value, ordering section nodes within a section in a monotonous manner. Consequently, all nodes of one section are ordered and the section itself gets an implied direction.
- Switches are modelled by three, crossings by four section-border nodes (one for each branch); interconnections between these nodes represent the possible routing throughout a switch or crossing, resp. (Figure 2).

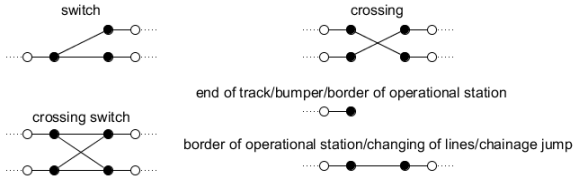


Figure 2: Modelling of interconnections between sections.

Upon these microscopic nodes technically secured routes are defined (Figure 3). These routes consist of tracks of a station or a line. In theory a route is a path in the graph consisting, inter alia, of a start and an end node, and the course of the route throughout the graph (defined by branching information for each switch passed by). With routes, the interlocking behaviour and dependency by signals is modelled.

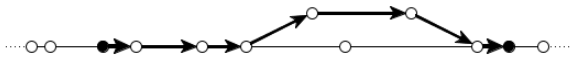


Figure 3: A route defined as a path in the graph, starting with an entry signal.

1.2 Distributed Structures and Data Sets

In this work, we consider the railway infrastructure of a large area, e.g. a whole country. We assume that the infrastructure data is distributed into regions, where each region is controlled by its responsible operational control centre. Every control centre has the infrastructure data of its own region and a small glimpse of the infrastructure data across its border to model the cross-border coherences. So the given infrastructure data of every control centre is its region internal infrastructure data extended by its own data of the infrastructure across its border. It is not ensured that the overlapping infrastructure data of two operational centres is synchronized, as every control centre only takes care of its own database. As the data sets are maintained independently the possibility of historically grown apart data in border regions is given. The infrastructure data of each control centre is considered to be consistent.

Figure 4: Regional data responsibility and data sets with overlapping region borders.

In addition to regional oriented infrastructure data, every control centre has its own content oriented data like train types, public holidays, braking tables etc. This data should be similar for every control centre, but e.g. deviation in naming might exist and some

control centres may have more extensive data sets than others or just data that is not relevant for others as a public holiday that is only regional.

2 Challenges and Migration Problems

Data sets of operational control centres – as representatives of regions – follow certain semantics: All centres have similar data, acting as master data of more or less static conditions that form a certain common data basis. This data set is considered as content oriented data and has to be treated in another way than region oriented data, where the maintenance responsibility can be clearly assigned to one control centre (except for cross-border coherences). The characteristics of these two principle data sets are described in the following sections. The handling when migrating depends on these data characters.

2.1 Content Oriented Data Sets

The content oriented data of an operational control centre contains all information considered as "static dictionary data", often called master or framework data. It includes for instance the list of train types, braking tables (braking percentage/LZB braking curve/braking delay/ETCS), referenced keys and data of tractive units. Content oriented data which exists in various regions should be equal in every regions data set to ensure a common perspective, e.g. a unique locomotive number should identify the same engine in all regions.

So an aggregating data set would secure the consistency of the perspectives of the regions. Even though there is data which may not be relevant for every region, a universal database for all regions would bring advantage as it is of importance for the optimization of cross-border process flows. A public holiday which does not exist in one region but in its adjacent region is important, as it may influence its workload or its timetable as example.

The main challenges for content oriented data sets is to determine how "new data sets" fit to the existing data sets, to detect content changes for the same data entries and to select the correct shaping of a data entry.

2.2 Regional Data Sets

Regional data sets considered within this paper consist primarily of infrastructure data following the detailed microscopic infrastructure data model (Chapter 1.1). One characteristic of regional data sets is that they can be assigned to a region uniquely, so exactly one operational control centre is responsible for its regions allocated subgraph.

Regional data sets might overlap at their bordering areas. This means, that for operational reasons a regional data set might contain data of topology and graph areas which are not in its operational control centres responsibility and vice versa.

While infrastructure data within a region is expected to be consistent and consistency is ensured by the legacy systems itself, migrated databases have to deal with bordering areas, where information from affected regions might contradict each other.

Additionally, not only the topology and infrastructure of two regions might be contradicting within the bordering area, but also the elements and information contained within sections of the bordering area, e.g. distances between nodes (e.g. distance signal to main signal), braking distances, maximum speeds, gradients, train control equipment etc.

2.3 Synchronization and Collaboration, Process Flows

One of the central problems when migrating similar and complementary data sets in the context of databases is the identity of data entries. Region identifiers of data entries might not be unique within the more global context of centralized and migrated databases anymore.

In other words, it should be possible to load different content versions of one data entry and to manage all its occurrences.

Another important aspect when migrating legacy databases are existing workflows. As mentioned, the migration should follow an evolving approach which directly implies, that legacy workflows remain similar and only change stepwise.

Therefore, migrating regional data sets also includes migrating workflows, e.g. the frequency of data version publication and propagation for each region and how new versions are integrated into the migrated database. Two principle approaches are possible:

- Direct reaction whenever new data versions are published by a region or
- Implementation of aggregation, enforcement of new workflows e.g. collection and propagation periods.

3 Centralizing and Migrating

In our approach every data delivery of a region is stored as one version. Data is transferred into an object-identity set and a shaping set, called splitted schema. The object-identity set contains the technical global keys of the new enlarged data set and columns forming semantical keys. Semantical keys are derived from regional data sets and remain equal in every regions delivery. So these keys identify entries throughout all regions, e.g. the combination of a tractive unit series number and a company identifier for a tractive unit (as only one of these attributes would not identify it uniquely). The shaping data set contains the remaining data content of every entry and references its object identity as outlined by Figure 5.

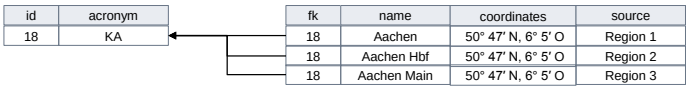


Figure 5: Separation of data entry identity from entry content (example OCP).

The object-identity is generated once for each semantical key, the shaping set grows with every delivery and is associated with the corresponding version.

Content oriented data from different regions is merged on behalf of common object identity and multiple shapings (Chapter 3.1). Region oriented data is restricted to entities related to the merged region (Chapter 3.2) therefore a methodology to ensure consistency at borders is developed (Chapter 3.3).

For both data types object dictionaries are introduced to map semantical keys of incoming entities to objects representing the target entities with their associated attributes. These object dictionaries represent the current state of the merged target data set as they get extended with every occurrence of a new semantical key while merging. Moreover, an entity-wise merging in a hierarchical order, derived from data entity dependencies, is implemented to ensure hierarchical data entity references.

3.1 Merge of Content Oriented Data Sets

For the merge of content oriented data sets the already merged content oriented data is loaded from the target data set and added to the dictionaries identified by their semantical keys. In this way, the information about already existing and known content oriented data is provided for further merging.

On the other side, delivered data has to be merged into the target data sets. First the imported data is converted into the splitted target scheme. Every splitted data entity gets saved into its object identity and shaping data set. While merging a data set, it is iterated pairwise through these sets for each delivered entity. For every considered pair the existence of its semantical key in the object dictionary is verified and depending on its occurrence it is acted (Figure 6).

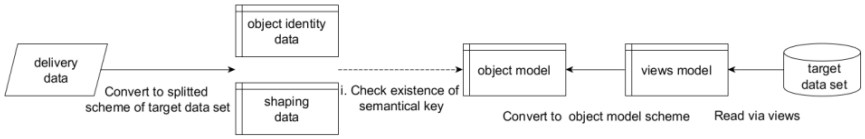


Figure 6: Merging components for content oriented data sets.

During the merge process the object dictionaries represent the current state of the merged target data set. By merging and saving into the target data set the local technical keys of the delivery data set are replaced recursively by the technical keys of the target data set on behalf of the dictionaries. The hierarchical merging enables correct re-referencing to the keys of earlier merged data entities (references which are part of the semantical key included). So with every merge-step first the re-referencing takes place.

In the next step, the existence of the entities identity – the semantical key – within the corresponding dictionary is checked.

If the identity is missing, the object identity is added to the target data set and the object identity of the entities splitted data gets updated by the new generated technical key. Additionally the object repository gets extended by the persisted entity.

If the identity already exists the entities object identity key is replaced by the mapped one of the target data set.

In both cases, the shaping data entity component is added to the target data set and associated to the updated object identity as well as to the version defining the data entities validity (Figure 7).

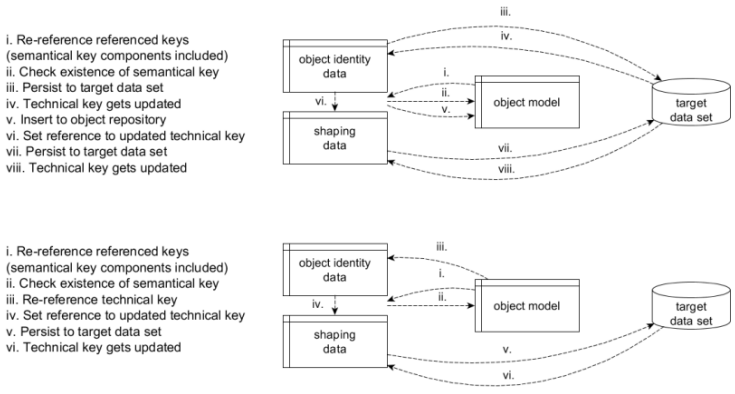


Figure 7: Merge of content oriented base data sets.
Above: Workflow for a new data entity.
Below: Workflow for an already existing data entity.

3.2 Merge of Region Oriented Data Sets

Region oriented data sets are merged in a different way than content oriented data sets: The regions infrastructure data of adjacent regions is excluded for every regions delivered data set. In consequence problems might occur while merging, if infrastructure data which belongs to the merged region references infrastructure data of another region. In this case, references are modelled by semantical keys which get saved as a substitute for later integration.

Again, like in the merge of the content oriented data, the delivered data entities get transferred into the splitted schema of the target data set due to the merge (object identity and shaping data). For region oriented data in turn the object identity data only contains the mapped technical keys of the target data set and the technical keys of its delivery data set.

As the amount of infrastructure data usually is very huge, key-maps get introduced which map the technical keys of a regions region oriented delivery data sets from earlier merges to the technical keys of the target data set. These resulting key-maps are loaded from the target data set as a first, preparing step.

Non-existing technical keys of the regions delivery data set in the key-map imply persistence of the object identity data to the target data set and hereby include the generation of new technical keys. By every new generation, the key-map gets extended. If the technical key already exists within the key-map, mapping can be performed directly.

In this way for all region oriented data, merging can be limited to key-map activities. Complete re-referencing is supported and the existing (shaping) infrastructure data of the target data set is not required. It can remain within the target data set without any access.

The content oriented data set is merged before the regional oriented data set. Due to this hierarchical treatment in the current state of the merge the object repository contains all content oriented data within the corresponding dictionaries for further usage. So re-referencing of referenced content oriented data can further on take place through the object repository.

With every merge-step the object repositories get extended by the re-referenced region oriented data entity as well, whereby in this case only for test purposes.

Independent from the existence of the data entities technical key in the key-map, the entities shaping data component is added to the target data set and associated to the object identity as well as a version defining the data entities validity (Figure 8).

Additionally, operation control points and lines are expected to be valid for all regions and are treated as overall data with given and predefined identities. Delivered data sets will only extend these data entities in the target data set with corresponding warnings.

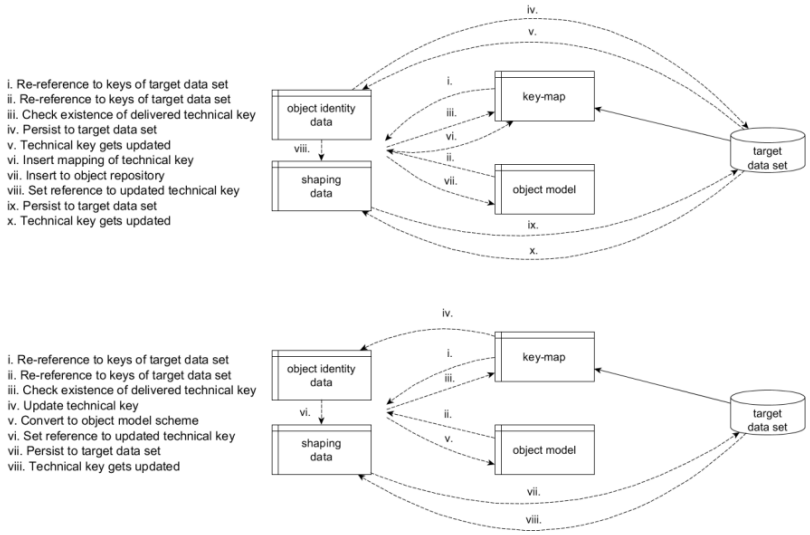


Figure 8: Merge of region oriented data sets.
Above: Workflow for a new data entity.
Below: Workflow for an already existing data entity.

3.3 Border Analysis of Region Oriented Data Sets

As described in Chapter 3.2., infrastructure data of all neighbour regions is filtered. When merging regional data, the bordering areas might become inconsistent, e.g. missing tracks within the neighbouring region. Border consistency is verified by a new, supporting data entity, the connectors. Connectors are elements which conclude information about the border conditions of the regions potential border crossings, identifying a position within the graph semantically. Two shaping data sets (border-node and border-route connectors) concretise infrastructure and route specific information at connector positions.

Bordering Infrastructure

Border-node connectors are introduced to represent infrastructure specific border information. A border-node connector references an associated border-node of the region and contains indicator values that should fit to the indicator values of an associated border-node connector belonging to the adjacent neighbour region. Indicator values are e.g. distances between the border-node and closest signals (distant and main) along the track, current braking distances or track characteristics like gradient and curve. Some values moreover are computed for inbound or outbound trains separately (e.g. closest

signals). Connectors are identified while merging the regional data sets due to semantical information and represent locations within the graph (border-nodes or track ends), where regions might join and where border consistency has to be checked.

After merging all region oriented data, a newly developed graph iteration algorithm determines the indicator values of all identified border-node connectors and assigns the values to them. The iteration determines distances considering mileage changes or mileage direction changes. Break conditions for the graph iteration are e.g. the reach of graph borders, exceeding of a defined maximum distance or the successful determination of all values. Figure 9 outlines this algorithm.

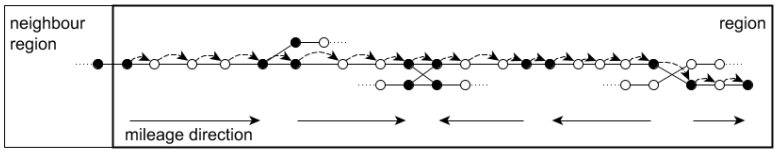


Figure 9: Graph iteration algorithm for indicator value determination of a border-node connector.

Bordering Routes

Cross-border routes have to be considered in the border analysis as well. Routes have to be split up at borders if they are belonging to different regions. With a region perspective the inner route segment might be an inbound or outbound segment starting or ending at a border-node referenced by a border-node connector. Border-route connectors are generated for these cross-border routes and semantical keys are derived, so matching route segments can be associated to each other at data retrieval time.

As for border-node connectors border-route connectors are enriched by route indicator values. These indicators contain e.g. the partial routing information. Figure 10 illustrates the algorithm which iterates throughout the route by its course and identifies the inner route course of the region. In dependence from the routes direction, the algorithm starts at the end- or the start node of the route, as regional data across the border was filtered, until it reaches the border-node of the associated connector.

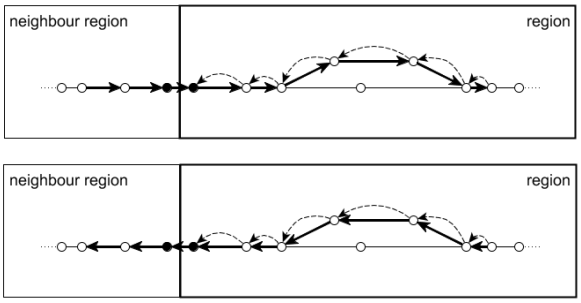


Figure 10: Graph iteration algorithm for indicator value determination of border-route connector.

Data Consistency

Data consistency checks for data of neighbouring regions are performed on behalf of border-node connectors and border-route connectors by their existence and their indicator values.

A first topology check determines the existence of fitting border-node connectors for adjacent regions (border-node to border-node or track-end without partner). If the topology of bordering nodes is consistent the next steps will evaluate indicator values with respect to infrastructure and routes.

Border-node connectors are evaluated with respect to reasonable indicator value matches, e.g. distance of cross-bordering distant and main signals, consistent train protection, tunnel cross-section, gradient and curve value consistency and more (Figure 11). Route-connectors are e.g. evaluated with respect to complementary partial routing.

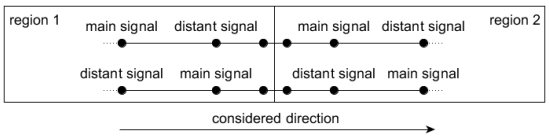


Figure 11: Consistency check example (signal positioning & border-node connectors).

4 Conclusion

This paper presented an approach to migrate and consolidate existing legacy infrastructure databases. With the approaches it becomes possible to centralize existing infrastructure data sets currently used in several operational IT systems with microscopic data models.

The approach therefore allows bringing together distributed and currently independent data sources and setting up new, centralized functionalities on top of new and enlarged data models.

Moreover, the paper outlines less technical aspects when migrating and implementing such IT systems. Anyway, final experiences and evaluations will not be possible until the new IT systems are in operation, the algorithms will be adopted and modified accordingly while implementing and evolving the system.

References

Kuckelberg, A., 2015. *Graph Databases and Railway Operations Research Requirements*, EDBT/ICDT Workshops, 183-188.

Kuckelberg, A., Seybold, B., 2013. *Adaptive Rule-Based Infrastructure Modelling*, Proc. of the 5th International Seminar on Railway Operations Modelling and Analysis, Copenhagen.

Hundt, R., 2005. *Betriebszentralen der DB Netz AG – Die Revolution in der Betriebsführung*.Signal + Draht. Vol 97, Nr. 1, ISSN 0037-4997, S. 3.

Hirschlach, J., 2016, *Aufbruch ins Computerzeitalter* Eisenbahn-Magazin. Nr. 3, ISSN 0342-1902, S. 45.

Bormet, J., Rausch, R., 2017, *Weichen stellen für die Zukunft der Betriebssteuerung*. Deine Bahn. Bahn Fachverlag GmbH

Evaluation of Train Operation with Prediction Control by Simulation

Taketoshi Kunimatsu [a,1], Takahiko Terasawa [b], Yoko Takeuchi [a]
[a] Signalling and Transport Information Technology Division, Railway Technical Research Institute
2-8-38, Hikari-cho, Kokubunji city, Tokyo 185-8540, Japan
[1] E-mail: kunimatsu.taketoshi.49@rtri.or.jp, Phone: +81 (0) 42 573 7311
[b] Railway Technical Research Institute (Former)

Abstract

In recent years, to increase transportation capacity, new intelligent signalling systems such as moving block have been proposed and put in operation. In addition, research and development on prediction control are now ongoing. Prediction control is a kind of optimization of train operation curves to minimize train headway, which leads to decrease the propagation of train delay. Hence, it is important to estimate the effects of new signalling systems or train control systems because replacement of current system may incur high costs. In this study, we first proposed and formulated new methods for applying prediction control for under both fixed and moving block. Then, we developed new functions on Train Operation/Passenger Behaviour Simulator to analyse the activity of trains with prediction control, taking into account the drivers' operational requirements. Finally, we applied the simulation system to an actual commuter line, aiming to evaluate the quantification of effects of moving block and prediction control. As a result, we confirmed that both moving block and prediction control are effective to decrease train headway, which lead to the faster recovery from delay.

Keywords

moving block, prediction control, train traffic control, simulation, passenger flow

1 Introduction

In railways, signalling systems are conventionally developed based on the concept "fixed block," under which only one train is allowed to enter a block section. Recently, new signalling systems, named "moving block," are now going to put into operation. Under moving block, as train headway become shorter, more trains can be set during the peak hours, and train delay is easily recovered (Baba et al. (2003)).

In addition, a new train control method named "Prediction Control," is proposed in the previous research (Hiraguri et al. (2004)). That is, based on the prediction for the departure time of the preceding train from the station, the succeeding train is controlled to arrive at the station with minimum headway. Prediction control can be applied whether the signalling system is fixed block or moving block.

When these new systems are considered to be installed, the existing system have to be replaced with train control system using radio communication It incurs high costs, and requires detailed design about location of radio base stations, and allocation of radio

frequency slots. So, it is desired to analyse cost effectiveness of the new systems in advance of installations.

In this research, we focus on developing simulation system which can estimate train traffic or passenger flow under new systems. The goal of our research is to realize the method to quantitatively evaluate effects of installing new signalling systems or train control systems, such as moving block or prediction control.

We first devised the fast estimation method for train operation curves under moving block. We then devised train control algorithm based on prediction control theory under both fixed and moving block to minimize headway between successive trains. After that, we implemented those methods to "Train Operation/Passenger Behaviour Simulator," which is developed by the authors to reproduce train traffic under a certain timetable (Takeuchi et al. (2015)). Finally, we evaluated effects of installing moving block and prediction control in an actual commuter line in Japan.

In addition to our previous work (Kunimatsu et al. (2018)), we devised a new method for estimating train operation curves, by which we can prevent unnecessary delay propagation for trains running after the succeeding train. The differences among train operation curves are discussed in the case study for an actual commuter line. We also discussed changes of effects when the departure delay of the preceding train is altered.

The rest of the paper is organized as follows. Section 2 describes our target problem and aim of the research. In section 3, we introduce the conventional simulator, Train Operation /Passenger Behaviour Simulator. Details of the devised train traffic simulation method with prediction control under fixed block is described in section 4, and that under moving block is in section 5. The application results of our method for an actual commuter line are described in section 6. We summarize and conclude our research in section 7.

2 Motivation and Aim

2.1 Moving Block

In railways, to avoid collision of trains and guarantee safety of train operation, signalling systems are developed based on the concept "block." Conventionally, fixed block signalling systems are used, under which only one train is allowed to enter a block section which is mainly set between two successive signals. The succeeding train is controlled to stop in front of the block section in which the preceding train is on, and the marginal stop point is moved forward discretely, according to the change of block section which the preceding train is on.

On the other hand, new signalling systems, named "moving block," are now going to be developed and put into operation (Fig.1). That is, the marginal stop point for the succeeding train is caught and updated repeatedly by the radio communication system, according to the

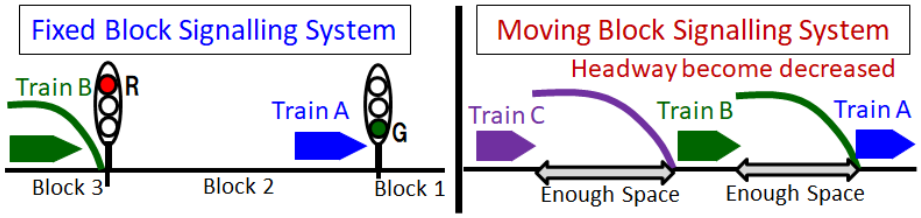


Fig. 1. Outline of fixed block and moving block

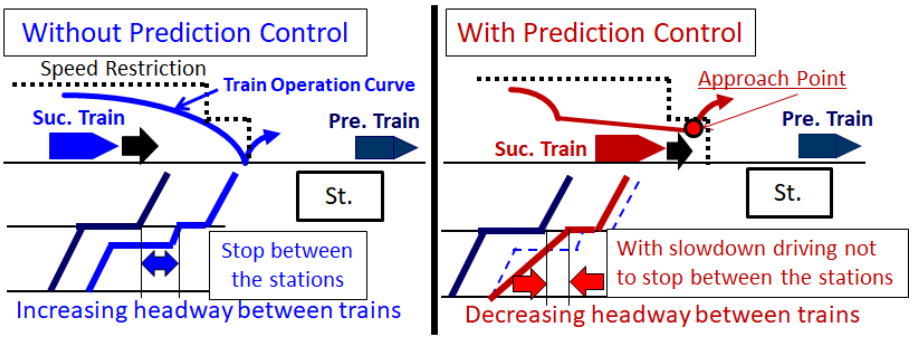


Fig. 2. Outline of prediction control

continuous change of the position of the preceding train. Under moving block, train headway become smaller. So, more trains can be set during the peak hours, and train delay is easily recovered.

2.2 Prediction Control

When a train is running close to its preceding train at the next station, if dwell time of the preceding train become longer, the successive train may stopped in front of the station platform. It may increase headway between the two trains, because it takes time for the successive train to restart again. To avoid this situation, an intelligent train control method of the succeeding train, called “Prediction Control”, is proposed (Hiraguri et al. (2004)). That is, if a train driver exactly know when the preceding train depart from the station, he can drive slowly to minimize train headway. However, it is difficult to evaluate effects of prediction control, because effects of that depends on train traffic condition. It is desired to evaluate quantitatively merits of prediction control, like decreasing headway and faster recovery from the delay.

In prediction control theory, to minimize headway between the two trains, there is a kind of target point for the succeeding train to pass. It is called “approach point.” The approach point consists of position, speed, driving operation and time of the succeeding train. The approach point is also on the train operation curve for the train to stop at the marginal stop point in front of the station platform. By controlling the succeeding train to pass the approach point, the headway is minimized if the preceding train departs from the station just on the predicted departure time. In case when the preceding train do not depart from the station, the succeeding train stops at the marginal stop point, avoiding bump into the rear of the preceding train.

When prediction control is applied for a rail line, it is necessary to install new intelligent train traffic control system, in which both train control and traffic control functions are implemented. It is necessary for trains to continuously communicate and exchange each other about detail information of their positions, velocity, and driving operations. This may be realized by train control systems using radio communications, which are going to be actually used. In addition, computers with calculation and information processing functions have to be installed on trains to create optimal train operation curves. Moreover, to control trains to run along with the optimal train operation curves, ATO (Automatic Train Operation) system or DAS (Driver Advisory System) is essential. Although there are several problems to be solved to realize these systems, in this research, we set the preconditions that prediction control can be realized. Under the preconditions

above, we developed methods for evaluating effects for train traffic and passenger flow by installing prediction control.

2.3 Purpose of Research

In this research, our goal is set to develop a simulation system which estimate both train traffic condition and passenger flow under moving block and prediction control. By using the simulator, we want to evaluate quantitatively total merits of prediction control, like decreasing headway and faster recovery from delays.

We first improved functions of existing "Train Operation/Passenger Behaviour Simulator" to reproduce train traffic under moving block. Then, we developed and implemented functions for prediction control under both fixed and moving block. After that, we applied the simulator for the existing rail lines, and evaluated effects of prediction control.

2.4 Related Works

There are some previous works about evaluation of moving block, or train control algorithm to decrease train headway. Kanda et al. analysed the extent of decrease of train delay when moving block is installed in commuter lines in Japan (Kanda et al. (2014)). D'Ariano et al. and Xu et al. proposed a method to optimize train headway or energy consumption by controlling train operation curves of group of trains (D'Ariano et al. (2005), Xu et al. (2015)). They optimize train traffic by using estimated positions or signal aspects of trains. But, they do not consider both train traffic and passenger flow in the target rail line.

In commuter lines in a big city like Tokyo, it is not sufficient for optimizing train headway or energy consumption to estimate train operation curves only by simulation. Trains may be delayed due to the excess of dwell time at stations caused by congestion. The delay may in turn affect the succeeding train, and optimized train operation curves cannot be realized. So, it is necessary for the simulation to incorporate estimation of passenger flow and dwell time at stations. The comprehensive simulation and optimization method of both train operation curves and passenger flow is not developed yet.

In our previous works, we developed "Train Operation/Passenger Behaviour Simulator," which can estimate both train operation curves and passenger flow (Takeuchi et al. (2015)). Then, by developing functions for optimizing train operation curves, and

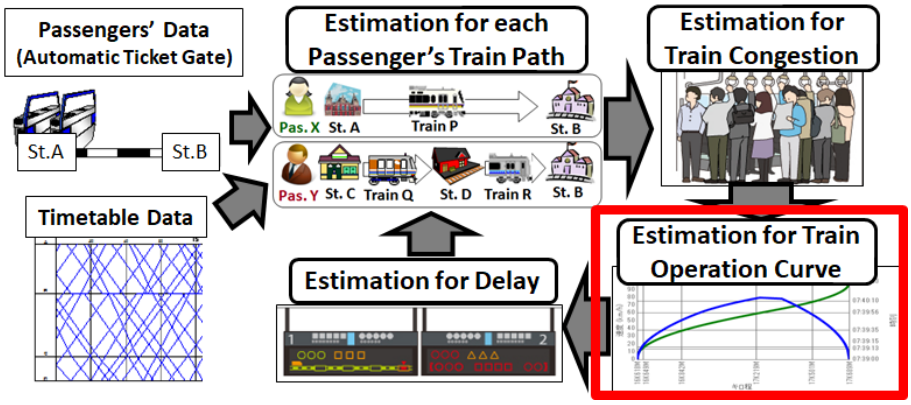


Fig. 3. Outline of train operation/passenger behaviour simulator

implementing that to the simulator, we realized train traffic and passenger flow simulation with prediction control (Kunimatsu et al. (2018)). In this research, we proposed a new method for estimating train operation curves to prevent unnecessary delay propagation. In addition, we discussed changes of effects when the departure delay of the preceding train is altered.

3 Train Operation/Passenger Behaviour Simulator

3.1 Fundamental Function of the Simulator

The overview of Train Operation/Passenger Behaviour Simulator is shown in Fig. 3. The inputs are, timetable data, passenger Origin-Destination data collected through the automatic ticket barriers, and signalling equipment data. The outputs are data for estimated train operation time, passenger train paths towards their destinations, and number of the passengers on board each train. The simulator also predicts train delays caused by congestion, and propagation of train delays. By estimating passengers' train paths, the number of passengers on board each train, and train delays successively, it is possible not only to evaluate timetables, but also various types of equipment, such as signalling systems. During morning rush hour on commuter lines in particular, the dwelling time of trains in stations is longer, because of the high number of trains being operated, and the extent of delay propagation depends on the design of the signalling system. The simulator can be used to design a signalling system to minimize train delay propagation.

Since Train Operation/Passenger Behaviour Simulator can estimate the route taken by passengers from the train operating timetables, it is possible to evaluate the timetable and signalling equipment design from the passenger point of view.

3.2 Estimation Function for Train Operation Curves under Fixed Block Systems

Train Operation/Passenger Behaviour Simulator can be applied to rail lines using fixed block systems. In the case of a fixed block system, the simulator first estimates the train

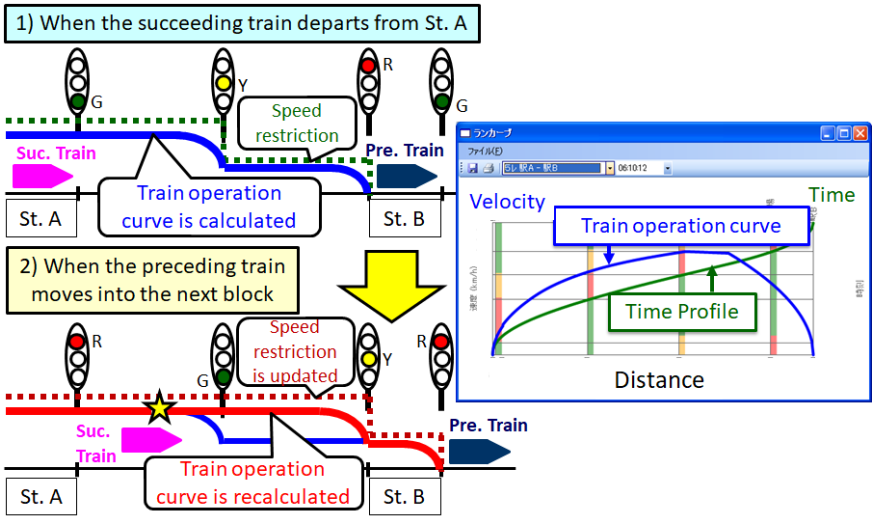


Fig. 4. Simulation of train operation curves under fixed block systems

operation curves when each train departs from a station, based on the signal aspect and all speed restrictions. Then, each train runs according to the estimated train operation curve. When the preceding train exits a block, and moves into the next block, the train operation curve is recalculated and updated (Fig. 4). The number of train operation curve recalculations is therefore equal to the number of blocks the train passes through. An approximate number of recalculations is given by the number of trains multiplied by the number of blocks. The overall simulation for an actual commuter line in Japan can be conducted within about 15 minutes with an ordinary personal computer.

The method for estimating train operation curves is the same as in SPEEDY (Yamashita (2006)), which was developed by RTRI and is used in practice for assessing train operating times. Train performance curves can therefore be rapidly estimated by predicting acceleration or deceleration of trains not only in the forward direction from the position of the train, but also in the opposite direction from where the train is stopped, which is determined by signal aspects. In addition, considering that it is difficult for trains to move from powering to braking, without coasting in between for a certain time, the estimation method of train operation curves can take into consideration these driving restrictions.

3.3 Efficient Recalculation Method for Train Operation Curves under Moving Block

In moving block systems, when the preceding train goes ahead, the marginal stop point for the succeeding train moves forward continuously. So, in a train traffic simulation under moving block, if the train operation curves are recalculated using the conventional method for fixed block systems, they would have to be recalculated for every simulation period. If the simulation period is one second, the approximate number of recalculations would be the product of the number of trains and the simulation time (sec.), which would far exceed the number of calculations for the fixed block system.

This research therefore adopts a new estimation method for train operation curves (Fig. 5). In this method, when the first train operation curve is estimated, the time when the train starts coasting to decelerate and stop at the marginal stop point is also predicted. After that, the train operation curve is not recalculated until the train starts coasting. Recalculation is not necessary during this time because the train operation curve will not be influenced by the continuous change in position of the preceding train. When the succeeding train is already located in a position closer to the preceding train than to the position where coasting starts, then the whole train operation curve may be affected by the preceding train, and so

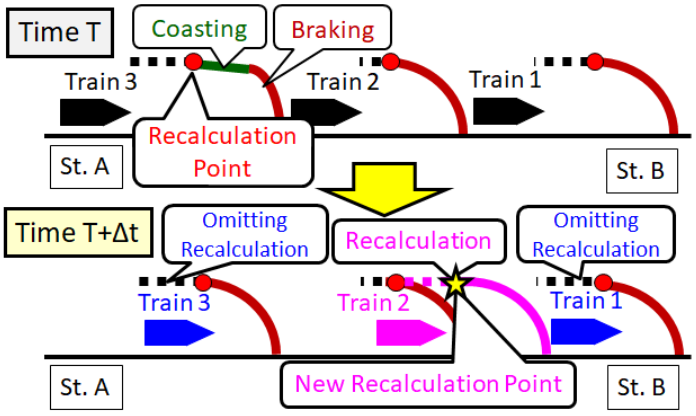


Fig. 5. Simulation of train operation curves for moving block systems

recalculation is conducted for each simulation period.

In the proposed method, the number of recalculations is lower than in the conventional method. The effect of the proposed method depends on the number of trains, or the headway of trains. If there are no succeeding trains that begin to coast to decelerate and stop at the marginal stop point, the approximate number of recalculations is equal to the number of trains multiplied by the number of times a succeeding train reaches the recalculation points. This number is much smaller than that in the conventional method.

4 Train Traffic Simulation with Prediction Control under Fixed Block

4.1 Preconditions

When we discuss prediction control under fixed block, the approach point needs to be set and calculated. It consists of position, velocity and time, on which the succeeding train is controlled to pass. In this research, considering possibility of changing signal aspects for the succeeding train during prediction control, we set the following preconditions to calculate the approach point properly.

1: The working acceleration rate for powering, coasting and braking are supposed to be constant, regardless of trains or conditions.
$Powering : \alpha\,[m/s^2]$, $Coasting : \beta[m/s^2]$, $Braking : \gamma[m/s^2]$
2: The necessary time to change driving operation from powering to coasting, or coasting to braking, are supposed to zero[sec].
3: The minimum continuous time for coasting is considered and supposed to constant time (t[sec]).
4: Prediction control is applied for a succeeding train only in cases which it and its' preceding train have stops at the next station.
5: Prediction control is applied for a succeeding train only when its' preceding train is on the block section of the platform of the next station.

4.2 Estimation of Train Operation Curves after the Approach Point

The approach point and train operation curves after the approach point satisfy the following conditions.

1: If the preceding train do not depart from the station even when the predicted departure time has come, the succeeding train stop at the marginal stop point in front of the block with the station platform. So, the approach point is on the braking pattern to stop at the marginal stop point.
2: If the preceding train depart from the station when the predicted departure time has come, the succeeding train passes the approach point, and stop at the designated point on station platform.
3: If the preceding train depart from the station when the predicted departure time has come, headway between two trains is minimized.

These conditions can be represented in Fig.6. By using the above three conditions, the position and velocity of the approach point can be represented as follows. The time of the approach point is when the preceding train pass signal 2, and the aspect of signal 1 become green in Fig.6.

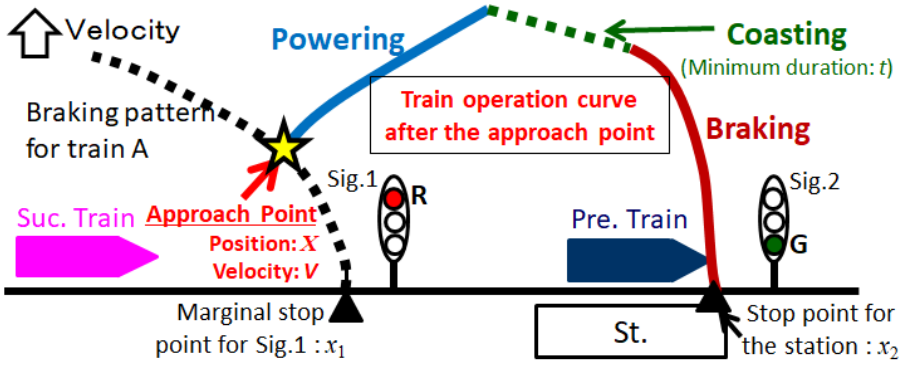


Fig. 6. Train operation curve with prediction control under fixed block

$$V = \sqrt{\frac{\beta^3\{(\gamma^2 + \alpha\beta - \beta\gamma - \gamma\alpha)t^2 - 2(\alpha - \beta)(x_2 - x_1)\}}{(\alpha - \beta)^2(\alpha - 2\beta)}} \tag{1}$$

$$X = x_1 + \frac{V^2}{2\beta} \tag{2}$$

where

x_1: Marginal stop point in front of the station platform

x_2: Designated stop point on the station platform

V: Velocity of the succeeding train at the approach point

X: Position of the succeeding train at the approach point

4.3 Estimation of Train Operation Curves before the Approach Point 1, "Energy Saving Strategy"

Based on the approach point solved in 4.2, train operation curves before the approach point is calculated based on the following preconditions and approaches.

1: The driving operation on the approach point is set as coasting. This is because the train have to change its' driving operation after the approach point, according to whether the preceding train depart from the station on the predicted time or not.

2: If the predicted departure time of the preceding train is changed due to the delay caused by congestion, (if possible) prediction control can be applied again by recalculating and updating the approach point.

In this section, we adopt the strategy for train operation curves to realize energy saving. That is, in the optimal train operation curves, we try to incorporate coasting operation as long as possible.

Estimation of train operation curves with prediction control is conducted by modifying the fastest operation curve. Slowdown driving operation, like coasting or braking, is added to that to meet the position, velocity, time of the approach point. The procedure of

estimating or updating train operation curves is different depending on the following conditions.

- Whether the preceding train is on the block section of the next station or not
- Whether the succeeding train already depart from the previous station or not
- Whether the predicted departure time become earlier or later

We developed the way of estimating train operation curves for each combination of the conditions above. In this paper, we describe the way of estimating train operation curves when the preceding train is on the block section of the next station, and the succeeding train already depart from the previous station.

Figure 7 illustrates the way to modify and update the train operation curves to meet the time of the approach point. Firstly, the train operation curve which passes the approach point with coasting operation, and stops at the marginal stop point in front of the station is created. Then, it is modified by extending the duration of the coasting in front of the approach point, until the train pass the approach point on the predicted time. In case when the predicted departure time of the preceding train become later, the duration of the coasting become longer to meet the updated time of passing the approach point.

Figure 8 illustrates another way to modify and update the train operation curves to meet the time of passing the approach point. If there is no room for the operation curve to extend the duration of the coasting, braking-coasting-powering operation is added to meet the time.

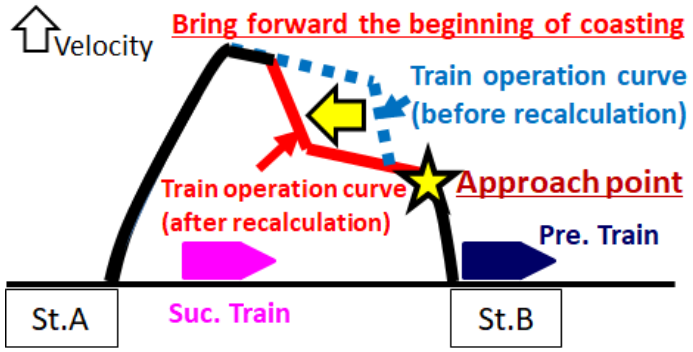


Fig. 7. Updates of train operation curve with prediction control (1)

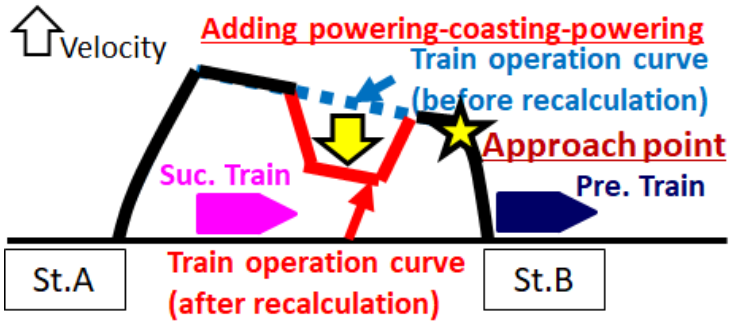


Fig. 8. Updates of train operation curve with prediction control (2)

If there is much time for the succeeding train to pass the approach point on the updated target time, it may stop between the stations, while it remains possible to pass the approach point on the updated target time by restarting.

Train operation curves created by these methods satisfy preconditions and approaches described above. There are coasting driving operation in front of the approach point. Prediction control can be applied again if the estimated time of passing on the approach point is updated. Train operation curves are energy saving ones because they adopt coasting operation as long as possible.

4.4 Estimation of Train Operation Curves before the Approach Point 2, ”Preventing Delay Propagation Strategy”

Although the train operation curve of the succeeding train described in the previous section can minimize train headway, they may influence the train after the succeeding train. To realize energy saving driving, the train operation curve of the succeeding train includes coasting operation as long as possible. This may in turn affect and bring in front for the marginal stop point of the train after the succeeding train. If the succeeding train runs as fast as possible under the condition that it passes the approach point, the influence for the train after the succeeding train may be decreased, and that prevents unnecessary delay propagation for trains running after the succeeding train.

In this section, we adopt the strategy for train operation curves to prevent delay propagation. To realize this idea, we devised another method for the train operation curve of the succeeding train. The left side of Fig. 9 illustrates each train operation curve for the succeeding train under the strategy described in 4.3 or 4.4. The running time between the two stations are the same. In the operation curve in 4.3, the duration time for coasting is long. On the other hand, in the operation curve in 4.4, there is braking-coasting-powering operation in front of the approach point. By this operation, the position of the succeeding train become forward to the next station, compared to that in the operation curve in 4.3. The difference is shown in the right side of Fig. 9, which is the time-space graph of the train operation curves described on the left side. By adopting this strategy, the train after the succeeding train can go ahead to the next station, and that leads to prevent or decrease delay propagation from the preceding train to the trains running after that.

When the predicted departure time of the preceding train is changed, the way to update the train operation curve to meet the new time of passing the approach point is the same as that described in 4.3.

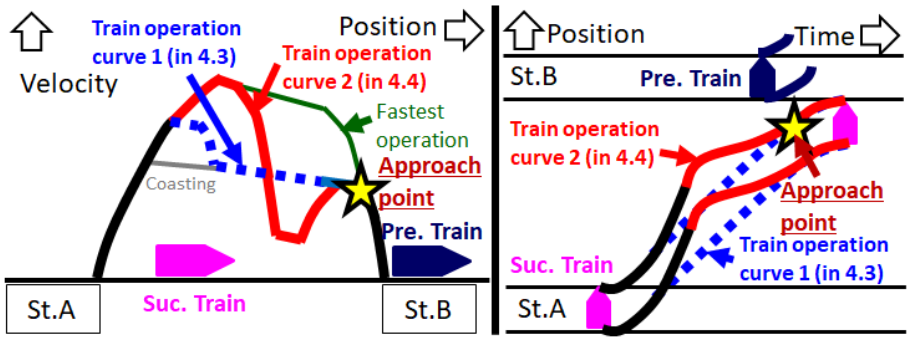


Fig. 9. Comparison of train operation curves under the two strategies

5 Train Traffic Simulation with Prediction Control under Moving Block

5.1 Precondition

Although the preconditions described in 4.1 are also adopted, as there are some differences between fixed block and moving block, it is necessary to analyse the traffic condition under moving block in which headway between trains is minimized.

Figure 10 illustrates the closest approach of successive two trains under moving block. There is the closest point at which distance between two successive trains is minimized. If we call that "contact point," we can calculate that and then the approach point, by setting and using the following preconditions in addition to those described in 4.1-4.4.

1: The driving operation of the succeeding train after the contact point is restricted to coasting and braking, until it stops at the station platform.
2: The driving operation of the succeeding train between the approach point and the contact point is restricted to coasting.
3: If the preceding train do not depart from the station even when the predicted departure time has come, the succeeding train stop at the marginal stop point in front of the station platform. So, the approach point is on the braking pattern to stop at the marginal stop point. The marginal stop point is set considering the buffer distance under moving block.

5.2 Estimation of Train Operation Curves after the Approach Point

The approach point, the contact point and train operation curve with prediction control satisfy the following conditions.

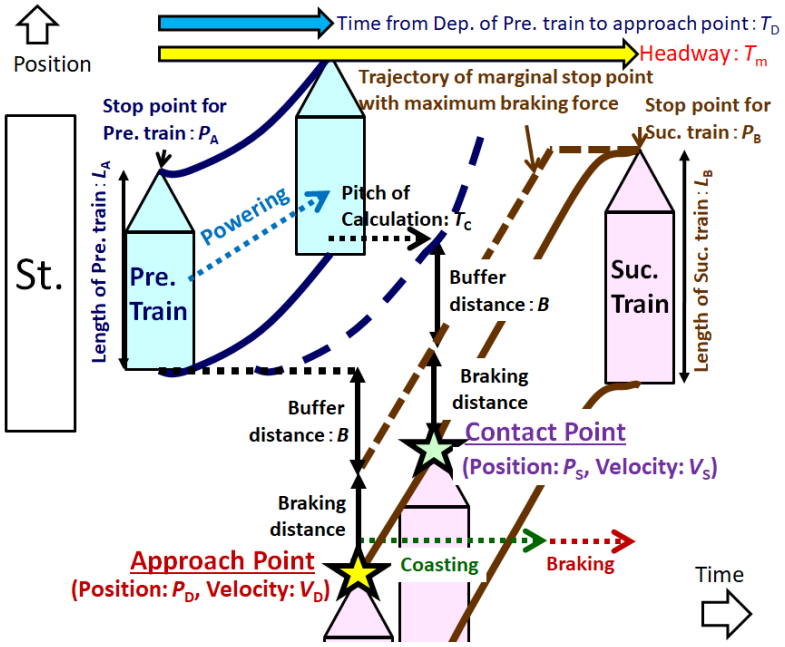


Fig. 10. Closest approach of successive two trains under moving block

1: At the contact point, the distance between the preceding train and succeeding train is minimized and the velocity of both trains become equal.
2: At the contact point, the distance between the preceding train and succeeding train is equal to sum of the buffer distance and braking distance necessary to stop by the maximum braking force.
3: If the preceding train depart from the station when the predicted departure time has come, the succeeding train passes the approach point, the contact point, and then stop at the designated point on station platform.
4: If the preceding train depart from the station when the predicted departure time has come, headway between two trains is minimized.

By using the above conditions, the position and velocity of the contact point can be calculated as follows. The minimum headway between the two trains are also calculated as follows.

$$V_s = \sqrt{\frac{-2\alpha\beta^3(\alpha-\gamma)^2 H}{\alpha^3(\beta-\gamma)(2\beta-\gamma)-\beta^3(\alpha-\gamma)^2-2\alpha^2\beta^2\gamma}} \tag{3}$$

$$P_B - P_s = H - \frac{(\alpha+\beta)V_s^2}{2\alpha\beta} \tag{4}$$

$$T_m = T_C + \frac{\gamma-\alpha}{\alpha\gamma}V_S - \frac{\sqrt{(\beta-\gamma)\left\{\left(\beta-\gamma-\frac{\beta\gamma}{\alpha}\right)V_S^2+2\beta\gamma H\right\}}}{\beta\gamma} \tag{5}$$

$$H = L_A + B - P_A + P_B \tag{6}$$

where

V_S: Velocity of the succeeding train at the contact point

P_A:Designated stop point on the station platform for the preceding train

P_B:Designated stop point on the station platform for the succeeding train

P_S: Position of the succeeding train at the contact point

T_m: Minimum headway between two successive trains

T_C: Cycle of calculation for train operation curves

L_A: Length of the preceding train

B: Buffer distance under moving block

The position, velocity and time of the approach point can also be calculated as follows.

$$V_D = \sqrt{\frac{\alpha\beta-\alpha\gamma-\beta\gamma}{\alpha(\beta-\gamma)}}V_S \tag{7}$$

$$P_A - P_D = L_A + B - \frac{V_D^2}{2\beta} \tag{8}$$

$$T_D = T_C + \frac{\gamma - \alpha}{\alpha\gamma} V_S + \frac{V_D^2}{\gamma} \quad (9)$$

where

V_D: Velocity of the succeeding train at the approach point

P_D: Position of the succeeding train at the approach point

T_D: Time period from departure of the preceding train from the station to passing of the succeeding train on the approach point

5.3 Confirmation of the Approach Point

To confirm the mathematical solutions described in 5.2, we calculated the approach point based on the parameters described in Table 1. In case 1, we suppose that train performance is general one, and the length of train is 200[m], which is typical in commuter lines in Tokyo. In case 2, high performance train, such as metros, is supposed to be operated in commuter lines in Tokyo. In case 3, the rail line in which the length of trains is short is supposed.

As the calculation results of minimum headway are realistic ones, we conclude that the calculation method is appropriate one.

Table 1. Calculation examples of the approach point

Parameters	Value		
	Case 1	Case 2	Case 3
$\alpha\ [km/h/s]$	1.6	3.0	1.6
$\beta\ [km/h/s]$	-1.8	-4.0	-1.8
$\gamma\ [km/h/s]$	-0.03	-0.05	-0.03
$L_A + B - P_A + P_B\ [m]$	210	210	100
$T_C\ [s]$	3.0	1.0	3.0
Calculation Results			
$V_S\ [km/h]$	29.9	43.0	20.6
$P_B - P_S\ [m]$	201	189	95.9
$T_m\ [s]$	54.4	36.6	38.4
$V_D\ [km/h]$	30.2	43.4	20.8
$P_A - P_D\ [m]$	280	275	134
$T_D\ [s]$	12.2	8.11	9.37

5.4 Estimation of Train Operation Curves before the Approach Point

Based on the approach point solved in 5.2, train operation curves before the approach point are calculated based on the same preconditions and approaches as those described in 4.3 or 4.4.

6 Test Calculation of Train Traffic under Commuter Line

6.1 Outline of the target rail line

In this paper, the effects of installing prediction control were evaluated. The railway line used in the evaluation had 19 stations, and about 1,000 trains in operation in a single day. The period used for the study was the morning rush hour between 7AM and 10AM, during which trains were running every 3 or 4 minutes. There were 208,335 passengers departing from origin stations between 7AM and 10AM.

We mainly analysed the train operation curves of the succeeding train supposing that the signalling system and train control system were as follows.

1) Fixed block without prediction control
2) Fixed block with prediction control
3) Moving block without prediction control
4) Moving block with prediction control

6.2 Scenario 1

We supposed that a small accident was occurred on the train in St. B at 8AM, and the train remain stopping during 3 minutes and 12 seconds. We estimated train traffic conditions under the scenario by using the improved simulator. When prediction control is applied, train operation curves before the approach point are estimated based on energy saving strategy described in 4.3.

Figure 11 illustrates the train operation curves of the succeeding train in case 1) and 4). In case 1), after the succeeding train depart from St. A, it stops between the stations due to the speed restrictions by the signal. Then, the preceding train departs from St. B, and the succeeding train restarts, and arrives at St. B. It takes 82 seconds in St. B from the departure of the preceding train to the arrival of the succeeding train.

On the other hand, in case 4), after the succeeding train depart from St. A, it drives slowly to St. B. It avoids intermediate stops between stations. When the preceding train depart from St. B, the succeeding train pass the approach point, and arrive at St. B. Headway becomes only 51 seconds. As a result, the extent of delay propagation to the succeeding train is reduced about 30 seconds. Figure 12 illustrates the train trajectory of both the preceding and succeeding train in case 1). Figure 13 illustrates that in case 4). In Fig.12, the succeeding train stops between St. A and St. B, and that leads to increase headway between trains. Moreover, as the signalling system is fixed block in 1), the succeeding train have to run within the speed restriction by the signal between St. B and St. C. It leads to increase of running time for the succeeding train. On the other hand, in Fig. 13, the succeeding train do not stop between St. A and St. B, and that leads to decrease headway between trains. Also, as the signalling system is moving block in 4), the succeeding train can run without the speed restriction by the signal between St. B and St. C.

Table 2 summarized the result of calculated headway in each case when the preceding train is delayed by 3 minutes and 12 seconds. It can be said that both moving block and prediction control is effective for the train traffic condition.

In addition, we evaluated effects of decreasing train headway under various conditions of departure delay of the preceding train. We set the preconditions that the time period of the preceding train remain stopping is varied from 60 sec. to 200 sec. Train headway between the two trains is calculated under each condition of block system and prediction control. The results are shown in Fig.14. By applying prediction control under moving block, train headway is minimized regardless of departure delay of the preceding train.

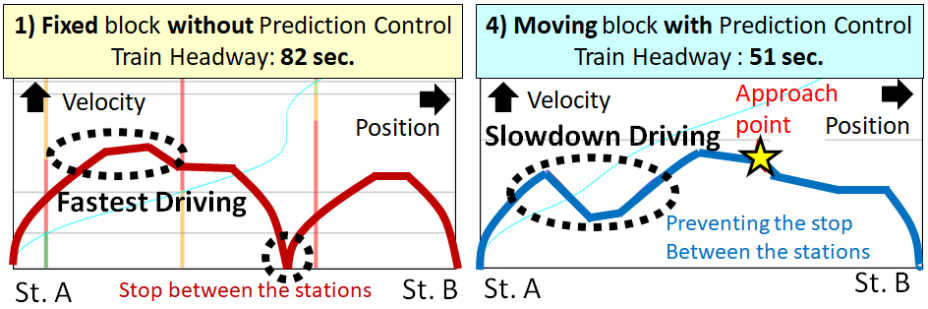


Fig. 11. Train operation curve of the succeeding train under each control system

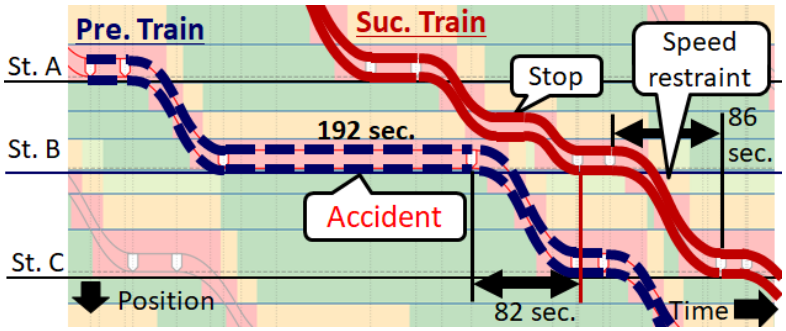


Fig. 12. Train trajectory under fixed block without prediction control

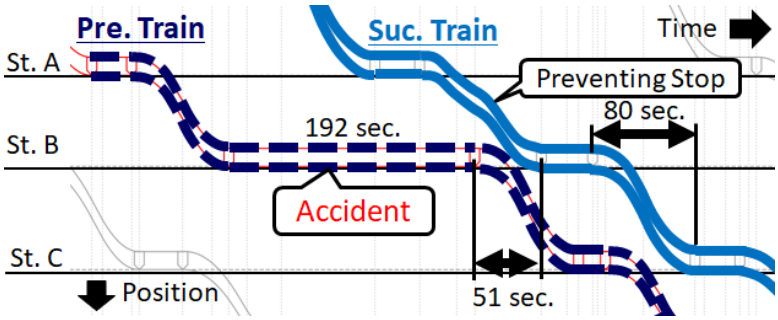


Fig. 13. Train trajectory under moving block with prediction control

Table 2. Comparison of headways

	Without prediction control	With prediction control 1: Energy Saving Strategy
Fixed block	1) 82 sec.	2) 80 sec.
Moving block	3) 56 sec.	4) 51 sec.

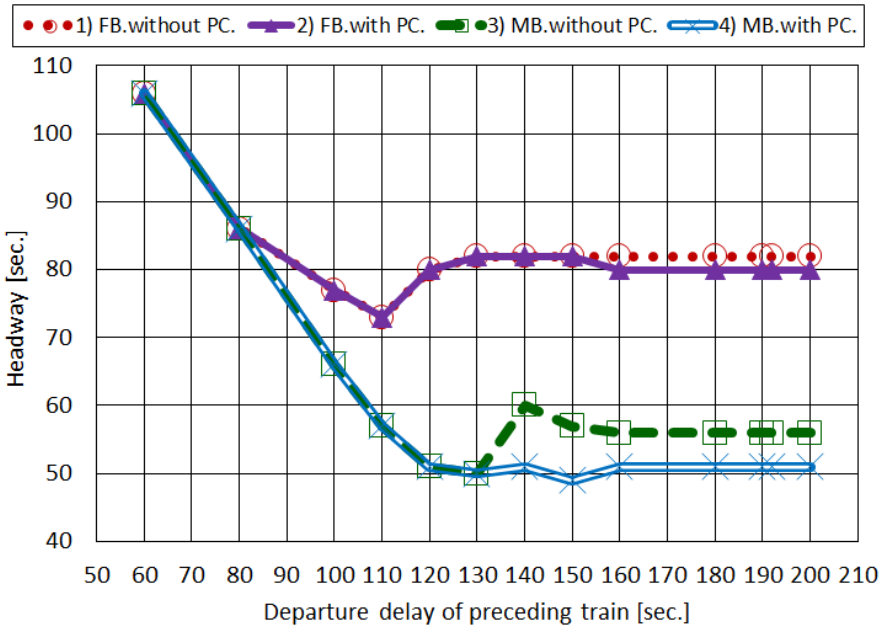


Fig. 14. Train headways under each condition

6.3 Scenario 2

We supposed that a small accident was occurred on the train in St. Q at 7:17, and the train remain stopping during 2 minutes and 5 seconds. We estimated train traffic conditions under the scenario by using the simulator. In this scenario, effects of prediction control under moving block is tested by comparing train operation curves under the conditions 3) or 4). When prediction control is applied, train operation curves before the approach point are estimated based on either strategy described in 4.3 or 4.4.

Figure 15 illustrates the train operation curve for the succeeding train under each condition. By applying prediction control, train headway become shorter. Comparing 4-1) and 4-2), although train headways are almost the same, the shape of train operation curves are different. In 4-2), the succeeding train can be operated halfway the same as that in 3). It decreases influences of delay propagation to the train after the succeeding train. So, we can select either strategy 4-1) or 4-2), along with the policies for train operation.

7 Conclusions

In this paper, we improved functions of “Train Operation/Passenger behaviour simulator” to reproduce train traffic under moving block and prediction control. We conducted test evaluation for train traffic in an actual commuter line in Japan, and confirmed the effects.

In particular, we proposed and formulated a new method for applying prediction control for trains under moving block. By combining the estimation function for passenger flow and train delay, we realized the simulation system by which prediction control can be applied repeatedly to minimize train headway, considering possibility of extension of dwell

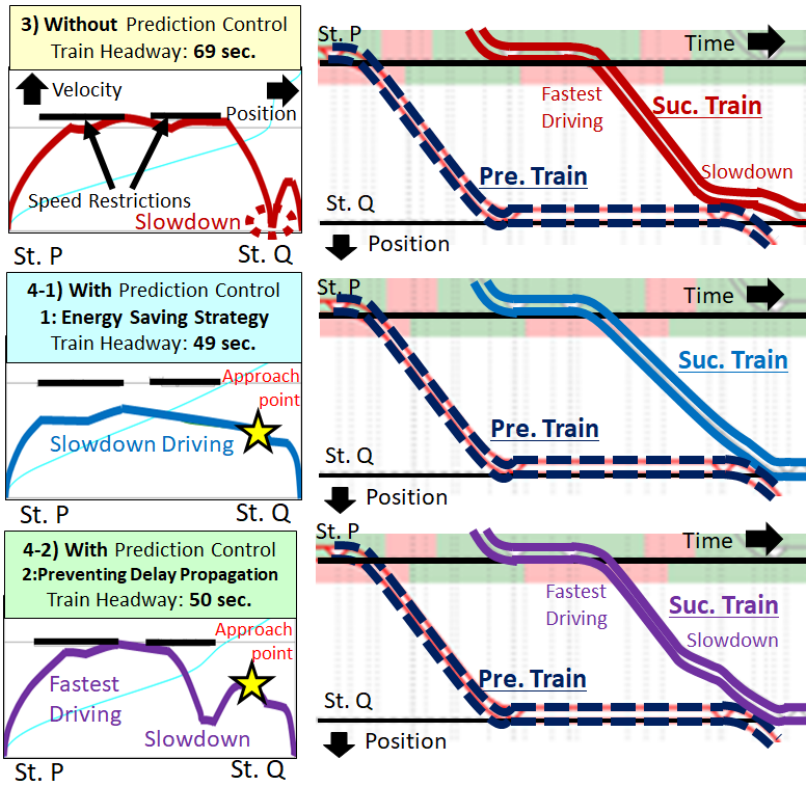


Fig. 15. Train operation curve of the succeeding train under each condition

time caused by passengers. Moreover, the simulation system can also reproduce the situation that prediction of departure of the preceding train was failed, and the succeeding train cannot be operated along with the optimized train operation curve by prediction control. At the left side of Fig. 16, when the preceding train arrive at the next station, the train operation curve for the succeeding train is optimized based on the planned dwell time of the preceding train. But, if the necessary time for boarding and alighting is estimated to be longer than the planned dwell time, the preceding train postpone the departure from the station. At the same time, as described on the right side of Fig.15, the train operation curve for the succeeding train is calculated and optimized again, based on the updated departure time of the preceding train. As there are many cases that predictions for dwell time are failed in commuter lines, we think the devised simulation method to reproduce the condition is one of the major contribution of this research.

For the future works, it is desired to evaluate under various scenario and conditions. We are also going to implement the method of predicting train delay based on past recorded data of actual delay (Nakabasami et al. (2019)). By combining the prediction method for train delay, we can utilize prediction control effectively, avoiding failure of prediction for the departure time of the preceding train. We will confirm and evaluate effects of the delay prediction by updating the simulator.

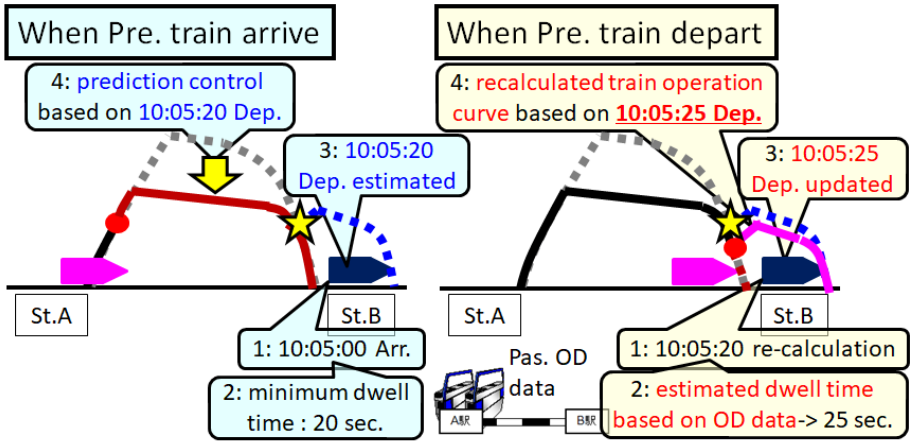


Fig. 16. Recalculation of train operation curves reflecting estimated dwell time based on passenger data

References

Baba, Y, et al., 2003. "Outline of Advanced Train Administration and Communications System" , Technical review, JR East (5), pp.31-38. (in Japanese)

D'Ariano, A, et al., 2005. "Conflict resolution and train speed co-ordination for solving timetable perturbations", Proc. of 1st International Seminar on Railway Operations Modelling and Analysis. (RailDelft)

Hiraguri, S, et al., 2004. "Advanced Train and Traffic Control Based on Prediction of Train Movement", JSME International Journal Series C Mechanical Systems, Machine Elements and Manufacturing, Vol.47, No.2. (2004)

Kanda, D, et al., 2014. "Simulation analysis of reduction of train delay under moving block in urban rail line", Proc. of the 69th annual Conference of JSCE, No.4-078. (in Japanese)

Kunimatsu, T. et al., 2018. "Development of Train Operation Simulator under Moving Block and Prediction Control", IEEJ Transactions on Industry Applications. Vol.138, No.4. (in Japanese)

Nakabasami, K, et al., 2019. "Prediction Method for Train Delay and Congestion Rate Using a Neural Network", IPSJ Journal. Vol. 60, No.4. (in Japanese)

Takeuchi, Y. et al., 2015. "Development of Detailed Model of Train Operation and Passenger Flow Simulation and Multicriteria Evaluation of Train Operation Plans", IEEJ Transactions on Industry Applications. Vol.135, No.4. (in Japanese)

Terasawa, T, et al., 2015. "Calculation Method of Train Performance Curve for the Train Prediction Control", Proc. of 2015 Annual Conference of IEEJ, No.5-146. (in Japanese)

Xu, F, et al., 2015. "Optimisation Framework for Rail Traffic Regulation at a Single Junction", Proc. of 6th International Conference on Railway Operations Modelling and Analysis (RailTokyo).

Yamashita, O, 2006. "System for Train Performance Evaluation, Drawing and Analysis," Transactions of Japan Train Operation Association, Vol.48, No.3. (in Japanese).

Infrastructure Capacity in the ERTMS Signaling System

Alex Landex [a,1], Lars Wittrup Jensen [a,2]
[a] Department of Planning & Rolling Stock, Ramboll
Hannemanns Alle 53, 2300 Copenhagen S, Denmark
[1] E-mail: alxl@ramboll.dk, Phone: +45 5161 1183
[2] E-mail: lawj@ramboll.dk, Phone: +45 5161 2148

Abstract
This article describes the main differences between level 1-3 in the new European signaling standard ERTMS and conventional signaling systems focusing on communication differences, the ability to look ahead and braking curves. Based on this description, the capacity differences between level 1 and 2 are investigated for theoretical as well as real-life cases using line headway calculation models developed for the study.

The results show ERTMS level 2 generally has shorter headways than level 1 and hence higher capacity. However, in homogeneous operation where the braking distance is well-adapted to the block lengths, level 1 can have shorter headways than level 2 due to less system delays. The results also show that Level 2 due to continuous update of the Movement Authority (MA), result in higher capacity than level 1 for longer block sections and lower speeds.

The article discusses that a 1:1 replacement of conventional signaling with ERTMS can lead to loss of capacity as the ERTMS braking curves are likely to be longer. The article also discusses how extra capacity can be gained with ERTMS as it is possible to look more block sections ahead.

Keywords
Railway capacity, Signaling, ERTMS, ETCS, Braking curves

1 Introduction

ERTMS is the new signaling standard in Europe but has also been adapted (with some modifications) in other parts of the world (mainly Asia). ERTMS consists of a standardized control system ETCS (European Train Control System) and a communication standard GSM-R (Global System for Mobile communication for Railways).

ERTMS exists in five different basic levels: Level 0-3 and level NTC (National Train Control). Level 0 enables trains equipped with ERTMS to operate on infrastructure not equipped with ERTMS, and where there is no alternative train protection or warning system. Level NTC enables trains with ERTMS to operate on infrastructure where the national train control system needs to be operated. The pure ERTMS levels range from the simplest at level 1 to the most advanced at level 3 – and some hybrid versions as well as adaptions to other markets like the Chinese. This article focuses on ERTMS level 1 and 2.

In general, ERTMS level 1 is similar to a conventional multi-aspect signaling system with ATP (Automatic Train Protection) where the train is updated discretely with new movement authority at balises (potentially with infill by balises, loops or radio). In ERTMS level 2, the communication between train and infrastructure is updated continuous allowing the train's movement authority to be continuously updated and shown to the driver. Level

3 is a moving block system with no (or only limited) train detection in the track needed why the position of the train is continuously sent from the train and a train integrity system is needed. The different signaling systems are compared in Table 1.

Table 1: Comparison of different signaling systems.

	Conventional	**Conventional multi-aspect**	**Level 1**	**Level 2**	**Level 3**
Train control	Possible	Possible	Included	Included	Included
Communication	Discrete (infill possible)	Discrete (infill possible)	Discrete (infill possible)	Continuous	Continuous
Signal aspects	2 (Red/green)	3+	Movement authority	Movement authority	Movement authority
Signal visibility	Needed	Needed	Usually needed	Not needed	No signals
Train detection in track	Needed	Needed	Needed	Needed	Limited (on train and turnouts)
Train integrity	Not needed	Not needed	Not needed	Not needed	Crucial
Train position	Known in block section	Known in block section	Known in block section	Known in block section but can be more exact	"Exact" position known

With few exceptions, higher levels of ERTMS result in increased level of capacity which is covered by numerous publications, e.g. UNIFE (2014). Capacity of different levels of ERTMS (and variations within different levels of ERTMS) is well examined e.g. UIC (2008). Higher levels of ERTMS generally leads to higher capacity as illustrated in Figure 1.

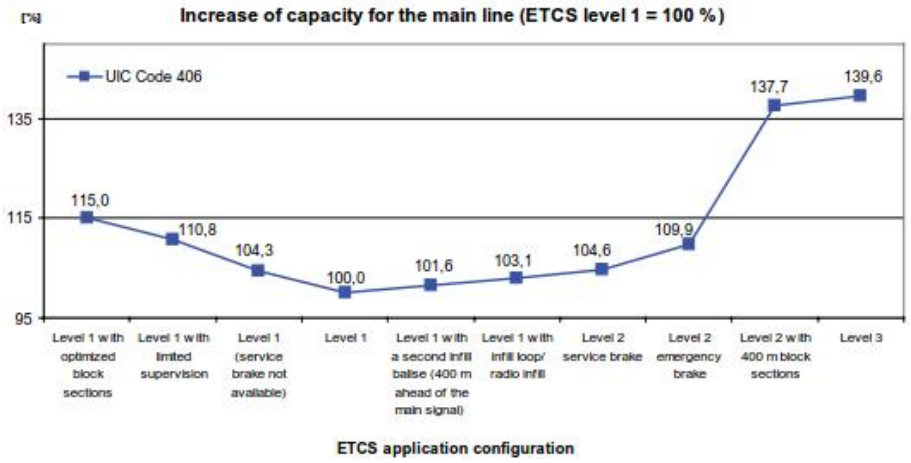


Figure 1: Influence of different ETCS levels on line capacity (UIC, 2008).

Increased capacity is often used as one of the selling points for implementing ERTMS.

However, capacity is often lost when going from multi-aspect conventional signaling to ERTMS (e.g. Goverde et al., 2013), especially if converting the signaling in a 1:1 ratio. This is mainly due to more conservative braking curves and because the multi-aspect conventional signaling system has been "optimized" to increase capacity. For the simpler single-aspect conventional signaling, there will usually be a gain in capacity when implementing ERTMS as the ability to read signal aspects further than one block section ahead is introduced.

This article describes the main differences in ERTMS that affects infrastructure capacity. Based on a line headway calculation models developed for this study, the article analyses infrastructure capacity of ERTMS level 1 and level 2 for both theoretical and practical cases.

2 ERTMS and Infrastructure Capacity

ERTMS is, as shown in Table 1, similar to conventional signaling. However, especially the differences in communication, the ability to look more block sections ahead and the braking curves result in changed infrastructure capacity. The following sections describe these parameters and their impact on infrastructure capacity

2.1 Communication

The biggest differences between the different levels of ERTMS (and to conventional signaling) is within the communication. The higher level of ERTMS, the more communication is required between the train and signaling system, cf. Table 2.

Table 2: Communication differences in ERTMS.

	Level 1	**Level 2**	**Level 3**
Communication between train and infrastructure	Line Electronic Units (LEUs) and Eurobalises	Eurobalises and RBC	Eurobalises and RBC
Role of Eurobalise	Position & signal state	Position	Position
Location of train	Track detection equipment	Mainly track detection equipment	Position information from train
Movement Authority	From Eurobalise	From RBC	From RBC
Radio	Voice	Voice and data	Voice and data

The differences in communication result in discrete update of the movement authority to the train driver in level 1 but continuous update in level 2 and 3. Increased communication of position as well as train integrity system in level 3 furthermore allows moving block. This leads to the possibility of shorter headways between the trains, cf. Figure 2.

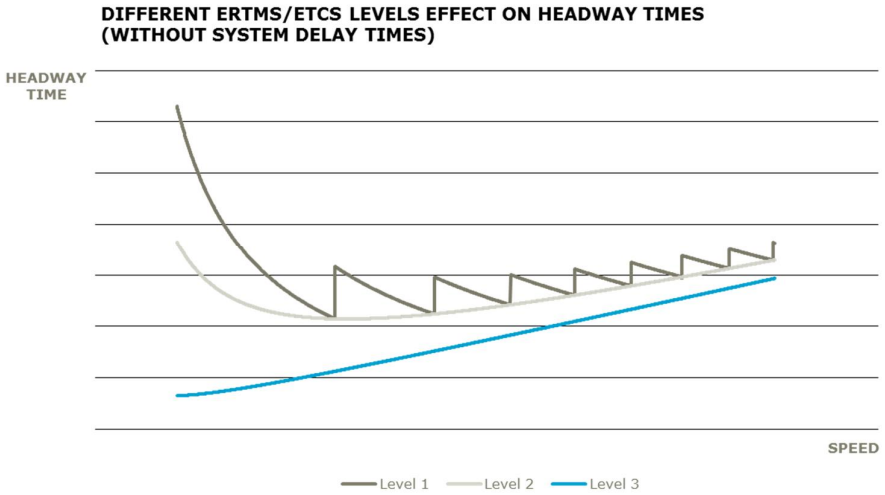


Figure 2: Headway time for different speeds and levels of ERTMS.

Figure 2 is a conceptual figure showing the minimum headway time for different levels of ERTMS depending on the speed. For level 1, the optimal headway times are when the braking distance is equal to the sum of blocks in the braking distance. Here, the headway time is the same for level 1 and 2 when system delays are not considered.

The optimal travel speed is when the minimum headway time is as short as possible. When the travel speed is below the optimal travel speed the minimum headway time can be reduced by speeding up since the block occupation time is too long. At travel speeds above the optimal travel speed, the braking distance has become too long, so that the block sections are reserved for too long time. It is not possible to have travel speeds which require looking more block sections ahead than the signaling system allows.

The increased and changed communication leads to higher communication times. The longer communication times in level 2 compared to level 1 (cf. Table 3) results in shorter headway times when the braking distance in level 1 matches the block lengths whereby level 1 in homogeneous operation can have shorter headways and hence more capacity than level 2. With infill for level 1, level 1 can result in more capacity than level 2 for a larger interval in braking distance (cf. Figure 1).

Table 3: Communication times used in the theoretical and practical calculations (Liikenevirasto, 2018).

Type	Level 1	Level 2
LEU (Lineside Electronic Unit)	0.7 sec	–
Communication delay (train to/from RBC)	–	2.65 sec
Interlocking delay (no turnouts)	5 sec	5 sec
EVC (European Vital Computer) + DMI (Driver Machine Interface)	1 sec	1 sec

2.2 Ability to Look Ahead

ERTMS gives the possibility to look more block sections ahead than conventional signaling. This due to more modern technology compared to mechanical and relay based signaling

systems, where it in the electronic signaling system is easier—and less expensive—to look more block sections ahead. Besides, ERTMS has cab signaling which ensure the Movement Authority (MA) is shown to the driver on the Driver Machine Interface (DMI) in the cab. The number of signal aspects that can be communicated to the driver is therefore no longer a restricting factor.

The length of block sections in conventional (multi-aspect) signaling systems, where it is only possible to look few block sections ahead, are determined by the need to be able to stop the train within the length of the block sections indicated to the driver. This restricts how short the block sections can be for conventional signaling systems. If a train is unable to stop within the signal aspect given, the train's speed will need to be limited.

For ERTMS, where it is possible to look more block sections ahead, it is possible to have shorter block sections allowing for shorter headways. Furthermore, it is no longer needed to limit a train's speed due to the signaling system. This can potentially allow for faster freight trains resulting in higher capacity and/or faster high-speed trains on the infrastructure.

2.3 Braking Curves

As an ATP system, ETCS monitors the train's speed and position to ensure that the train does not run above the allowed speed or pass a given movement authority. This is achieved by calculating a braking curve for the train taking the braking performance, gradients, uncertainties and various correction factors into account. If the driver does not brake the train within the supervised limits of the calculated ETCS braking curve, the onboard ETCS equipment will intervene to brake the train.

In ETCS, the braking curve calculated is denoted the emergency brake deceleration (EBD) curve. It is also possible to use the (full) service brake deceleration (SBD) curve before emergency braking is initiated. This is preferred for comfort and as the emergency brake can damage the rolling stock and the track. However, in ETCS it is not a requirement to use the SBD curve.

In Figure 3 an example deceleration is shown including the EBD curve and the different supervision limits and interventions. When the train approaches a speed restriction the driver will be given an indication (I) that tells the driver to initiate braking to prevent driving faster than the permitted speed (P) as the permitted speed shown to the driver is decreasing. If the driver fails to brake according to the permitted speed an additional audible warning (W) is given before the onboard equipment intervenes and either initiate full service braking intervention (SBI) or emergence braking intervention (EBI). From the intervention to the EBD curve is reached, time is added to account for speed measurement inaccuracies and a possible acceleration during the brake build up time before the full braking performance is achieved. Furthermore, additional distance (time) is added for inaccuracies in the location of the train. The onboard equipment calculates a location confidence interval that ensures a safe location of the train as shown in Figure 3 (max safe front). The confidence interval is calculated as (up to) ±(5m+5%s) where s is the distance travelled since the last location balise (where the location confidence is reset) (UNISIG, 2015).

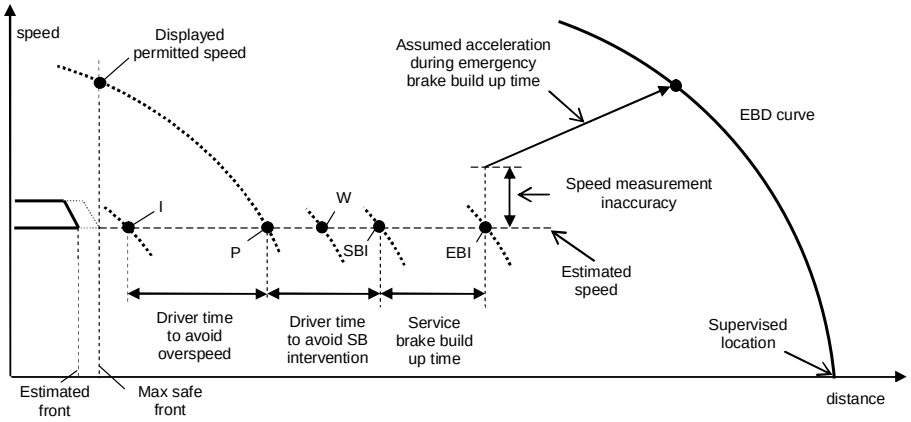


Figure 3: Emergency brake deceleration (EBD) and supervision limits (ERA, 2016).

The emergency brake deceleration (or full-service brake deceleration) curve itself is not easily calculated. Traditionally, brake weight percentage (BWP) has been used to define the braking performance of trains. This means that the nominal braking performance of train in terms of deceleration values (m/s^2) are not always available. Furthermore, the braking curves of many conventional signaling system are calculated based on the brake weight percentage, e.g. the Danish ATC system. To ensure easier transition to ETCS, ETCS offers two models for calculation of the braking curve (ERA, 2016; ERA UNISIG, 2016). One is the lambda model based on the brake weight percentage (denoted lambda, λ), the other is the gamma model based on the nominal braking performance of the train. Both yields safe braking curves that are subsequently corrected for gradients based on data from the trackside equipment. The gamma model is used for all trains that have well defined train characteristics, i.e. train sets and push-pull trains with a defined set of cars. The lambda model is used for freight trains and trains where the nominal braking characteristics cannot be obtained.

For both models, input values are given by the railway undertaking (the train data) and the infrastructure manager (the national values). The national values supplied by the infrastructure manager may differ from country to country to account for different national safety practices. This means that the same train running from one country to another on the same kind of infrastructure might have different braking curves due to national values, although the maximum braking effort of the train does not change.

The lambda model is based on a conversion model that converts the brake weight percentage of the train (λ) to converted deceleration values in m/s^2 for different speed intervals (A_brake_converted). These deceleration values are subsequently corrected by the integrated correction factors (Kv_int, and Kr_int) from the infrastructure manager based on the train type (passenger or freight), P or G braking and the length as shown in Figure 4. The deceleration values obtained (A_brake_tuned) ensures a safe braking due to the integrated correction factors (national values) and the conversion model that has been validated trough braking tests (ERA, 2016). As the lambda model yields a conservative braking, it is likely that the EBD curve is longer than in a conventional signaling system where the braking curve calculation has been optimized as mentioned in Section 1.

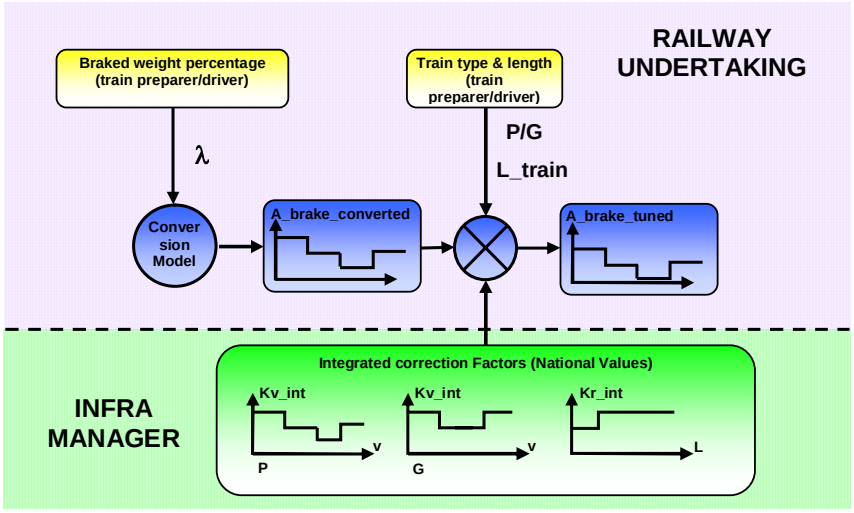


Figure 4: Braking curve estimation using the lambda model. (ERA, 2016).

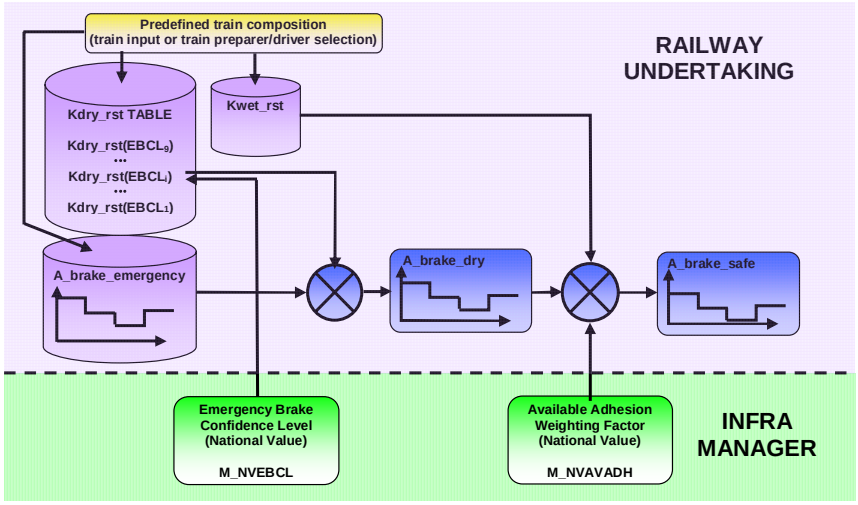


Figure 5: Braking curve estimation using the gamma model. (ERA, 2016).

The gamma model is based on the nominal braking performance of the train in terms of deceleration values m/s^2 for different speed intervals (A_brake_emergency or A_brake_service). This gives a more precise representation of the train's braking performance than the lambda model. As shown in Figure 5, the nominal emergency braking performance of the train is adjusted by the Kdry_rst value to obtain the safe braking performance on dry track (A_brake_dry). The Kdry_rst value is selected from a table of values. Each value in the table provides increasing safer brake performance. Basically, each value is tied to the probability that the train will brake according to nominal emergency brake performance multiplied by Kdry_rst on dry track. The Kdry_rst value is chosen based on the emergency brake confidence interval (M_NVEBCL), a national value provided by

the infrastructure manager. This value spans from 50% to 99.9999999%, where the first ensures safe braking every second time on dry track and the latter essentially every time. Low values will result in shorter braking curves than higher values which results in improved capacity by shorter headways as well as the possibility of stopping faster at stations e.g. towards buffer stops and end of MA. For lower values (shorter braking curves) , it may be needed to addoverlaps to maintain/improve safety. The M_NVEBCL value may be optimized for capacity while maintaining overall safety when overlaps are used using a Monte Carlo approach as described by Meyer et al. (2011).

The dry brake performance of the train is subsequently adjusted to obtain A_brake_safe using both a factor for available adhesion from the infrastructure manager (M_NVAVADH) and a factor for the train describing the train's braking performance on rails with reduced adhesion (Kwet_rst).

In addition to the national values described in this section, there exist more national values that also has impact on the braking curves. We will not go into depth with these in this paper. Table 4 summarizes and compare the two different braking curve calculation models.

Table 4: Differences between lambda and gamma braking curves.

Type	**Lambda**	**Gamma**
Precision	Low/limited	High
Number of parameters	Few	Many
Generally used for	Freight trains as exact braking parameters are not known	Train units as braking performance is well-defined

The length of the braking deceleration curve, whether calculated as gamma or lambda, and the associated supervision limits has a large impact on the infrastructure capacity as the time required for braking constitutes a larger proportion of the minimum headway time compared to other signaling related parameters. This is especially the case at high speeds as shown by Abril et al. (2008). For capacity planning, the permissive (P) is mainly used for the headway calculations. The indication (I) may also be used as this will give a conservative braking length estimate in normal operation (cf. Figure 3). If the train driver drives more aggressively or ATO is used, i.e. closer to the warning speed, a decrease in headway, and thus improved capacity, can be obtained.

Comparing ECTS braking curves with conventional signaling systems, the ETCS braking curves tend to be more conservative. This is a consequence of the calculation models and the associated correction factors (national values) chosen. An infrastructure manager migrating from a conventional signaling system to ERTMS may choose nation values that result in braking curves (for different rolling stock) as close as possible to the ones in the existing conventional system. However, this will result in some braking curves being longer than in the conventional system and some braking curves being shorter. The latter is a problem as it means that safety is reduced. The national values are thus (in early evaluation phases) chosen to ensure that no braking curves are (significantly) shorter than in the conventional system. This result in ETCS braking curves being generally longer resulting in capacity loss.

3 Methodology

Our analyses of infrastructure capacity using ERTMS is divided in two parts. The first part consists of theoretical calculations for various fixed block lengths traversed at constant speed while the second part consists of practical calculations on lines with varying block lengths and speeds. In both parts we analyze the capacity gains and losses between ERTMS level 1 and 2. As described in section 2.3, the braking curves are a crucial input for both the theoretical and practical calculations. For our capacity analyses, we calculate the braking curves using the ERA brake calculation tool (ERA, 2018) as input to our headway calculation. The data and methodologies for the two parts are described in the following three sections.

3.1 Train Data

For our analyses, three types of trains are used: a freight train, an IC train and a fast/express train. Braking curves for the latter are calculated using both the lambda and the gamma model. The data for the trains is shown in Table 5. For the theoretical calculations, we only use the IC and Freight trains while all types are used in the practical calculations.

Table 5. Train data for the analyses.

Train	**Freight**	**IC**	**Fast (Lambda)**	**Fast (Gamma)**
Weight [t]	2006	462	328	
Length [m]	515	177	159	
Maximum speed [km/h]	90	200	220	
Start acceleration (m/s^2)	0.19	0.8	0.4	
Avg. deceleration (gamma) [m/s2]				1.05
Brake weight percentage (lambda) [m/s^2]	54	135	135	
Used in theoretical calculations	•	•		
Used in practical calculations	•	•	•	•

3.2 Theoretical Calculations

The purpose of the theoretical calculations is to map the capacity gains and losses between ERTMS level 1 and level 2 for various fixed block length sizes. As part of the theoretical calculations, a sensitivity analysis is also conducted for the communication delay to the radio block center (RBC) in ERTMS level 2. In this sensitivity analysis the nominal delay of 2.65 seconds (cf. Table 3) is compared with an increased delay of 7 seconds.

The calculations are defined as theoretical as the trains travel at their maximum line speed and all block sections on the line are equal in length. Thus, acceleration and braking are not considered, and the calculations are therefore most realistic on the middle of a line, not at the ends of the line.

An automated Excel tool has been set op for the theoretical calculations. The tool iterates through all combination of parameters for a line headway calculation with fixed block lengths and speeds. The minimum line headway is the minimum separation time between two trains on a line the ensures that both trains can run on the line unhindered. The

parameters combined for headway calculation to form the mapping of capacity gains (and losses) are:

- Speeds: 60 to 200 km/h in increments of 10 km/h (although ERTMS can handle increments of 5 km/h)
- Block lengths [m]: 500, 750, 1000, 1200, 1600, 2000, 2500, 3000, 4000

The calculations are carried out for IC and Freight trains as described in Section 3.1. The two train types are combined to simulate both homogenous and heterogenous operation. For the headway calculations, the IC and freight trains are thus combined in all four possible ways, i.e. (1st train, 2nd train): (IC, IC), (IC, Freight), (Freight, IC), and (Freight, Freight). For ERTMS, the system reaction times shown in Table 3 are used.

3.3 Practical Calculations

As described in Section 3.2, the theoretical calculations do not take acceleration, braking, and varying block lengths into consideration. We have therefore also carried out calculations for two real-life lines in the Nordics where the acceleration, braking, all block lengths, dwell times and speed profiles are taken into consideration.

The two lines are divided into seven respectively two line sections (denoted line section 1-7 and 8-9). The line speed is in the range 130-200 km/h for the first line and 220 km/h for the second line.

As shown in Table 5, we use an IC train and a freight train as in the theoretical calculations, but also a high-speed train is used. As in the theoretical calculations, we map the capacity gains and losses between the different signaling systems for different combinations of trains taking the actual block lengths, speed profile and timetable into account. Again, the line headway forms the basis for the capacity estimation. To estimate the minimum line headway, we have developed a calculation model in C++ that uses blocking time theory (Happel, 1959) to estimate the block occupations and subsequently calculate the line headway between trains as described in Pachl (2008). Blocking time theory is the same approach as used in commercial tools (e.g. OpenTrack and RailSys). To estimate the time spent by a train in each block, we use the running time estimation model described in Jensen (2015). This model takes acceleration, braking, and the speed profile into account. The estimated train running times include timetable supplements recommended by UIC (2000). Complex train movements in junctions are not considered in the model. This is to be implemented at a later stage to make it possible to analyze the capacity gains and losses with ERTMS in major junctions in detail.

4 Results

Results of the theoretical and practical calculations are described in the next two sections based on the methodologies described in Section 3.

4.1 Theoretical Results

Based on the theoretical calculations it is confirmed that the continuous update of level 2 generally result in higher capacity than discrete update for level 1. However, the larger systems and communication delays in level 2 (cf. Table 3) decreases the capacity gain of continuous update. In case of the braking length matching the block lengths, level 1 can result in shorter headways than level 2 as continuous update has no immediate effect and level 1 has less system and communication delays, cf. Figure 6.

Homogeneous operation

Average capacity improvement with ERTMS level 2 (IC trains only)										
Speed/Block length	500 m	750 m	1000 m	1200 m	1600 m	2000 m	2500 m	3000 m	4000 m	Average
60 kph	27,5%	9,5%	20,7%	27,8%	38,8%	46,9%	54,5%	60,1%	68,1%	39,3%
70 kph	17,7%	2,3%	13,6%	20,9%	32,4%	40,9%	49,0%	55,1%	63,8%	32,8%
80 kph	9,0%	32,1%	7,0%	14,4%	26,2%	35,1%	43,6%	50,1%	59,5%	30,8%
90 kph	1,2%	23,5%	0,9%	8,4%	20,3%	29,5%	38,4%	45,2%	55,3%	24,7%
100 kph	17,6%	15,8%	33,2%	2,7%	14,7%	24,1%	33,3%	40,5%	51,0%	25,9%
110 kph	9,5%	8,6%	25,6%	37,2%	9,4%	18,9%	28,3%	35,8%	46,8%	24,5%
120 kph	2,3%	2,1%	18,6%	30,0%	4,4%	14,0%	23,5%	31,2%	42,7%	18,8%
130 kph	14,4%	21,7%	12,2%	23,3%	-0,3%	9,3%	18,9%	26,8%	38,6%	18,3%
140 kph	6,5%	13,9%	5,6%	16,4%	34,7%	4,3%	14,1%	22,0%	34,2%	16,9%
150 kph	-1,1%	6,4%	-0,9%	9,5%	27,5%	-0,7%	9,1%	17,1%	29,6%	10,7%
160 kph	-6,7%	-0,6%	18,7%	-3,1%	20,6%	35,3%	4,2%	12,3%	25,1%	13,9%
170 kph	-0,6%	12,2%	11,5%	24,7%	14,3%	28,6%	-0,4%	7,8%	20,6%	13,2%
180 kph	5,5%	5,2%	4,9%	17,6%	8,3%	22,4%	37,3%	3,4%	16,4%	13,4%
190 kph	-1,4%	-1,3%	-1,3%	10,9%	2,6%	16,4%	31,1%	-0,9%	12,2%	7,6%
200 kph	3,4%	8,4%	15,0%	4,9%	25,8%	10,8%	25,3%	37,7%	8,1%	15,3%
Average	7,9%	10,7%	12,2%	16,8%	18,6%	22,4%	27,3%	29,6%	38,1%	20,4%

Heterogeneous operation

Average capacity improvement with ERTMS level 2 (IC & freight trains)										
Speed/Block length	500 m	750 m	1000 m	1200 m	1600 m	2000 m	2500 m	3000 m	4000 m	Average
60 kph	15,2%	6,8%	20,3%	29,3%	21,6%	30,3%	38,7%	45,3%	55,0%	29,2%
70 kph	13,9%	10,5%	11,2%	20,1%	34,8%	23,2%	32,0%	39,0%	49,4%	26,0%
80 kph	4,0%	18,4%	3,1%	11,8%	26,4%	16,6%	25,7%	32,9%	43,8%	20,3%
90 kph	2,4%	19,3%	8,2%	4,3%	18,6%	30,3%	19,6%	27,0%	38,4%	18,7%
100 kph	10,0%	15,7%	23,5%	1,6%	16,0%	27,7%	17,2%	24,8%	36,4%	19,2%
110 kph	6,3%	12,4%	19,9%	18,0%	13,4%	25,5%	14,8%	22,5%	34,3%	18,5%
120 kph	2,9%	9,3%	16,7%	14,6%	11,1%	22,9%	12,5%	20,3%	32,4%	15,9%
130 kph	8,6%	18,6%	13,6%	11,5%	8,8%	20,7%	10,3%	18,2%	30,4%	15,6%
140 kph	4,9%	15,0%	10,5%	8,2%	25,6%	18,3%	8,0%	15,9%	28,3%	14,9%
150 kph	1,3%	11,4%	7,4%	4,9%	22,2%	15,8%	5,6%	13,5%	26,0%	12,0%
160 kph	5,0%	8,0%	16,8%	1,8%	18,9%	33,3%	3,2%	11,2%	23,8%	13,6%
170 kph	1,6%	14,2%	13,3%	12,5%	15,8%	30,1%	1,0%	9,0%	21,7%	13,2%
180 kph	4,5%	10,8%	10,2%	8,8%	13,0%	27,0%	19,3%	6,9%	19,6%	13,3%
190 kph	1,2%	7,7%	7,2%	5,6%	10,2%	24,2%	16,3%	4,8%	17,5%	10,5%
200 kph	3,5%	12,4%	14,1%	2,7%	21,5%	21,5%	13,5%	23,6%	15,6%	14,3%
Average	5,7%	12,7%	13,1%	10,4%	18,5%	24,5%	15,8%	21,0%	31,5%	17,0%

Figure 6: Capacity improvement for level 2 compared to level 1 for homogeneous operation (left) and heterogeneous operation (right) for different block lengths and line speeds.

The results in Figure 6 show that level 2 generally results in the highest capacity gains for longer block sections and lower speeds. This is because the continuous update of level 2 has more effect when trains occupy the block sections for longer time.
Comparing homogenous and heterogeneous operation, it can be seen from Figure 6 that higher capacity gains are achieved with homogeneous operation for long block sections and low speeds (red circles in Figure 6), while higher capacity gains are achieved with heterogenous operation for short block sections and high speeds (blue circles in Figure 6). This is due to variation in block occupation times where it for short block occupation times is less likely that the trains in heterogeneous operation will have braking distances matching the block lengths, while it for long block occupation times increases the probability that some trains will have braking distances better matching the block lengths.

Figure 7 shows cumulative distributions of capacity improvement from level 1 to level 2 for homogenous operation (IC trains) and heterogenous operation with four different train combinations (IC, IC), (IC, Freight), (Freight, IC), and (Freight, Freight) as described in Section 3.2.

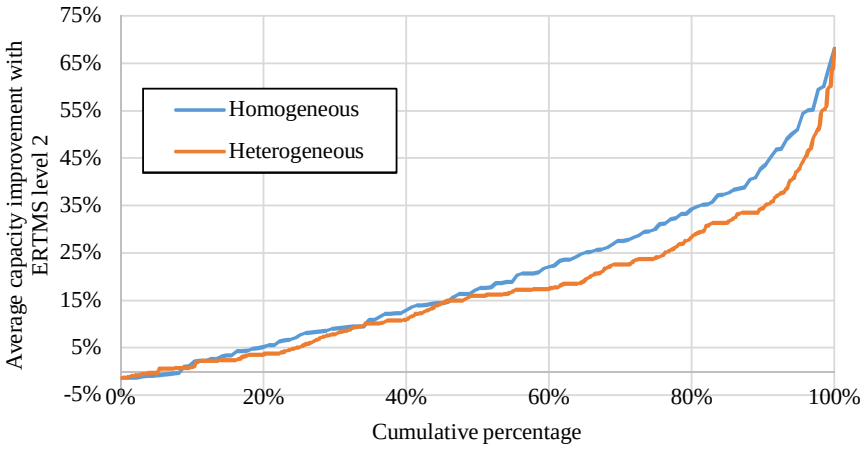


Figure 7: Cumulative capacity improvements for level 2 compared to level 1 for homogeneous operation and heterogeneous operation.

From Figure 7 and the corresponding Table 6, it is seen that the capacity gain of ERTMS level 2 vs level 1 is higher for homogeneous than heterogeneous operation. While significant capacity gains are possible, it can be observed that most train combinations have

more moderate gains, and only few combinations result in loss of capacity for level 2 compared to level 1.

Table 6: Capacity improvement for level 2 compared to level 1 in percent.

	Min	5%	25%	50%	75%	95%	Max	Std. dev	Avg.
Homogeneous	-1.4	-0.6	8.0	17.6	30.5	52.1	68.1	16.3	20.4
Heterogeneous	-1.4	-0.2	5.2	15.9	24.1	42.7	68.1	13.6	17.0

ERTMS level 2 is more sensitive to delays in the communication system then level 1 as the Movement Authority (MA) is received by radio. If the movement authority is not received timely, the train headway will increase, and in worst case, the train will be emergency braked. The sensitivity of the system delay is illustrated in Figure 8 for 2.65 seconds (cf. Table 3) and 7 seconds.

Homogeneous operation – 2.65 seconds system delay

Speed/Block length	500 m	750 m	1000 m	1200 m	1600 m	2000 m	2500 m	3000 m	4000 m	Average
60 kph	27,5%	9,5%	20,7%	27,8%	38,8%	46,9%	54,5%	60,1%	68,1%	39,3%
70 kph	17,7%	2,3%	13,6%	20,9%	32,4%	40,9%	49,0%	55,1%	63,8%	32,8%
80 kph	9,0%	32,1%	7,0%	14,4%	26,2%	35,1%	43,6%	50,1%	59,5%	30,8%
90 kph	1,2%	23,5%	0,9%	8,4%	20,3%	29,5%	38,4%	45,2%	55,3%	24,7%
100 kph	17,6%	15,8%	33,2%	2,7%	14,7%	24,1%	33,3%	40,5%	51,0%	25,9%
110 kph	9,5%	8,6%	25,6%	37,2%	9,4%	18,9%	28,3%	35,8%	46,8%	24,5%
120 kph	2,3%	2,1%	18,6%	30,0%	4,4%	14,0%	23,5%	31,2%	42,7%	18,8%
130 kph	14,4%	21,7%	12,2%	23,3%	-0,3%	9,3%	18,9%	26,8%	38,6%	18,3%
140 kph	6,5%	13,9%	5,6%	16,4%	34,7%	4,3%	14,1%	22,0%	34,2%	16,9%
150 kph	-1,1%	6,4%	-0,9%	9,5%	27,5%	-0,7%	9,1%	17,1%	29,6%	10,7%
160 kph	6,7%	-0,6%	18,7%	3,1%	20,6%	35,3%	4,2%	12,3%	25,1%	13,9%
170 kph	-0,6%	12,2%	11,5%	24,7%	14,3%	28,6%	-0,4%	7,8%	20,6%	13,2%
180 kph	5,5%	5,2%	4,9%	17,6%	8,3%	22,4%	37,3%	3,4%	16,4%	13,4%
190 kph	-1,4%	1,3%	-1,3%	10,9%	2,6%	16,4%	31,1%	-0,9%	12,2%	7,6%
200 kph	3,4%	8,4%	13,0%	4,9%	25,8%	10,8%	25,3%	37,7%	8,1%	15,3%
Average	7,9%	10,7%	12,2%	16,8%	18,6%	22,4%	27,3%	29,6%	38,1%	20,4%

Homogeneous operation – 7 seconds system delay

Speed/Block length	500 m	750 m	1000 m	1200 m	1600 m	2000 m	2500 m	3000 m	4000 m	Average
60 kph	21,7%	5,2%	16,5%	23,8%	35,1%	43,5%	51,4%	57,3%	65,7%	35,6%
70 kph	12,0%	-2,0%	9,3%	16,7%	28,4%	37,2%	45,6%	52,0%	61,2%	28,9%
80 kph	3,4%	26,2%	2,7%	10,2%	22,1%	31,3%	40,0%	46,8%	56,7%	26,6%
90 kph	-4,1%	17,8%	-3,3%	4,1%	16,2%	25,5%	34,7%	41,8%	52,2%	20,5%
100 kph	11,3%	10,2%	27,4%	-1,5%	10,6%	20,1%	29,5%	36,9%	47,9%	21,4%
110 kph	3,6%	3,2%	19,9%	31,4%	5,3%	14,9%	24,5%	32,1%	43,5%	19,8%
120 kph	-3,3%	-3,0%	13,1%	24,3%	0,3%	10,0%	19,7%	27,5%	39,3%	14,2%
130 kph	8,1%	15,5%	6,9%	17,8%	-4,3%	5,3%	15,1%	23,0%	35,2%	13,6%
140 kph	0,6%	8,1%	0,5%	11,1%	29,3%	0,4%	10,3%	18,3%	30,7%	12,2%
150 kph	-6,5%	1,0%	-5,6%	4,6%	22,3%	-4,5%	5,4%	13,5%	26,2%	6,3%
160 kph	1,0%	-5,6%	13,1%	-1,5%	15,7%	30,2%	0,6%	8,8%	21,6%	9,3%
170 kph	-5,8%	6,6%	6,3%	19,1%	9,5%	23,7%	-3,9%	4,3%	17,3%	8,6%
180 kph	0,0%	0,0%	0,0%	12,3%	3,8%	17,7%	32,5%	0,0%	13,1%	8,8%
190 kph	-6,5%	-6,1%	-5,8%	6,0%	-1,6%	11,9%	26,4%	-4,1%	8,9%	3,2%
200 kph	-1,8%	3,2%	7,8%	0,2%	20,6%	6,6%	20,8%	33,1%	5,0%	10,6%
Average	2,2%	5,4%	7,3%	11,9%	14,2%	18,3%	23,5%	26,1%	35,0%	16,0%

Heterogeneous operation – 2.65 seconds system delay

Speed/Block length	500 m	750 m	1000 m	1200 m	1600 m	2000 m	2500 m	3000 m	4000 m	Average
60 kph	15,2%	6,8%	20,3%	29,3%	21,6%	30,3%	38,7%	45,3%	55,0%	29,2%
70 kph	13,9%	10,5%	11,2%	20,1%	34,8%	23,2%	32,0%	39,0%	49,4%	26,0%
80 kph	4,0%	18,4%	3,1%	11,8%	26,4%	16,6%	25,7%	32,9%	43,8%	20,3%
90 kph	2,4%	19,3%	8,2%	4,3%	18,6%	30,3%	19,6%	27,0%	38,4%	18,7%
100 kph	10,0%	15,7%	23,5%	1,6%	16,0%	27,7%	17,2%	24,8%	36,4%	19,2%
110 kph	6,3%	12,4%	19,9%	18,0%	13,4%	25,3%	14,8%	22,5%	34,3%	18,5%
120 kph	2,9%	9,3%	16,7%	14,6%	11,1%	22,9%	12,5%	20,3%	32,4%	15,9%
130 kph	8,6%	18,6%	13,6%	11,5%	8,8%	20,7%	10,3%	18,2%	30,4%	15,6%
140 kph	4,9%	15,0%	10,5%	8,2%	25,6%	18,3%	8,0%	15,9%	28,3%	14,9%
150 kph	1,3%	11,4%	7,4%	4,9%	22,2%	15,8%	5,6%	13,5%	26,0%	12,0%
160 kph	5,0%	8,0%	16,8%	1,8%	18,9%	33,3%	3,2%	11,2%	23,8%	13,6%
170 kph	1,6%	14,2%	13,3%	12,3%	15,8%	30,1%	1,0%	9,0%	21,7%	13,2%
180 kph	4,5%	10,8%	10,2%	8,8%	13,0%	27,0%	19,3%	6,9%	19,6%	13,3%
190 kph	1,2%	7,7%	7,2%	5,6%	10,2%	24,2%	16,3%	4,8%	17,5%	10,5%
200 kph	3,5%	12,4%	14,1%	2,7%	21,5%	21,5%	13,5%	23,6%	15,6%	14,3%
Average	5,7%	12,7%	13,1%	10,4%	18,5%	24,5%	15,8%	21,0%	31,5%	17,0%

Heterogeneous operation – 7 seconds system delay

Speed/Block length	500 m	750 m	1000 m	1200 m	1600 m	2000 m	2500 m	3000 m	4000 m	Average
60 kph	11,1%	3,5%	16,9%	26,0%	18,8%	27,6%	36,3%	43,0%	53,0%	26,2%
70 kph	9,7%	6,9%	7,9%	16,8%	31,5%	20,5%	29,4%	36,6%	47,2%	22,9%
80 kph	-0,1%	14,2%	-0,2%	8,5%	23,1%	13,8%	23,0%	30,4%	41,5%	17,1%
90 kph	-1,6%	15,0%	4,7%	1,0%	15,3%	27,0%	16,8%	24,4%	36,0%	15,4%
100 kph	5,6%	11,5%	19,2%	-1,6%	12,6%	24,4%	14,4%	22,1%	33,9%	15,8%
110 kph	2,0%	8,2%	15,7%	14,0%	10,1%	21,9%	12,0%	19,8%	31,8%	15,1%
120 kph	-1,3%	5,3%	12,5%	10,7%	7,8%	19,6%	9,7%	17,6%	29,8%	12,4%
130 kph	4,1%	14,1%	9,6%	7,6%	5,5%	17,3%	7,5%	15,4%	27,8%	12,1%
140 kph	0,6%	10,6%	6,5%	4,4%	21,7%	15,0%	5,1%	13,1%	25,6%	11,4%
150 kph	-2,8%	7,2%	3,6%	1,3%	18,3%	12,6%	2,8%	10,8%	23,4%	8,6%
160 kph	0,7%	4,0%	12,6%	-1,7%	15,2%	29,4%	0,5%	8,5%	21,2%	10,0%
170 kph	-2,5%	9,9%	9,3%	8,3%	12,2%	26,3%	-1,7%	6,3%	19,1%	9,7%
180 kph	0,3%	6,7%	6,3%	5,0%	9,4%	23,4%	16,0%	4,2%	17,0%	9,8%
190 kph	-2,8%	3,7%	3,5%	2,0%	6,8%	20,6%	13,1%	2,2%	15,0%	7,1%
200 kph	-0,6%	8,3%	10,1%	-0,8%	17,6%	18,0%	10,3%	20,4%	13,1%	10,7%
Average	1,5%	8,6%	9,2%	6,8%	15,1%	21,2%	13,0%	18,3%	29,0%	13,6%

Figure 8: Capacity improvement for level 2 compared to level 1 with system delay of 2.65 seconds (left) and 7 seconds (right) for different block lengths and line speeds, for homogeneous operation (top) and heterogenous operation (bottom).

Figure 8 shows higher system delays reduce the capacity gain of level 2 compared to level 1, and higher system delays have significant impact on short block sections and high speed. This is because the system delay has higher impact when trains occupy the block sections for shorter time.

4.2 Practical Results

Applying the headway calculation model on real-life railway lines and timetables with freight, IC and fast trains, it is seen in Table 7 that the decrease in capacity consumption for level 2 vs level 1 is limited to 1-10% with an average of 3%.

Table 7: Capacity consumption for different line sections for level 1 and 2.

	Level 1	**Level 2**	**Difference**
Line section 1	45%	44%	1%
Line section 2	46%	43%	3%
Line section 3	59%	49%	10%
Line section 4	50%	47%	3%
Line section 5	41%	39%	2%
Line section 6	56%	54%	2%
Line section 7	60%	59%	1%
Line section 8	61%	59%	2%
Line section 9	43%	42%	1%
Average	51%	48%	3%

The results in Table 7 are for different line sections on two different railway lines that have had a 1:1 replacement of conventional signals with ERTMS. The reasons for the limited capacity gain for level 2 compared to level 1 are short block sections on the line sections and high degree of heterogeneity with less potential for improving the headways, cf. Figure 9.

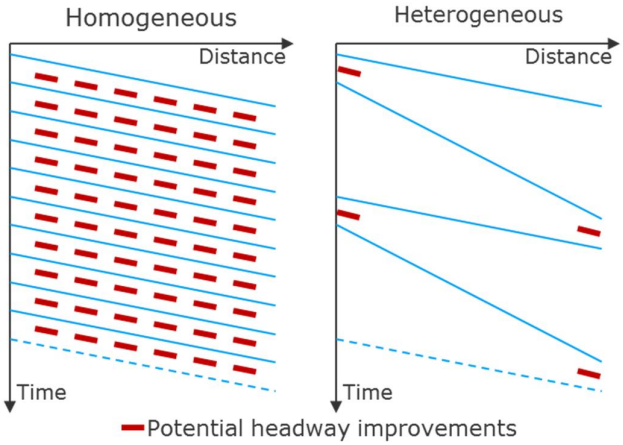


Figure 9: Potential for headway (block occupation) improvements for homogeneous and heterogeneous operation.

As an example, for line section 4, homogeneous operation would have yield 18%, 14%, 45%, and 12% for operation with purely Freight, IC, Fast (Lambda) and Fast (Gamma) respectively, cf. Table 8.

Table 8: Improvement in line headway for a real-life case example for level 2 vs level 1 for line section 4 (cases with homogeneous operation marked with bold).

1st train \ 2nd train	Freight	IC	Fast (Lambda)	Fast (Gamma)
Freight	**18%**	3%	1%	2%
IC	19%	**14%**	24%	8%
Fast (Lambda)	27%	26%	**45%**	12%
Fast (Gamma)	27%	26%	45%	**12%**

Two different ETCS braking curve calculation methods exist, lambda and gamma, as described in section 2.3. The lambda calculation is used when nominal brake performance data is not available for the more detailed gamma calculation. However, as seen in Table 8 for the fast train, the choice of braking curve model can have significant influence on the headways and hence capacity. In Table 8, the Fast train has well-adapted braking distance for the block lengths with the gamma braking curve. However, if the braking curves is calculated with the lambda braking curve model, a more significant capacity improvement from level 1 to level 2 would have been observed for this specific train type. This is due to slightly longer braking curves calculated using the lambda model resulting in reservation of an extra block section.

The higher capacity gain from level 1 to level 2 for the lambda braking curves thereby illustrate the effect of the continuous update of the Movement Authority (MA) in level 2 in the current case.

5 Discussion

The theoretical results have shown potential for large capacity gains for level 2 compared to level 1, especially for infrastructures with long block sections and low speed. For homogenous operation, we have observed moderate capacity gains in our real-life case examples when comparing level 1 and level 2. However, only small realized gains in capacity from level 1 to 2 have been identified in our real-life case examples for heterogeneous operation. It is therefore relevant to examine solutions for capacity improvements for different line sections e.g. by adding infill (loops, radio and/or balises) to level 1.

This article has focused on the differences between ERTMS level 1 and 2 systems. However, when changing from conventional signaling to ERTMS, it is essential to examine the potential loss in capacity when the braking curve calculations change which generally lead to longer braking curves and hence loss in capacity. Here it is important to choose the right national values for the braking curves to ensure as high capacity as possible with ERTMS – or limit the capacity loss converting to ERTMS. An example of an important national value is the emergency brake confidence level used in the gamma calculation – both important for infrastructure capacity and safety. Another parameter that greatly affects the braking in ERTMS, and thus the capacity, is the use of the service braking interface. Not using this interface improves capacity as braking intervention is initiated later, although this is not advisable as describe in Section 2.3.

When deciding on the ERTMS architecture, 1:1 replacement of the conventional signaling or overlay may not be options due to loss of capacity. For heavily utilized line sections, ERTMS' advantages over older signaling systems can be used. This is especially the possibility of shorter and more flexible block sections as ERTMS generally can look

further ahead and optical signals are not needed. This both allow shorter headways and possibility of higher speed for the (freight) trains with reduced braking capabilities.

ERTMS can potentially be used with Automatic Train Operation which can reduce operation cost for the TOC and lead to increased infrastructure capacity. The extra capacity is achieved by a more uniform driving behavior, including braking, making it possible to optimize block lengths and driving behavior.

6 Conclusions

This article has described the main differences between ERTMS level 1-3 and conventional signaling systems. Based on this description, the capacity differences between level 1 and 2 have been investigated for theoretical as well as real-life cases using a line headway calculation model developed for the study.

The results illustrate that ERTMS level 2 generally has shorter headways than level 1 and hence higher capacity. In homogeneous operation where the braking distance is well-adapted to the block lengths, level 1 can have shorter headways than level 2 due to less system delays.

Level 2 has the highest capacity gains over level 1 for longer block sections and lower speeds. This is because the continuous update of level 2 has more effect when trains occupy the block sections for longer time. Heterogeneous operation generally reduces the capacity gain for level 2 compared to level 1 in case of long block sections and low speed while the capacity gain for homogeneous operation is increased for short block sections and high speed as the disadvantage of discrete update of the Movement Authority (MA) in level 1 thereby is reduced.

1:1 replacement of conventional signaling to ERTMS can lead to loss of capacity as the braking curves are likely to be longer for ERTMS why longer headways occur. However, extra capacity can be gained with ERTMS as it is possible to look more block sections ahead resulting in shorter and more flexible block sections and potentially higher speed for (freight) trains with reduced braking capabilities.

References

Abril, M., Barber, F., Ingolotti, L., Salido, M.A., Tormos, P., Lova, A., 2008. “An assessment of railway capacity”. *Transportation Research: Part E.*

ERA, 2016. *Introduction to ETCS Braking curves.*

ERA UNISIG, 2016. *SUBSET-026-3: System Requirements Specification*.

ERA, 2018. *ERA Braking curves tool handbook*.

Goverde, R.M.P., Corman, F., D’Adriano, A., 2013. “Railway line capacity consumption of different railway signalling systems under scheduled and disturbed conditions”, *Journal of Rail Transport Planning & Management.*

Happel, O., 1959. “Sperrzeiten alse Grundlage der Fahrplankonstruktionen”, *Eisenbahntechnische Rundschau*, 8(H.2), pp. 79-80.

Jensen, L.W., 2015. *Robustness indicators and capacity models for railway networks*, PhD Thesis, Technical University of Denmark, Lyngby, Denmark.

Liikenvirasto, 2018. ERTMS/ETCS-tason 2 kapasiteettihyödyt kaksiraiteisilla radoilla”, *www.liikenvirasto.fi*

Meyer,P., Chavagnat, R., Bourgeteau, F., 2011. *Computation of the Safe Emergency*

Braking Deceleration for Trains Operated by ETCS/ERTMS using the Monte Carlo Statistical Approach. Proceedings in: the 9th World Congress on Railway Research (WCRR). Lille.
Pachl, J., 2008. "Timetable Design Principles", Railway Timetable & Traffic, pp. 9-42, EurailPress.
UIC, 2000. Leaflet 451-1: Timetable recovery margins to guarantee timekeeping - Recovery margins.
UIC, 2008. *Influence of ETCS on line capacity – Generic study*.
UNIFE, 2014. "Increasing infrastructure capacity - How ERTMS improves railway performance (factsheet)", *www.ertms.org*
UNISIG, 2015. *SUBSET-041: Performance Requirements for Interoperability*.

Taking Driver Advisory Systems to the next level

Per Leander [1)], Andreas Törnblom
Transrail Sweden AB
Västmannagatan 3, 111 24 Stockholm, Sweden
[1] E-mail: per.leander@transrail.se, Phone: +46 (0) 70 511 80 69

Abstract
There is constantly increasing pressure on railways globally to provide greater capacity and improved service performance, whilst reducing investment, operational and energy costs. This drives demand for improved traffic management and train operation systems.

Driver Advisory Systems (DAS) is a fairly new technology within railway/metro operations where Transrail Sweden AB has developed and markets a product called CATO. The technology gives very strong support to increased punctuality, increased traffic capacity and reduced operational cost (e.g. reduced energy consumption). In fact, the system can be used for Intelligent Cruise Control and for highly efficient ATO operation, better than the systems found on the market today.

This paper brings a description/outline of the technology and development trends beyond current standalone Driver Advisory Systems, i.e. C-DAS, Intelligent Cruise Control (ICC) and usage of the technology for ATO.

Keywords
Driver Advisory System, C-DAS, Intelligent Cruise Control, ATO, Sustainability

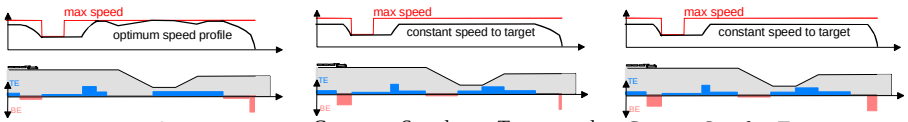
1 Problem and Objectives

There is constantly increasing pressure on railways globally to provide greater capacity and improved service performance, whilst reducing investment, operational and energy costs. This drives demand for improved traffic management and train operation systems.

Driver Advisory Systems (DAS) are finding favour around the world as a means of optimising the performance of individual trains to reduce energy consumption while ensuring close adherence to the timetable.

The definition of a DAS system: A system that assists a train driver to drive on time and with an economic driving style.

The performances vary between DAS products and depend on the strategies on which they build as well as how they are implemented. For example; if you know the distance to the next stop and the arrival time, a simple strategy would be to calculate the constant speed to arrive on time. You may understand that the issue of efficient driving is far more complex than using this simple strategy. The driving profile shall for example be calculated depending on the track profile (speed limits, gradients, curves), the train (weight, length, motoring/braking performance etc) as well as possible timing restrictions along the journey. Some alternative DAS strategies are presented in Figure 1. The Optimal Speed Profile (CATO300) makes full use of a train's character as a roller coaster.



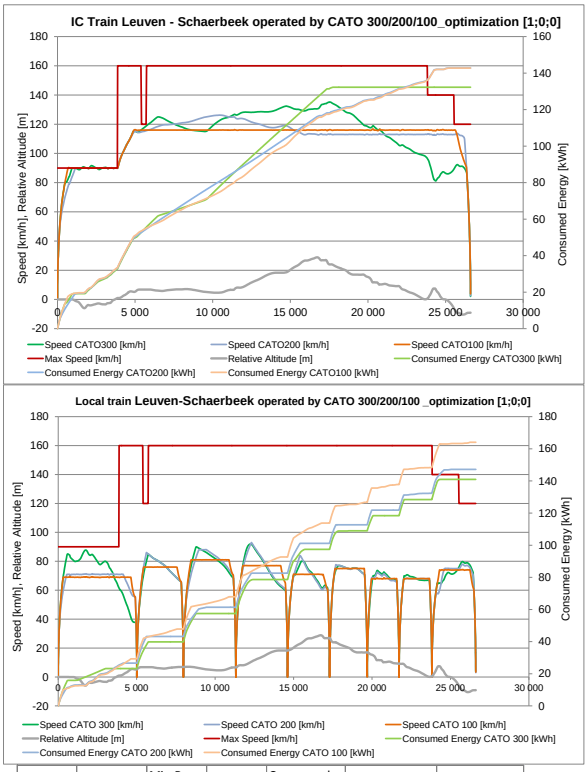
Optimal Speed Profile (CATO300)

Constant Speed to Target and coasting (CATO200)

Constant Speed to Target (CATO100)

Figure 1: Typical DAS strategies, also available as CATO DAS versions

The difference between the strategies might not appear very substantial, but Figure 2 and Figure 3 illustrate their performances as regards energy savings. Their improvements as regards reduced friction brake energy are even larger.



Train type	CATO version	Min Run Time [min]	Run Time [Min]	Consumed Energy [kWh]	Dynamic Brake Energy [kWh]	Friction Brake Energy [kWh]
IC train	CATO300	12,4	15,0	132,3	7,3	7,3
	CATO200			142,8	9,1	11,4
	CATO100			142,8	12,9	7,8
Local train	CATO300	20,5	25,6	141,0	17,4	39,1
	CATO200			147,2	20,6	40,6
	CATO100			164,1	33,2	41,0

Figure 2: Example of performances of various DAS strategies in IC and Local train operation over a line.

Figure 3 shows the result of a CATO300 Benefit Analysis for the Stockholm commuter train operation.

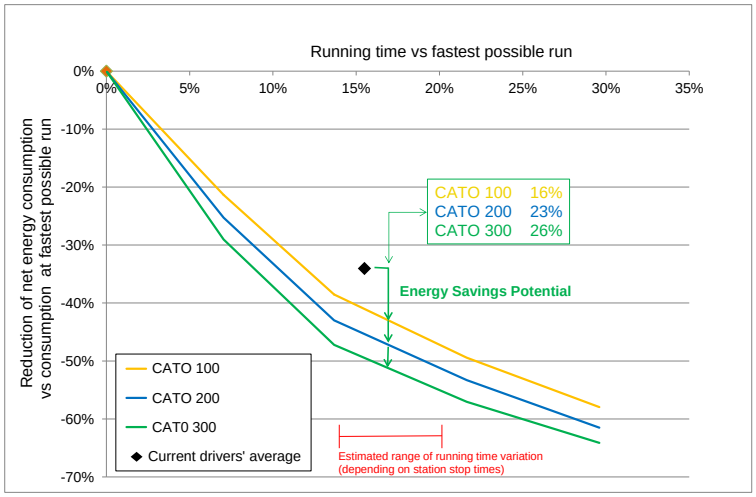
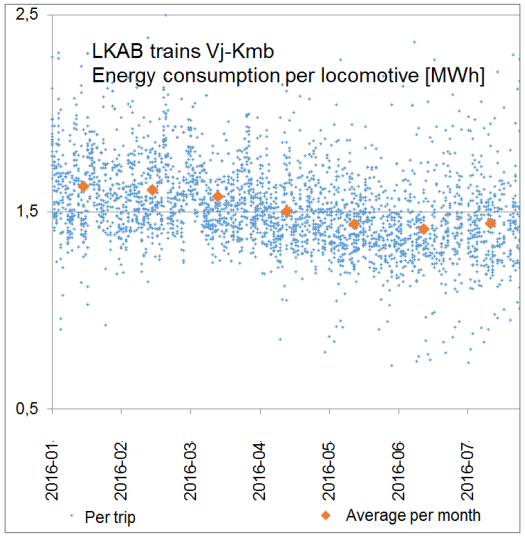


Figure 3: Performance Diagram showing the saving potentials of the various CATO versions in Stockholm commuter traffic. The X-axis corresponds to "slack" e.g. runtime above minimum runtime (MRT). The Y-axis shows energy savings. With more slack increased savings can be achieved. Drivers' average is the black square.

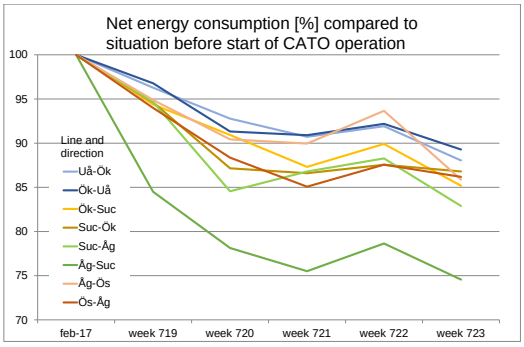
CATO makes use of an optimisation based on achieving the targeted arrival time and minimising a cost function, which for example may be defined according to Equation 1: Example of CATO cost function used for the optimisation algorithm: The cost function may include any variables and weights which may be changed at any time.

$$C = 1 \times Consumed\ Energy\ \ kWh\ - 0{,}5\ \times RegeneratedEnergy\ \ kWh\ + 4\ \times MechBrake\ Energy\ [kWh]$$

Equation 1: Example of CATO cost function used for the optimisation algorithm



LKAB iron ore trains operated in Northern Sweden,
Line Vj-Kmb (C-DAS). CATO implemented in February/March.



Tågkompaniet regional trains (S-DAS) Average/week on various lines

Figure 4: Some examples of results when introducing CATO300 based on energy consumption measured by the onboard energy meters, illustrating energy before implementation of CATO300 and the performance shortly afterwards.

A further effect brought by the CATO optimisation is shown in Figure 5; Example of how energy consumption is divided into different components. In this case, the train will gain potential energy (marked gray), which can be recovered run in the opposite direction. Energy for heating, ventilation and auxiliary power has not been included.. The bar graphs show that both the gross energy, i.e. the energy drawn from the power supply, and the net energy are reduced. It is interesting, albeit natural, to note that the optimization selects a driving profile that minimizes the use of friction brakes as well as the use of regenerative brakes. Coasting is preferred whenever possible depending on the available journey time. The decrease in gross and regenerated energy means that the catenary power load is reduced.

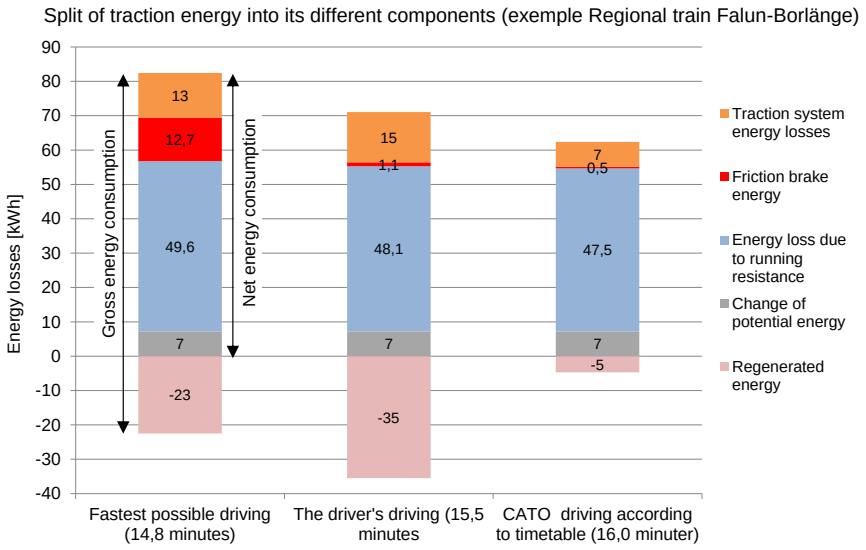


Figure 5; Example of how energy consumption is divided into different components. In this case, the train will gain potential energy (marked gray), which can be recovered run in the opposite direction. Energy for heating, ventilation and auxiliary power has not been included.

2 Connected (C-DAS) and Semi-Connected (SC-DAS) Solutions

Most DAS systems currently in operation are stand-alone (S-DAS) technologies, using the planned timetables for timing of the advice, and a strategy comparable to CATO100.

Technology like this can be deployed by the operator without the need to establish a real-time interface with the TMS, nor does it need advanced TMS functionality, but information about the actual real traffic situation, and its needs according to the real time traffic plan, is lost.

2.1 C-DAS

Connected Driver Advisory System (C-DAS) is conceptually an IM-RU system where data from defined master data sources may be changed dynamically during the journey. IM (the Infrastructure Manager) defines timing requirements, as regards schedule and adherence. RU (the Railway Undertaking) defines the economic train driving style within the limits of the timing requirements.

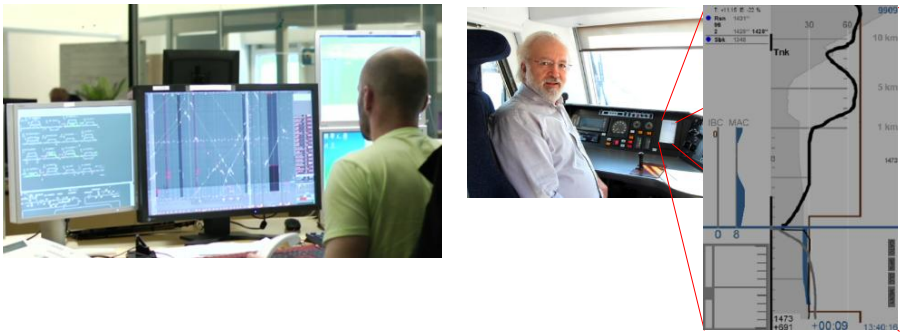


Figure 6: To the left, an example of a TMS system connected to CATO for dynamic submission of information on the real time traffic plan, route, speed restrictions, track possessions etc. To the right, the Cato Driver Machine Interface (CDMI) with advice on the optimal speed profile and information on the current schedule as well as other trains on the line..

It is obvious that C-DAS takes system optimisation a step further by providing a communications link between the DAS and the Traffic Management System (TMS). In fact C-DAS could be seen as consisting of two subsystems, a DAS and a TMS.

The TMS defines real time timing requirements with regards to scheduling and adherence to the timetable, while the DAS defines the optimum driving style within the limits of the timing requirements. Scheduling, routing and speed restriction updates are communicated to the train in real time, while information from the train enhances traffic regulation decisions at the TMS end. The TMS need to focus on steering the traffic in terms of timings and not only controlling their routes.

Simplifying the complex railway operational environment, there are three main components: rolling stock, infrastructure and operating rules (timetables). A flexible railway is simply the ability to implement ad-hoc timetable changes, as infrastructure and rolling stock are essentially fixed parameters. It is this "enabler" function of C-DAS that delivers the flexibility to change the timetable according to Traffic Management (TM) needs, i.e. enabling both 'communication' and 'active correction' to the driver. DAS is an enabler of TM as it provides the medium upon which TM decisions can be communicated to the drivers, as well as providing the channel for the increased resolution of train location and timing necessary for TM to make better decisions. Transrail has introduced the term Train Tango for the C-DAS operation connecting TMS and DAS, dispatchers and drivers. Operational flexibility by Train Tango is an enabler to efficiently solve situations of traffic disturbances and to increase traffic capacity on the railway.

Integrating driver advisory systems with traffic management systems unlocks a number of opportunities for optimizing operational efficiency, but making this connection is not without its complications. There are so far only few C-DAS systems on the market worldwide, mainly due to the current TMS products and their inability to efficiently support a DAS. Still, the advantages are obvious and these solutions will evolve. Transrail's LKAB implementation is an early example of C-DAS.

2.2 SC-DAS

Semi-Connected DAS is conceptually a "C-DAS" but with connection only to the

signaling system. This is a solution developed by Transrail for situations where legacy TMS systems cannot handle C-DAS.

SC-DAS may be used to calculate the DAS advice not only from the timetable, but also based on the motion of other trains on the line. The solution is very strong when trains are running after each other on a line. The trains can be run on a green wave with minimum headways. It is also useful if there are trains on the line, which are not fitted with a C- or SC-DAS.

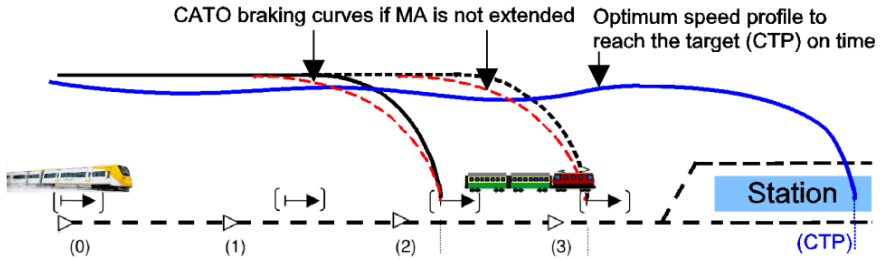


Figure 7: CATO optimal speed profile (CMP) and ATP movement authorities (MA). The optimal driving profile is calculated based on forecasts of the time when the MA will be extended due to the motion of trains ahead. The figure also explains the fundamental difference between an ATP system and DAS. ATP informs on the current status of the interlocking system (the current MAs). The DAS driving profile needs to predict the time when the MAs will be extended.

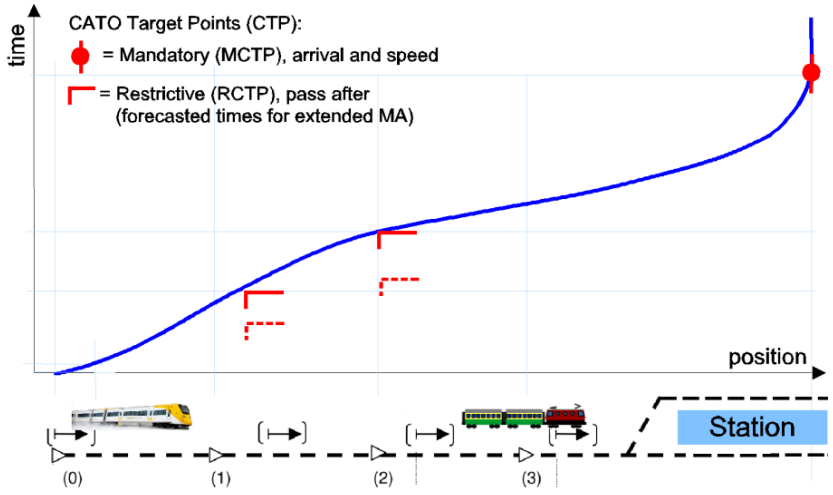


Figure 8: Situation as in Figure 7with delayed extensions of movement authorities and described in a distance-time diagram.

3 Intelligent Cruise Control (ICC) and Automatic Train Operation (ATO)

One of many aspects in the deployment of a DAS is the need to consider human factors in order to successfully harness the full potential of the technology. DAS is only a tool, so the drivers need to be able to use it correctly to get the desired results. A good C-DAS system relies on the driver's ability to read and respond to the advice it generates. An example of an early CATO Driver Machine Interface is shown in Figure 6. The CATO solution has proved that drivers can easily drive very heavy trains within seconds according to the traffic plan and even on difficult line profiles.

The intricacies of DMI design will become less important as C-DAS technologies move towards the integration of cruise functionality and ATO.

3.1 Intelligent Cruise Control (ICC), CATO Cruise

Many locomotives and multiple units of today have constant speed cruise control, very similar to what is available on ordinary cars. The next step is to use C- and SC-DAS to make the cruise intelligent. The driver presses a button and the train moves forwards in accordance with the CATO algorithm. Intelligent cruise is going to be a big thing for operators and rolling stock suppliers and it will be really popular with drivers. All the merits of C-DAS can be incorporated in the ICC.

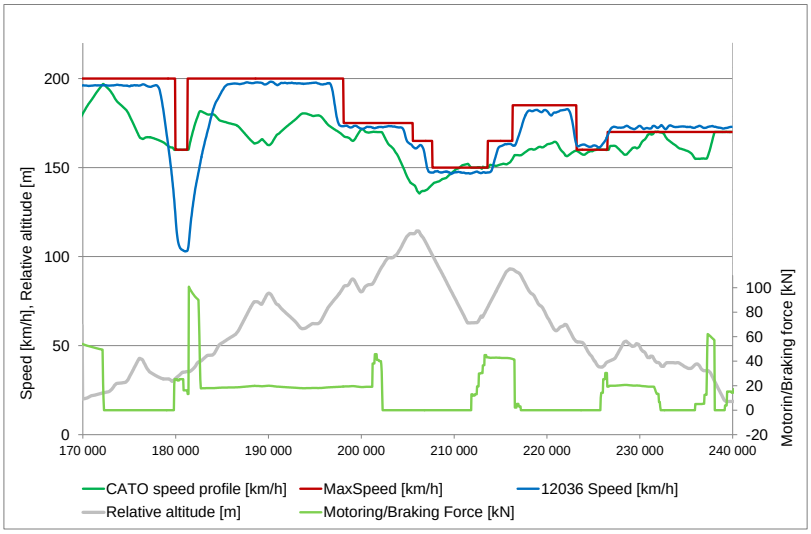


Figure 9; Example of CATO speed profile compared to a typical train run using conventional constant speed cruise control. 12036 speed is the conventional constant cruise speed control used by the driver.

It is not yet demonstrated, but we are confident that the CATO solution can efficiently drive any train from one stop to the next, from one platform to the next, and this can be

done with a performance that will surpass what most drivers can achieve with a DAS.

3.2 Automatic Train Operation (ATO)

In the long perspective, ICC will lead to efficient ATO (Automatic Train Operation). CATO Cruise can be used already today as GOA2 ATO for the line haul.

Many studies have been done on ATO but often the algorithms that govern traction and braking are rudimentary, so energy usage can actually go up in ATO mode. What DAS and CATO has done so far is to enrich the technology with solutions for optimised driving profiles. ATO will always need intelligent algorithms on the train that can optimise conditions with dynamic data.

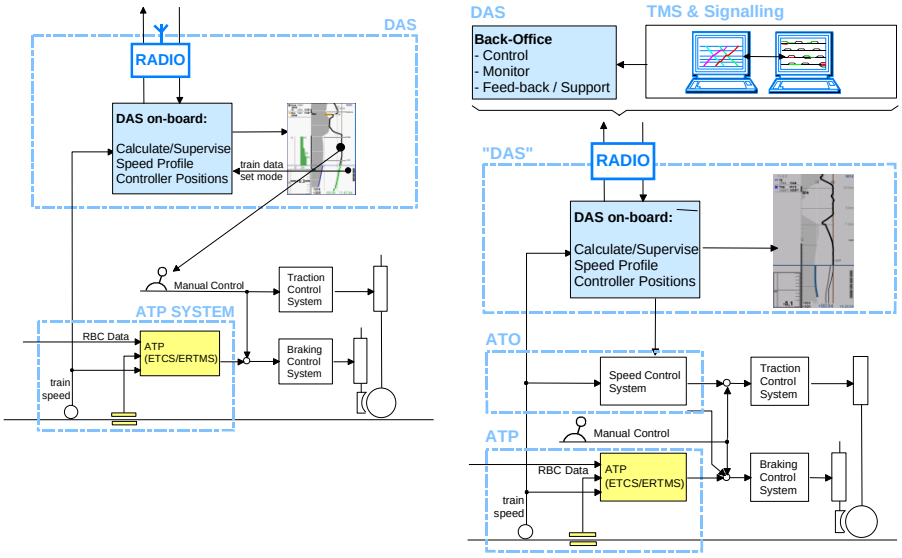


Figure 10; To the left, the general on-board architecture of DAS and ATP systems. Manual driving. To the right, the principle system architecture for an ICC or an ATO system using CATO as the "ATO engine".

Modelling the Prohibition of Train Crossings in Tunnels with Blocking Time Theory

Wiebke Lenze [a,1], Nils Nießen [a]
[a] Institute of Transport Science, RWTH Aachen University
Mies-van-der-Rohe-Straße 1, 52074 Aachen, Germany
[1] E-mail: lenze@via.rwth-aachen.de, Phone: +49 (0) 241 80 25657

Abstract

Preventing both passenger and freight trains from crossing each other in double-track railway tunnels is a fire safety measure required by the German railway authority in order to prevent fatal accidents. The prohibition poses a restriction on infrastructure usage that has to be incorporated in rail traffic planning. While it has already been implemented in timetabling and simulation tools, its effects on line capacity in long-term strategic planning has not been investigated so far. This paper presents a method to incorporate restrictions on simultaneous track usage in the blocking time calculation and minimum headway time estimation. The effects on line capacity are analysed quantitatively based on the STRELE approach, which is an analytical method for strategic long-term capacity planning currently used by German railway infrastructure manager DB Netz AG. Results are validated by comparison to delay increase in microscopic simulation of train operations.

Keywords

Analytical methods, Capacity, Safety, Train crossings, Tunnels

1 Introduction

Tunnels are critical elements for safe operation of rail traffic. Even though accidents occur less often inside railway tunnels, the damage caused by fire in such a closed environment with limited accessibility can be catastrophic. Especially trains carrying dangerous goods pose a fire hazard and should not be scheduled to cross oncoming passenger trains in order to avoid fatal fires (UIC (2003)). For German infrastructure, the federal railway authority Eisenbahn-Bundesamt (2008) prohibits all freight trains from crossing passenger trains in new tunnels that are longer than 500 meters.

Timetabling and simulation tools used for the German market such as RUT-K (Brünger and Gröger (2003)), LUKS® (Janecek and Weymann (2010)) or RailSys® (Radtke and Bendfeldt (2001)) need to incorporate the prohibition of passenger and freight trains from crossing each other in tunnels. In simulations with the software LUKS® both directions of a double-track railway line are evaluated simultaneously and restrictions on simultaneous track usage in tunnels is already implemented. Whenever a prioritized passenger train passes through the tunnel, it occupies the other direction for freight trains. Freight trains need to wait in front of the entry signal for the passenger train to leave the tunnel.

The effects on line capacity in long-term strategic planning have not been investigated so far. UIC (2004) defines capacity as the number of trains in a fixed time period, which can be operated with market-orientated quality. Evaluating the existing and future capacity is necessary for the recognition of bottlenecks. It is essential to make optimal use of the rail network and to expand the infrastructure where necessary in order to meet the constantly

increasing demand for transportation.

This paper presents a method to include restrictions on simultaneous track usage in the blocking time calculation. Mean obstructions caused by the prohibition of train crossings in tunnels are characterized by extended blocking times. Blocking times are required to calculate minimum headway times, which are important input parameters for long-term capacity planning. To assess the effects quantitatively, the modelling is included in the blocking time based STRELE formula by Schwanhäußer (1974). This method is a strategic planning framework based on stochastic prognosis of knock-on delays and is the standard method for capacity planning of railway lines used by German infrastructure manager DB Netz AG (DB Netz AG (2009)).

The following chapter 2 gives a detailed overview of existing methods to evaluate capacity. Chapter 3 presents the new method to modify blocking times. This method is applied for capacity analysis with the STRELE approach in chapter 4 with results being validated by comparison to delay increase in microscopic simulation of train operations.

2 Capacity Assessment

This chapter gives an overview about different methods for capacity assessment such as simulations or analytical approaches. Essential for these methods is a basic knowledge about the blocking time theory, which is provided in advance.

2.1 Fundamentals

The infrastructure occupancy can be described based on blocking times (Happel (1959), UIC (2013)). The train's operational occupancy of a section takes longer than the purely physical occupancy. Before the train runs through a section, it is already blocked for the route setup time t_{setup}, the signal watching time t_{sight} and the approach time $t_{approach}$. After the actual running time $t_{running}$, the clearing time $t_{clearing}$ and the release time $t_{release}$ block the section before the next train movement can occupy it (Pachl (2014)). The sum of these time elements represents the entire blocking time, which is illustrated in Figure 1. Blocking time theory can even be applied for different train control and signalling systems, such as ETCS (Wendler (2009)).

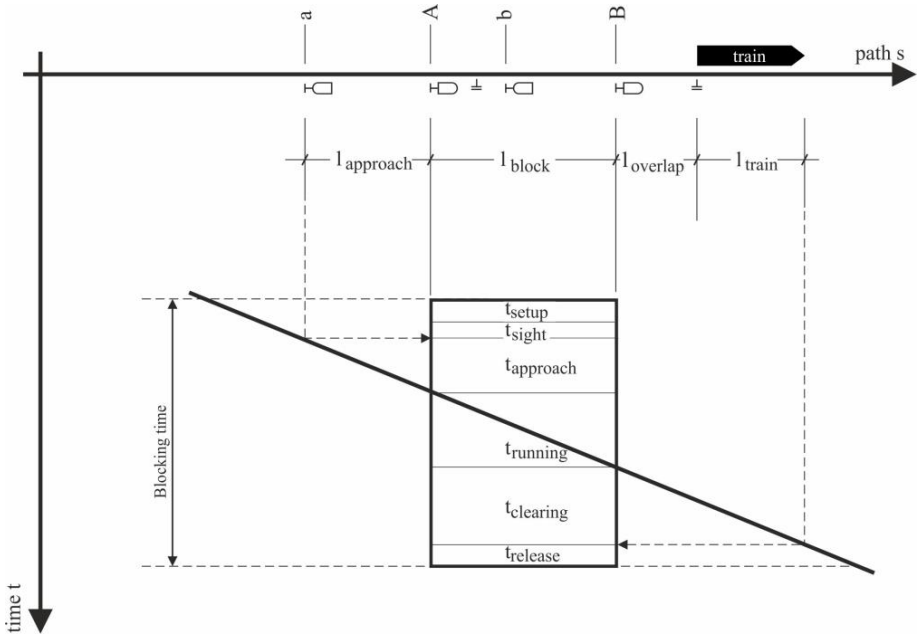


Figure 1: Blocking time and its elements

The graphic sequence of blocking times forms a blocking time stairway. Blocking time stairways of two trains demonstrate the minimum temporal distance in which they can follow each other free of obstruction. This duration is called minimum headway time and is measured for each overtaking section. The minimum headway time starts at the beginning of the blocking time of the preceding train and ends at the beginning of the blocking time of the subsequent train (see Figure 2). Minimum headway times refer to the common itinerary on an overtaking section of two trains. The overtaking section with the largest minimum headway time is decisive for the entire track (Nießen (2014)).

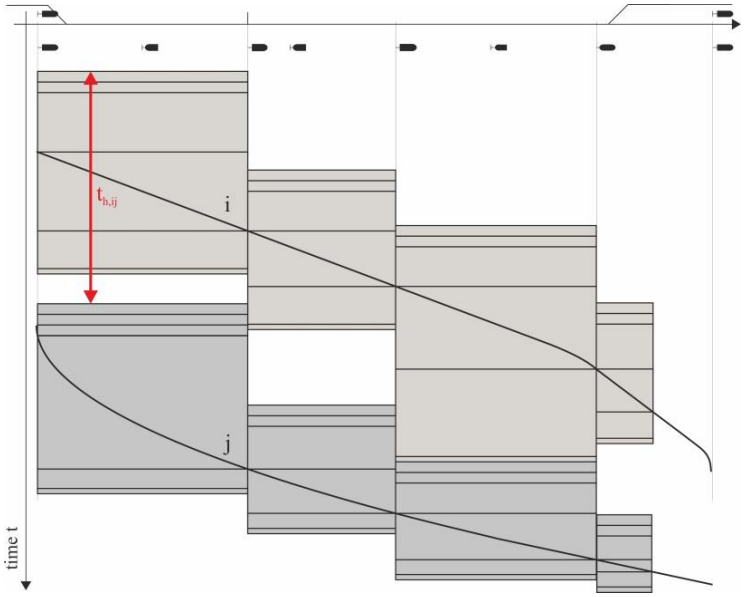


Figure 2: Minimum headway time $t_{h,ij}$

In scheduled timetables, running time supplements and buffer times are added to absorb smaller train delays. A train delay is a deviation from the timetable. According to the location and cause of generation, delays can be classified: Primary delays are not caused by other trains but are due to disruptions such as technical failures, large passenger volumes or bad weather conditions. Following these primary delays, a delayed train might hinder other trains and cause so-called knock-on delays (Yuan and Hansen (2007)).

2.2 Simulations

A simulation imitates the real operation process in a way that a given timetable is perturbed randomly by primary delays in many different runs and models the resulting propagation of train delays in the railway network. It is possible to include special characteristics of the infrastructure or the operating program. Thus, simulations are especially suitable for complex track layouts and timetables, which are known in detail. Modelling the infrastructure and timetable with the simulation tool requires extensive work. The results, such as delay developments and punctuality, are only valid for the examined timetable. Calculating general performance indicators is only possible by iteratively simulating a large number of different timetables (Watson and Medeossi (2014)).

Microscopic simulation models are generally divided into synchronous and asynchronous models. In synchronous simulations, all trains are modelled simultaneously. The operation process is reproduced in time steps and in each time step, concurrent occupations are resolved under consideration of priorities. Synchronous models allow a realistic representation of train traffic with all trains interacting between each other (Jacobs (2008), DB Netz AG (2009)).

Asynchronous simulations perform within a strictly descending hierarchical structure:

Trains are modelled ordered by their priority. First, trains with the highest priority are modelled and occurring conflicts are solved with a "first come, first serve" strategy. Resulting infrastructure occupations are fixed and stored. After that, the next priority group of trains is added to the time-distance-diagram and simulated in the same way (Watson and Medeossi (2014)).

2.3 Analytical Capacity Assessment

For the management and operation of railway systems, it is extremely important to evaluate the capacity of railway infrastructure. This knowledge constitutes the necessary basis to decide which measure – changing the infrastructure or its usage – is most effective to satisfy a growing traffic demand.

In Europe, the timetable compression method proposed by the International Union of Railways (UIC (2013)) is common to evaluate the capacity of a railway line. This method is based on the blocking time theory. Compressed blocking time stairways use the infrastructure during the occupancy time. The so-called concatenated occupancy rate then results from the ratio of the occupancy time to the investigation period. UIC (2013) recommends values for occupancy time rates for three different types of lines. Adding extra trains until the recommended occupancy time rate is reached leads to the line's capacity.

Capacity consumptions of a timetable consist not only of infrastructure occupation but also of timetable stability. Timetable stability is the ability to absorb delays. Ideally, delayed trains return to their scheduled train path by using the time allowances in the timetable. Goverde (2005) developed an analytical approach to evaluate network dependencies on timetable stability. The max-plus analysis approach is used to model a scheduled railway system and has been implemented in the software tool PETER (Goverde and Odijk (2002)).

Another software tool to assess the quality of timetables is OnTime. It combines the stochastic mapping of delay and the analytical calculation of delay propagation (Büker and Seybold (2012)).

In long-term strategic planning only limited knowledge about the future timetable is available, which requires stochastic tools to evaluate capacity. Schwanhäußer (1974) introduced an approach based on queueing theory for capacity evaluation. Since this approach is used for the case study (chapter 4), it is described in detail below. Wendler (2007) aims to predict the scheduled waiting time by means of a semi-Markovian queueing model. A discussion about queueing based approaches to assess the capacity of railway lines in Germany can be found in Weik et al. (2016).

Several papers focus on the capacity assessment of the railway system as a whole. A queueing network model is provided in Huisman et al. (2002). Mussone and Calvo (2013) present an analytical method based on an optimization model to assess the capacity of a railway system.

Analytical Method STRELE

In Germany, the timetable-independent analytical method by Schwanhäußer (1974) and Schwanhäußer (1994), which aims to determine the capacity of a railway line by calculation of expected waiting times, is widely used. A line is decomposed into overtaking sections.

Between two overtaking stations, the STRELE formula estimates mean knock-on delays

$$\bar{K} = \left(p_{del} - \frac{p_{del}^2}{2}\right) \cdot \frac{\bar{t}_{del}^2}{\bar{b} + \bar{t}_{del} \cdot \left(1 - e^{-\frac{\bar{t}_h}{\bar{t}_{del}}}\right)} \cdot \left[p_{eq} \cdot \left(1 - e^{-\frac{\bar{t}_{h,eq}}{\bar{t}_{del}}}\right)^2 + \left(1 - p_{eq}\right) \cdot \frac{\bar{t}_{h,diff}}{\bar{t}_{del}} \cdot \left(1 - e^{-2 \cdot \frac{\bar{t}_{h,diff}}{\bar{t}_{del}}}\right) + \frac{\bar{t}_h}{\bar{b}} \cdot \left(1 - e^{-\frac{\bar{t}_h}{\bar{t}_{del}}}\right)^2 \right]. \quad (1)$$

Input parameters for this formula are

p_{del}	mean probability of primary delays,
$\bar{t}_{del}$	mean time of delay of delayed trains,
$\bar{b}$	mean buffer time,
p_{eq}	probability of two trains with equal rank,
$\bar{t}_h$	mean minimum headway time,
$\bar{t}_{h,eq}$	mean minimum headway time between trains with equal rank and
$\bar{t}_{h,diff}$	mean minimum headway time between trains with different rank.

A defined level of service, which regulates the maximum admissible sum of knock-on delays, has been defined based on a statistical analysis of operation data to assess the optimal quality in operation. Calculated knock-on delays are compared with permissible waiting times in order to determine the capacity of the investigated railway line. DB Netz AG (2009) specifies quality levels for Germany. Admissible knock-on delays $adm\sum K$ on railway lines are defined as

$$adm\sum K = t_I \cdot q \cdot 0{,}260 \cdot e^{-1{,}3 \cdot ptr}. \quad (2)$$

Input parameters for this formula are the investigation period t_I, the quality factor q ($q=1$ for optimal quality) and the ratio of passenger trains p_{ptr}. Equating the STRELE formula to the level of service specified by Eq. (2) and solving for the buffer time, the minimum required buffer time b_{req} can be calculated. The corresponding number of trains n is obtained by

$$n = \frac{t_I}{\bar{t}_h + b_{req}}. \quad (3)$$

The STRELE formula is implemented in software tools such as LUKS®, which is the standard tool for capacity calculation in Germany. Even though the method is mainly used in Germany, it is transferable to any other infrastructure manager or analyst.

This approach is mainly used in long-term planning since it does not require an existing timetable. Merely little knowledge about the timetable e.g. train frequencies is necessary. Thus, it is suitable for comparing different infrastructure designs regardless of the precise operation concept. Compared with simulations, it takes less computing time to determine the capacity of a railway line. The performance indicators are easy to compare with defined limits and possess a validity extending far into the future.

3 Method

This chapter shows how the prohibition of train crossings can be included when calculating blocking times and how this transfers to the capacity assessment with the STRELE formula. In many cases, passenger trains have priority over freight trains. For an easier understanding, passenger trains are defined as priority trains in the following text, except otherwise stated. The method is applicable accordingly if the priorities are defined differently.

3.1 Restrictions for Tunnel Utilization

For German infrastructure, the federal railway authority defined which tunnels are affected by the prohibition of freight trains crossing passenger trains (Eisenbahn-Bundesamt (2008)). The prohibition applies for new double-track tunnels with a length of more than 500 m. When the tunnel length exceeds 1000 m, separate tubes for each direction are recommended.

In order to prevent passenger and freight trains from crossing each other in a tunnel, freight trains need to stop and wait at the tunnel's entry signal until the passenger train has cleared the infrastructure. As long as passenger trains are prioritized over freight trains, this prohibition supposedly only disturbs freight trains but may cause knock-on delays to more trains.

When two tunnels are built closely together and the distance between them is too short for a freight train to stop, they cannot be occupied separately. These tunnels are modelled as one continuous tunnel.

The following sections describe the method quantifying the effects on freight trains using modified blocking times. In tunnel blocks, blocking times are extended by the mean time a freight train needs to stop and wait for the prioritized passenger train to leave the tunnel. In section 3.2, the occupancy time of passenger trains is determined. Section 3.3 estimates the number of freight trains, which get disturbed and need to wait for passenger trains to leave a tunnel. With this information, it is possible to calculate new blocking times for the affected freight trains in section 3.4. Extended blocking times in relevant blocks lead to longer minimum headway times (section 3.5), which are input parameters of the STRELE formula. Mean knock-on delays increase with longer minimum headway times and reduce the capacity of a line.

3.2 Occupancy Time Rate

As long as a passenger train drives through a tunnel, freight trains need to wait at the tunnel's entry signal. During the occupancy time $t_{o,i}$, the passenger train i occupies the tunnel and prevents freight trains from entering. The occupancy time applies for the whole tunnel's length, which in the following example extends into two blocks. In Figure 3, solid lines show the division of block sections in the direction Node A – Node B and broken lines show the division in the opposite direction. The overall occupancy time of the tunnel begins at the tunnel's entry signal and ends at the location of the entry signal for the opposite direction. At this location, there is usually a clearing point to control that a train has left the tunnel. If this is not the case, the occupancy time prolongs up to the next clearing point.

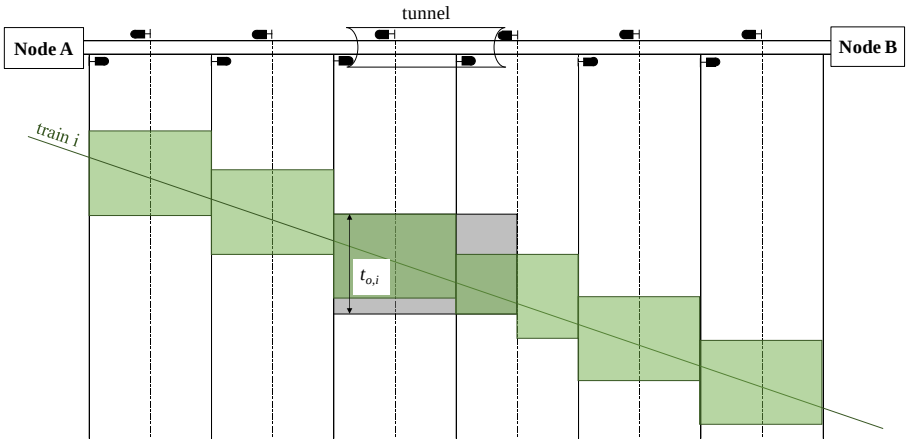


Figure 3: Occupancy time $t_{o,i}$

Occupancy times may overlap when trains of the same category follow each other through the tunnel. This occurs particularly in tunnels divided into several blocks. Two passenger trains with overlapping occupancy times prevent a freight train from entering the tunnel concurrently. In Figure 4, prioritized passenger trains *i* and *j* occupy the tunnel at the same time. The time during which both trains prevent a freight train from entering the tunnel is called overlapping time. The overlapping time reduces the total occupancy time of passenger trains and therefore the obstruction of freight trains. With this information, a mean occupancy time $\bar{t}_o$ can be calculated, which describes how long passenger trains occupy the tunnel on average.

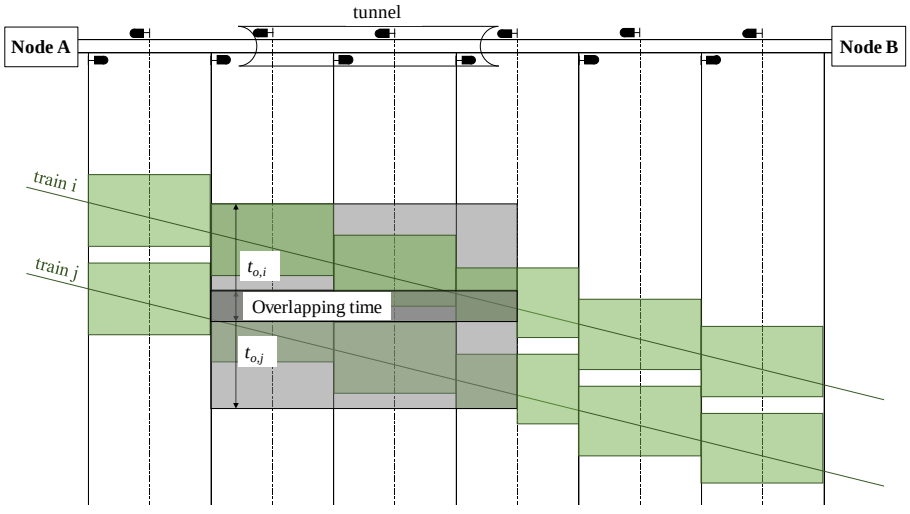


Figure 4: Overlapping of the occupancy times $t_{o,i}$ and $t_{o,j}$

When two passenger trains follow each other within a short period of time, a freight train with lower priority lets both trains pass in order not to disturb the passenger trains. To respect this priority, an operational service time supplement t_S, which extends the time frame during which a low priority train gets disturbed, is included. During this additional time, the tunnel is already blocked for the prioritized passenger train even though it has not entered the tunnel yet. Like occupancy times, service time supplements of two prioritized trains can overlap. Overlapping happens mainly in long tunnels. During the overlapping of the service time supplement $\bar{t}_{s,o}$, both passenger trains disturb the freight train. This reduces the total occupancy time of prioritized trains $T_{o,prio}$, which includes occupancy times and service time supplements for these trains:

$$T_{o,prio} = n_{prio} \cdot \left(\bar{t}_o + \bar{t}_s - \bar{t}_{s,o}\right) \tag{4}$$

with

n_{prio}	number of prioritized trains,
$\bar{t}_o$	mean occupancy time,
$\bar{t}_s$	mean service time supplement and
$\bar{t}_{s,o}$	mean service time supplement overlap.

The time period during which a freight train has to wait before entering a tunnel depends on the priority of passenger trains running in the opposite direction. If the freight and passenger trains had the same priority, the freight train would only have to let passenger trains with an earlier arrival at the tunnel pass ("first come, first serve" principle). The total occupancy time of trains with equal priority

$$T_{o,eq} = n_{eq} \cdot \bar{t}_o \tag{5}$$

is the product of the number of trains with equal priority n_{eq} and the mean occupancy time $\bar{t}_o$.

The total occupancy time for freight trains results from the occupancy times caused by equal and prioritized passenger trains. The occupancy time rate ρ^* is the ratio of the total occupancy time per investigation period T:

$$\rho^* = \frac{(T_{o,prio} + T_{o,eq})}{T} \tag{6}$$

with

ρ^*	occupancy time rate,
$T_{o,prio}$	total occupancy time of prioritized trains,
$T_{o,eq}$	total occupancy time of trains with equal priority and
T	investigation period.

3.3 Disturbed and Undisturbed Trains

The occupancy time rate ρ^* represents the ratio of the total occupancy time per investigation period for one tunnel. It is assumed that the occupancy time rate ρ^* equals the rate of disturbed trains. Thus, the rate of disturbed trains per tunnel ρ_t^* is known at this point. Given that there are several tunnels on one railway line, obstructions might occur in more than one tunnel. To determine the rate of disturbed trains for a whole line including several tunnels, it is necessary to use probability calculus. Assuming that both directional tracks are uncorrelated, a possible obstruction in tunnel 1 does not affect whether the train is disturbed in tunnel 2.

Formulas for two relevant tunnel blocks are shown in Figure 5. With the known rate of disturbed trains per tunnel (ρ_{t1}^* and ρ_{t2}^*) it is possible to calculate the rate of disturbed trains in only one specific tunnel (ρ_{t1} and ρ_{t2}). Accordingly, the rate of trains which are disturbed in both tunnels ($\rho_{t1,2}$) or in none (ρ_u) can be determined.

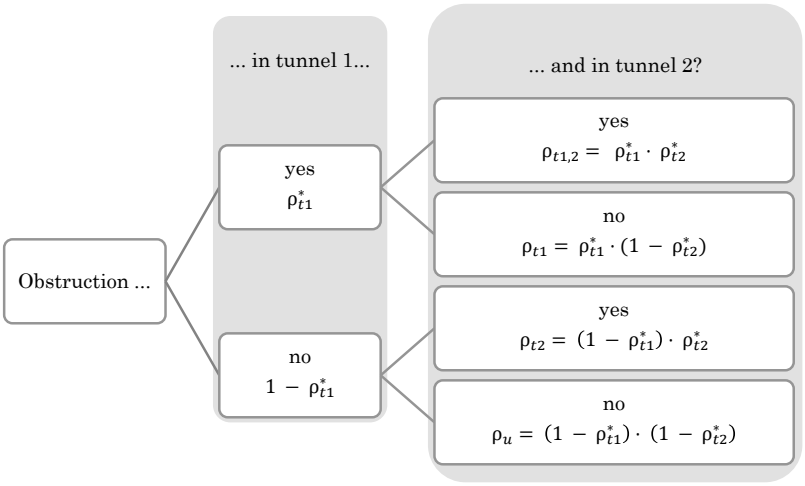


Figure 5: Probabilities of disturbed trains on a line with two tunnels

3.4 Blocking Time Modification

The blocking time extensions for freight trains represent the mean time a train has to wait before entering the tunnel in order not to disturb a prioritized passenger train. If the passenger train and the freight train had equal priority, the first train arriving at the tunnel would run first. In that case, the freight train would have to wait for at most the occupancy time $t_{o,i}$ of the passenger train i (see Figure 6). The mean blocking time extension caused by trains with equal priority

$$\bar{t}_{ext,eq} = \frac{1}{2} \cdot \frac{n_{eq}}{n_{eq} + n_{prio}} \cdot \bar{t}_o \tag{7}$$

is the product of the probability of needing to let a train with equal priority pass and the corresponding waiting time $\bar{t}_o$. Only the number of trains with equal priority n_{eq} and the number of prioritized trains n_{prio} are part of the formula since trains with lower priority do not cause disruptions and blocking time extensions are solely considered for disrupted trains.

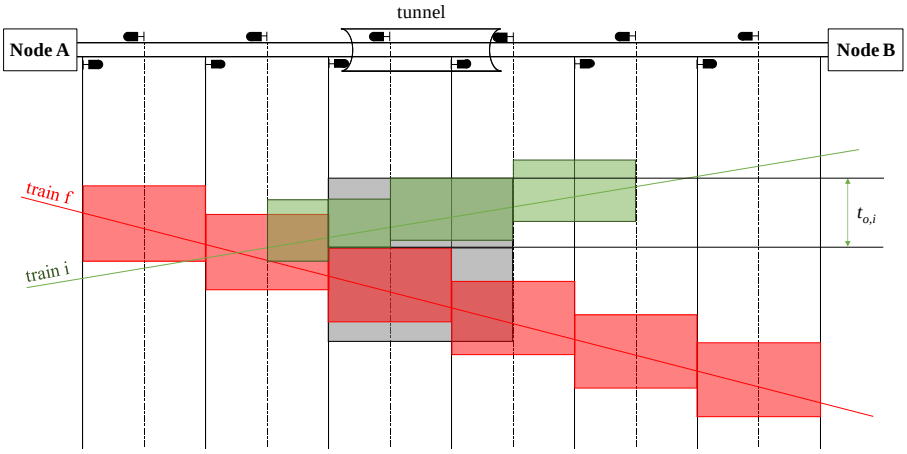


Figure 6: Maximum waiting time for freight train f caused by passenger train i with equal priority

If passenger trains are prioritized, freight trains must not disturb them. Thus, the freight train needs to let passenger trains, which are already driving through the tunnel and also those which are about to enter the tunnel, pass. The maximum waiting time to let one prioritized passenger train i pass consists of the occupancy time $t_{o,i}$ extended by the operational service time supplement t_s (see Figure 7).

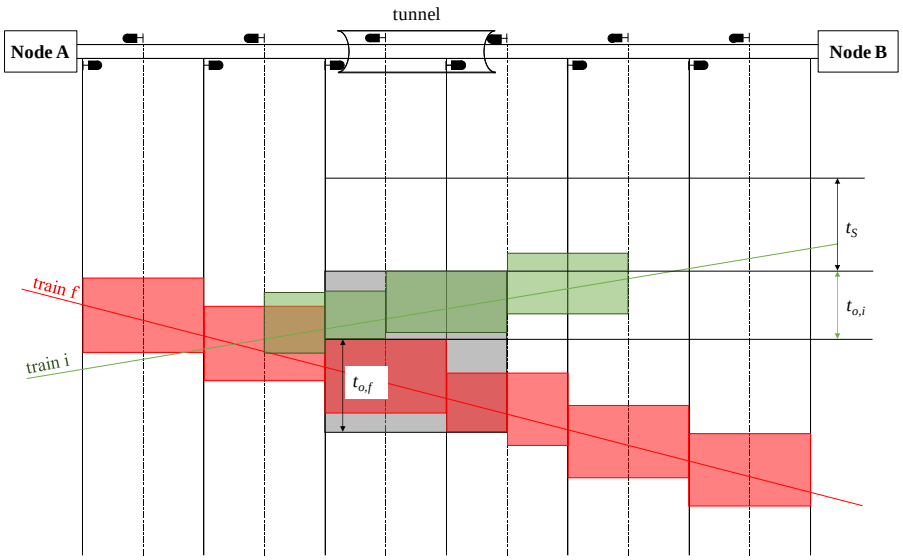


Figure 7: Maximum waiting time for freight train f caused by prioritized passenger train i

A freight train might have to let several priority trains pass before it can enter the tunnel without disrupting any priority trains. If it has to let more than one passenger train pass, the waiting time lengthens. For each of the expected additional passenger trains $E[n_{wait}]$, the waiting time is extended by the passenger trains' mean occupancy time $\bar{t}_o$ and the expected buffer times between them $E[b]$. This buffer time is shorter than the time that is needed by the freight train to drive through the tunnel without disrupting prioritized passenger trains. The mean blocking time extension caused by prioritized trains $\bar{t}_{ext,prio}$ is the product of the probability of letting a certain number of trains pass and the corresponding additional waiting time.

$$\bar{t}_{ext,prio} = \frac{1}{2} \cdot \frac{n_{prio}}{n_{eq} + n_{prio}} \cdot (\bar{t}_o + t_S) + \frac{n_{prio}}{n_{eq} + n_{prio}} \cdot (\bar{t}_o + E[b]) \cdot E[n_{wait}] \quad (8)$$

with

n_{prio}	number of prioritized trains,
n_{eq}	number of trains with equal priority,
$\bar{t}_o$	mean occupancy time,
t_s	service time supplement,
$E[b]$	expected value for the buffer time b between two prioritized passenger trains for which the freight train needs to wait and
$E[n_{wait}]$	expected number of passenger trains for which the freight train needs to wait in order to let them pass first.

The entire mean blocking time extension

$$\bar{t}_{ext} = \bar{t}_{ext,eq} + \bar{t}_{ext,prio} \tag{9}$$

consists of the mean blocking time extension caused by trains with equal rank $\bar{t}_{ext,eq}$ and those caused by prioritized trains $\bar{t}_{ext,prio}$.

3.5 Modification of Minimum Headway Times

The minimum headway times for undisturbed trains remain unchanged whereas the minimum headway times for disturbed trains receive a supplement. The rate of disturbed trains and the blocking time extension for each tunnel are necessary input variables to calculate minimum headway times of disturbed trains. As shown in Figure 8, the blocking time extensions cause the blocking time to begin earlier in tunnel blocks.

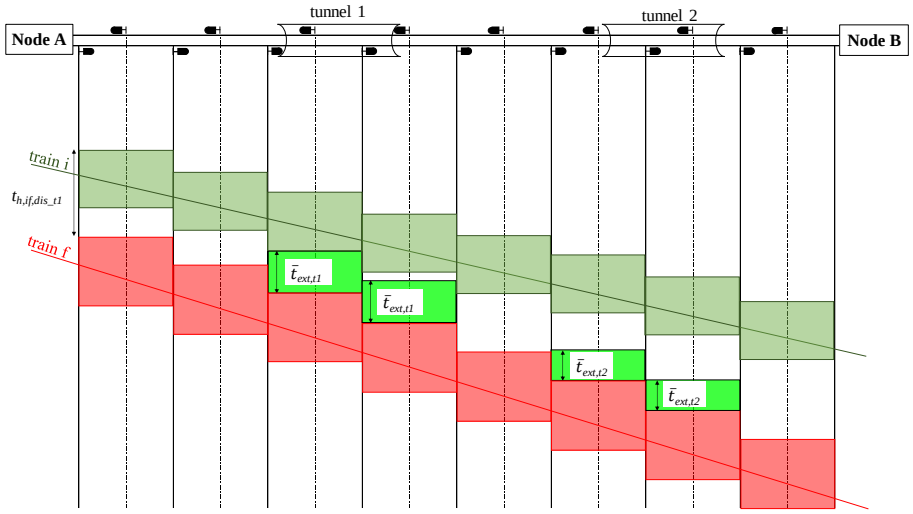


Figure 8: Minimum headway time t_{h,if,dis_t1} considering blocking time extensions

In this example, the first block of tunnel 1 is relevant for the modified minimum headway time between passenger train *i* and freight train *f*. The minimum headway times of undisturbed trains and those of disturbed trains are weighted according to the rate of disturbed trains. For a line with two tunnels the modified minimum headway time $\bar{t}_{h,mod,if}$ is calculated as

$$\bar{t}_{h,mod,if} = \rho_u \cdot t_{h,if} + \rho^*_{t1} \cdot t_{h,if,dis\_t1} + \rho_{t2} \cdot t_{h,if,dis\_t2}\,. \tag{10}$$

Variables in this formula are

ρ_u	rate of undisturbed trains,
$t_{h,if}$	minimum headway time between trains i and f,
ρ_{t1}^*	rate of all in tunnel 1 disturbed trains,
t_{h,if,dis_t1}	minimum headway time of in tunnel 1 disturbed trains,
ρ_{t2}	rate of only in tunnel 2 disturbed trains and
t_{h,if,dis_t2}	minimum headway time of only in tunnel 2 disturbed trains.

The formula is extended accordingly if a line includes more or less than two tunnels. With the help of modified minimum headway times, obstructions caused by the prohibition of train crossings can be included when applying the STRELE formula. Longer minimum headway times increase the calculated knock-on delays and therefore decrease the capacity.

4 Case Studies

This chapter presents the application of the method to include the prohibition of train crossings in tunnels on two exemplary regional railway lines in order to validate the method's plausibility. The capacity of the lines with and without tunnels is calculated using the described method. Furthermore, the same lines are simulated to evaluate the influence of the prohibition on the operating quality.

4.1 Line 1

The 80 km long double-track railway line, which is used for the case study, comprises three tunnels. The shortest tunnel has a length of 650 m and the longest of 1393 m as indicated in Figure 9.

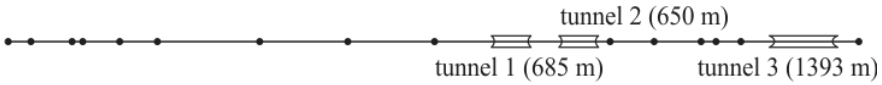


Figure 9: Line 1 and position of double-track tunnels

Table 1: Operating program on Line 1

	Passenger trains	**Freight trains**	**In total**
Direction 1	47	8	55
Direction 2	48	7	55

The operating program, which includes 110 trains per day, is depicted in Table 1. 34 of the trains per direction run through tunnel 1 and 2. 33 trains in direction 1 and 34 trains in direction 2 use tunnel 3.

Using the software LUKS® and the method presented in chapter 3, the capacity is determined analytically with and without tunnels. Table 2 shows the results for both directions separately.

Table 2: Capacity in trains per day with and without the prohibition of train crossings

Line Capacity	Without tunnels	With tunnels	Difference
Direction 1	71	66	- 7 %
Direction 2	70	67	- 4 %

The line capacity of 70 and 71 trains per day declines by 3 to 5 trains per direction when including the prohibition.

In total, the operating program on the existing line includes 55 trains per direction. With as well as without the prohibition of train crossings the capacity exceeds the actual number of trains significantly. Consequently, the line has a moderate utilization rate. The operating quality is respectively high.

Additionally, each scenario is simulated 200 times with the help of the software LUKS®. Without considering the prohibition of train crossings, the simulation results in a total delay of 56 minutes per day. When including the prohibition, the total delay increases by 6 minutes (Table 3).

Table 3: Total delay in minutes per day with and without the prohibition of train crossings

Total delay	Without tunnels	With tunnels	Difference
Both directions	56	62	+ 11 %

4.2 Line 2

The second examined line is an approximately 100 km long double-track railway line with one 698 m long tunnel (see Figure 10).

Figure 10: Line 2 and position of the tunnel

Table 4: Operating program in the tunnel on Line 2

	Passenger trains	Freight trains	In total
Direction 1	36	23	59
Direction 2	34	24	58

The operating program is depicted in Table 4. In total, 117 trains of which 47 are freight trains are scheduled to drive through the tunnel.

Using the software LUKS® and the presented method, the capacity is determined analytically with and without the prohibition. Table 5 shows the results for both directions separately.

Table 5: Capacity in trains per day with and without the prohibition of train crossings

Line Capacity	Without tunnels	With tunnel	Difference
Direction 1	111	106	- 5 %
Direction 2	97	97	0 %

The line capacity in direction 1 declines from 111 to 106 trains per day, which means by 5 trains when including the prohibition in the tunnel. The minimum headway times on the tunnel's line section are extended by the prohibition of train crossings. For the whole line though, another section is relevant for the decisive minimum headway time that leads to unchanged line capacity in direction 2.

Without considering the prohibition of train crossings, the simulation results in a total delay of 87 minutes per day. When including the prohibition, the total delay of all trains increases by 3 minutes (Table 6). Since the line has only a low utilization rate, the effects on the operating quality by the prohibition are rather low.

Table 6: Total delay in minutes per day with and without the prohibition of train crossings

Total delay	Without tunnels	With tunnel	Difference
Both directions	87	90	+ 3 %

4.3 Evaluation

It can be seen that with the modified analytical method as well as with simulations, the prohibition of train crossings in tunnels reduces the line's operating quality. Corresponding to the low utilization rate of the examined lines, the simulation shows only a slightly increased delay caused by the additional obstruction. The analytical method also shows a slight deterioration of the capacity.

A comparison of the results from the analytical method and from simulations is only possible to a limited extent. One specific timetable on each of the two lines has been used for simulations. Changes in the timetable such as different arrival and departure times will change the results. However, this does not affect the results of the timetable-independent analytical method. Thus, the case study is incapable of proving that the presented method reproduces the impact of the prohibition perfectly, but it still validates the plausibility of the results. The results of the simulations and analytical method show similar relative changes of the capacity and the total delay.

The presented method is going to be implemented in the software LUKS®. Since this will reduce the calculating time significantly, it will easily be possible to calculate the effect on a larger number of generic and existing railway lines, including those with high traffic loads.

5 Conclusions

This paper presents a new method to include the prohibition of passenger and freight train crossings in double-track railway tunnels by modifying blocking times of disturbed trains. The blocking time extensions represent the mean time a train has to wait before entering the tunnel in order not to disturb a prioritized train. Changes in blocking times influence minimum headway times, which are input variables for the capacity calculation. The extent as to which the prohibition of train crossing in tunnels influences line capacity is shown in a

case study based on the STRELE method. Longer minimum headway times caused by the infrastructure constraint increase knock-on delays and thereby reduce the capacity. This gives the opportunity to calculate the effects the prohibition of train crossings in tunnels has on line capacity and thereby helps to improve the results of analytical methods. The exemplary application of the presented method on two railway lines validates the plausibility of the results. After being implemented in the software LUKS®, the method can easily be applied on numerous lines with different operating programs to make sure the results are also plausible for these scenarios.

6 References

Brünger, O. and Gröger, T., 2003. "Fahrplantrassen managen und Fahrplanerstellung simulieren", in *Proc. 19. Verkehrswissenschaftliche Tage*, Dresden.

Büker, T. and Seybold, B., 2012. "Stochastic modelling of delay propagation in large networks". *Journal of Rail Transport Planning & Management*, Vol. 2 No. 1, pp. 34–50.

DB Netz AG, 2009. *Richtlinie 405 – Fahrwegkapazität*.

Eisenbahn-Bundesamt, 2008. *Anforderungen des Brand- und Katastrophenschutzes an den Bau und den Betrieb von Eisenbahntunneln*.

Goverde, R.M.P., 2005. "Punctuality of Railway Operations and Timetable Stability Analysis", Delft University of Technology, Delft, 2005.

Goverde, R.M.P. and Odijk, M.A., 2002. "Performance evaluation of network timetables using PETER", in Allan, J., Hill, R.J., Brebbia, C.A. (Ed.), *Computers in Railways VIII*, WIT Press, Southampton, pp. 731–740.

Happel, O., 1959. "Sperrzeiten als Grundlage der Fahrplankonstruktion". *Eisenbahntechnische Rundschau (ETR) No.* 2, pp. 79–90.

Huisman, T., Boucherie, R.J. and van Dijk, N.M., 2002. "A solvable queueing network model for railway networks and its validation and applications for the Netherlands", in *European Journal of Operation Research 142*, pp. 30–51.

Jacobs, J., 2008. "Rescheduling", in Hansen, I.A., Pachl, J. and Albrecht, T. (Eds.), *Railway timetable & traffic: Analysis, modelling, simulation*, 1. ed., Eurailpress, Hamburg, pp. 182–191.

Janecek, D. and Weymann, F., 2010. "LUKS - Analysis of lines and junctions", in *Proc. of the 12th World Conference on Transport Research (WCTR)*, Lisbon.

Mussone, L. and Calvo, R.W., 2013. "An analytical approach to calculate the capacity of a railway system". *European Journal of Operational Research*, Vol. 228 No. 1, pp. 11–23.

Nießen, N., 2014. "Queueing", in Hansen, I.A. and Pachl, J. (Eds.), *Railway Timetabling & Operations: Analysis, modelling, optimisation, simulation, performance evaluation*, 2. rev. and extended ed., Eurailpress, Hamburg, pp. 117–131.

Pachl, J., 2014. "Timetable Design Principles", in Hansen, I.A. and Pachl, J. (Eds.), *Railway Timetabling & Operations: Analysis, modelling, optimisation, simulation, performance evaluation*, 2. rev. and extended ed., Eurailpress, Hamburg, pp. 13–46.

Radtke and Bendfeldt, 2001. "Handling of railway operation problems with RailSys", in *Proc. of the 5th World Congress on Rail Research*, Cologne.

Schwanhäußer, W., 1974. "Die Bemessung der Pufferzeiten im Fahrplangefüge der Eisenbahn", Verkehrswissenschaftliches Institut, RWTH Aachen University, 1974.

Schwanhäußer, W., 1994. "The status of German railway operations management in research and practice". *Transportation Research Part A: Policy and Practice*, Vol. 28 No. 6, pp. 495–500.

UIC, 2003. *Code 779-9 Safety in railway tunnels*.
UIC, 2004. *Code 406 Capacity*.
UIC, 2013. *Code 406 Capacity*.
Watson and Medeossi, 2014. "Simulation", in Hansen, I.A. and Pachl, J. (Eds.), *Railway Timetabling & Operations: Analysis, modelling, optimisation, simulation, performance evaluation*, 2. rev. and extended ed., Eurailpress, Hamburg, pp. 191–215.
Weik, N., Niebel, N. and Nießen, N., 2016. "Capacity analysis of railway lines in Germany – A rigorous discussion of the queueing based approach". *Journal of Rail Transport Planning & Management*, Vol. 6 No. 2, pp. 99–115.
Wendler, E., 2007. "The scheduled waiting time on railway lines". *Transportation Research Part B: Methodological*, Vol. 41 No. 2, pp. 148–158.
Wendler, E., 2009. "Influence of ETCS on the capacity of lines", in *Compendium on ERTMS, European Rail Traffic Management System*, 1. ed., DVV Media Group Eurailpress, Hamburg.
Yuan, J. and Hansen, I.A., 2007. "Optimizing capacity utilization of stations by estimating knock-on train delays". *Transportation Research Part B: Methodological*, Vol. 41 No. 2, pp. 202–217.

A Study of the Performance and Utilization of High Speed Rail in China based on UIC 406 Compression Method

Jie Li [a,b], Dian Wang[a] ,Qiyuan Peng[a,1] , Yuxiang Yang [a]
[a] School of Transportation and Logistics, Southwest Jiaotong University
High-tech Zone West Campus, 611756, Chengdu, China
[b] Civil, Buildings and Environmental Engineering Department, SAPIENZA Università di Roma
Via Eudossiana 18, 00184 Roma, Italy
[1] E-mail: qiyuan-peng@swjtu.cn, Phone: +86 13808061287

Abstract

UIC Code 406 is an easy and effective way of calculating the capacity consumption. Based on the UIC 406 capacity method, the capacity consumption of railway infrastructure can be measured by compressing the timetable. Regarding the UIC 406 capacity leaflet as a framework, an optimal method are proposed to compress the real-record timetable for practical capacity consumption, with respect to train orders, overtaking and crossing on the given timetable. The proposed method is applied to evaluate the capacity consumption of Wuhan-Guangzhou HSR in China. Firstly the Wuhan-Guangzhou HSR is divided into several sections according to the station class on the line. Then each section can be handled separately by the UIC 406 capacity method and the capacity consumption can be got. Based on the result the temporal-spatial uneven of capacity utilization and the capacity bottleneck of the line can be defined. It can be concluded that the temporal-spatial uneven of capacity consumption of Wuhan-Guangzhou HSR is obvious. The capacity consumption in the early times during one day is high, and the section from Guangzhou South station to Yueyang East station is easy to be a bottleneck due to the layout of the HSR. Besides, the analysis shows that the capacity consumption on railway lines is very responsive section examined. Therefore, the division of the lines into sections is of major importance for the results of capacity consumption.

Keywords

UIC Code 406, High speed railway, Capacity consumption, Capacity bottleneck

1 Introduction

Many HSRs in China are struggling to accommodate necessary train services on the limited infrastructure. In this regard, efficient management and planning for measuring capacity are necessary. Railway faces capacity constraints on their main infrastructure as well as their nodal bottlenecks, hence comprehensive overview of capacity is necessary (Landex (2008)).

UIC Code 406 is an easy and effective way of calculating the capacity consumption. In the past years, the UIC method has been applied in a number of studies (Whalborg (2004) and Kaas (2006)). Landex and Schittenhelm (2008) described how the UIC 406 methodology was expounded in Denmark. Lindner (2011) summarized the main contents of UIC406 and discussed several different problems result from applying UIC Code 406. Landex (2009) discussed the differences between capacity analyses of double track lines

and single track lines using the UIC 406 capacity method. Pavlides and Chow (2016) measured the utilization of track capacity by using the occupation measure specified in the UIC 406 'Capacity' code.

In summary, the UIC 406 capacity method can be expounded in different ways and has been applied to some European countries. However, seldom researches have applied the UIC 406 method on Chinese HSR. In this paper, the UIC 406 method will be used to evaluate the capacity utilization of Chinese HSR.

Wuhan-Guangzhou HSR is one of the busiest railways in China. By using the UIC 406 method and the compress timetable method, the capacity utilization for Wuhan-Guangzhou HSR is analysed. The real-record train operation data from several different databases (supplied by the China Railway Administration) has been processed. The data consists of the scheduled timetable, real-record timetable and operational data such as recorded delays, train weights and train lengths, from January, 2015 to December, 2016.

The remainder of this paper is organized as follows.First, in section 2 the UIC 406 method is introduced and the optimal method based on UIC 406 is proposed. In section 3 the UIC 406 is applied on the Chinese HSR and the capacity consumption is evaluated. Then the capacity performance of the HSR is analyzed and the bottleneck is identified. Besides, how the division of the lines into sections affects the capacity consumption is discussed. Finally, conclusion and future envisions are discussed in section 4.

2 Method

2.1 The UIC 406 Capacity Calculation Model

The UIC 406 code defines railway capacity as "the total number of possible paths in a defined time window, considering the actual path mix or known developments respectively and the ... own assumptions" (Cordeau (1998)). Based on the UIC 406 method, the capacity consumption of railway infrastructure can be measured by compressing the timetable graphs so that the buffer times are equals to 0, as well as considering the safety headway of trains. Meanwhile, the train sequence and the timetable structure remained the same as in the real-record timetable. Some researches proposed that the total capacity consumption can be valued in a simple analytical way by the sum of the infrastructure occupation time in minutes, the buffer time in minutes, the supplement for single track lines and the supplement for maintain. The study aims to evaluate the capacity utilization and identify the capacity bottleneck of the Wuhan-Guangzhou HSR, which is a double-track line. For simply, the total capacity consumption in one section is just evaluated by the train occupation time in the compressed timetable. The percentage capacity consumption R can be calculated as the quotient of total consumption time K and chosen time window T , which is shown in equation (1). The capacity consumption represents the chained occupation rate, as the compression does not have to be done for a partition consisting of only one specific block interval, and an examination partition can consist of more than one block interval.

$$R = \frac{K}{T} \tag{1}$$

For a given timetable, the objective of compressing timetable is to minimize the train occupation time of all the trains involved during the time window in a section, meanwhile follows the principles below.

Principle 1: Both the train order and the travel speed in the compressed timetable

should be maintained as in the real-record timetable.

Principle 2: It is allowed to reduce the dwell time of trains; however the dwell time should be large enough for the necessary operation at the stations.

Principle 3: The buffer time in the compressed timetable is not necessary. The headway for trains should be guaranteed for a safety train operation.

In this paper, timetable compress process is treated as an optimal problem with the object of minimizing the train occupation time, which is detailed introduced in the following part.

Parameters and Decision Variables

All the symbols and parameters used in the formulation process are given as follows in Table 1.

Table 1: Symbols and parameters used in the model

Symbol	Definition
$G=\{S,H\}$	Physical railway network
$S=(s_1,s_2,\ldots,s_n)$	Set of stations distributed in the HRS line
$H=(h_1,h_2,\ldots,h_k)$	Set of sections in the HRS line
$L=(l_1,l_2,\ldots,l_j)$	Set of trains running on the HRS line
$S(l_j)=(s_j,s_{j+1},\ldots,s_{j+n})$	Denotes the train operation configuration that can be set along a linear corridor connecting $n+1$ stations $(s_j,s_{j+1},\ldots,s_{j+n})$
$\theta_{S_k}^{l_j}$	Binary variation to identify whether train l_j stops at station S_k, variable $\theta_{S_k}^{l_j}=1$ if train l_j is scheduled to stop at station S_k, and 0, otherwise.
yd_{ijk}	Indicator of departure order for trains l_i and l_j from station s_k ,if train l_i departs from station s_k before train l_j , yd_{ijk} =1, yd_{ijk} =0 otherwise
ya_{ijk}	Indicator of arrival order for trains l_i and l_j from station s_k ,if train l_i arrives at station s_k before train l_j , ya_{ijk} =1, ya_{ijk} =0 otherwise
$dwell_{\min s_k}^{l_j}$	The minimum dwelling time of each train l_j at station s_k
$dwell_{\max s_k}^{l_j}$	The maximum dwelling time of each train l_j at station s_k
I_a	The minimum arrival headway for each two consecutive trains
I_d	The minimum departure headway for each two consecutive trains
$r_{s_k s_{k+1}}^{l_j}$	The running time of train l_j from station s_k to station s_{k+1}

The model intends to get a minimum train occupation time considering the train operation safety and the given train order in the real-record timetable. Thus, two types of decision variables are proposed as follows in Table 2.

Table 2: Decision variables used in the model

Decision variables	Definition
$td_{s_k}^{l_j}$	The time train l_j departing from station s_k, $s_k \in S(l_j)$
$ta_{s_k}^{l_j}$	The time train l_j arriving at station s_k, $s_k \in S(l_j)$

Systematic constraints

In this subsection, a series of systematic constraints are formulated to provide the necessary services and guarantee the safety of trains in the compressed timetable. The involved constraints are formally formulated as follows.

Since the train order and the timetable structure in the compressed timetable is consistent with that of the real-record timetable, the operation zone constraints of trains, occupation uniqueness of blocks are satisfied. The compressing timetable model just considers the running time constrains, the dwell time constraints and the headway constraints.

(1) Running time constraints

$$td_{s_k}^{l_i} + r_{s_k s_{k+1}}^{l_i} = ta_{s_{k+1}}^{l_i}, \forall l_i \in L, s_k \in S(l_i), s_{k+1} \in S(l_i) \quad (2)$$

Equation (2) guarantee a continuous time-space path for the train, that is the arrival time $ta_{s_{k+1}}^{l_i}$ of train l_i at station s_{k+1} equals to the sum of the depart time $ta_{s_{k+1}}^{l_i}$ at the previous station and the running time $r_{s_k s_{k+1}}^{l_i}$ (including the departing additional time and the arriving additional time) in the section h_k. $r_{s_k s_{k+1}}^{l_i}$ can be calculated according to the real-record timetable.

(2) Dwell time constraints

$$td_{s_k}^{l_i} - ta_{s_k}^{l_i} \geq \theta_{s_k}^{l_i} \bullet \qquad \forall l_i \in L, s_k \in S(l_i) \quad (3)$$

$$td_{s_k}^{l_i} - ta_{s_k}^{l_i} \leq \theta_{s_k}^{l_i} \bullet \qquad \forall l_i \in L, s_k \in S(l_i) \quad (4)$$

For train which is scheduled to stop at station s_k, the dwell time is required for trains to conduct the necessary operation, such as the alighting and boarding of passengers, the shift handover of crews and so on. The dwell times of trains at different stations are various. According to the investment of dwell times in the real-record timetable, the dwell times vary within a range, and the distribution of dwell times is shown in Figure 1. Thus the dwell time should be long enough for the train operation as well as no more than the upper limitation. The minimum dwelling time for train operation at station s is guaranteed by equation (3). In addition, the dwell time for trains stop at some stations should no more than the maximum time, subjected to equation (4).

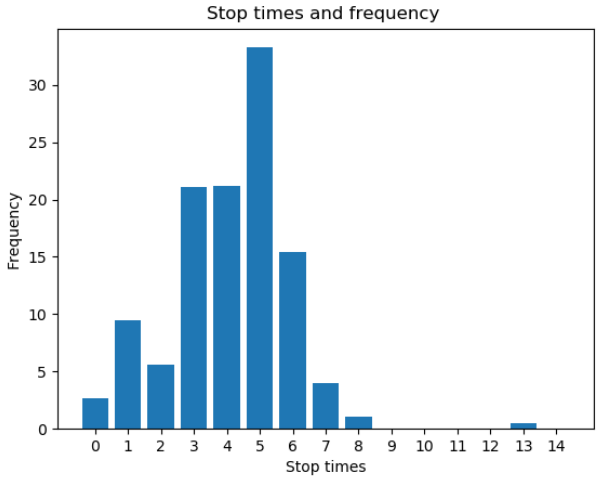


Figure 1: The distribution of dwell times on Wuhan-Guangzhou HSR

(3) Headway constraints

$$td_{s_k}^{l_i} + I_d \le td_{s_k}^{l_j} + M \bullet \left(1 - yd_{ijk}\right), \quad \forall l_i, l_j \in L, s_k \in S \tag{5}$$

$$td_{s_k}^{l_j} + I_d \le td_{s_k}^{l_i} + M \bullet yd_{ijk}, \quad \forall l_i, l_j \in L, s_k \in S \tag{6}$$

$$ta_{s_k}^{l_i} + I_a \le ta_{s_k}^{l_j} + M \bullet \left(1 - ya_{ijk}\right), \quad \forall l_i, l_j \in L, s_k \in S \tag{7}$$

$$ta_{s_k}^{l_j} + I_a \le ta_{s_k}^{l_i} + M \bullet ya_{ijk}, \quad \forall l_i, l_j \in L, s_k \in S \tag{8}$$

The headway constraints aimed to guarantee all the involved trains to keep the minimum safety headway for each of the two trains arriving or departing at the same station. There are two types of headway for each two consecutive trains, the headway of trains in section and the headway of trains at stations. For simply, the study just considers the headway of trains at stations, the safety headway of trains in section can be guaranteed by keeping the minimum safe headway for each of the two consecutive trains when they depart from or arrive at each station. Equation (5) and equation (6) is used to ensure the minimum departure time interval between the adjacent trains at stations while the minimum arrival time interval between the adjacent trains are guaranteed by equation (7) and equation (8). In detail, if train l_i departs from the station s_k earlier than train l_j, then $yd_{ijk} = 1$ and just the equation (5) is effective and the safety headway can be guaranteed. Meanwhile, the M in equation (6) is large enough to keep the equation reasonable. On the other hand, as train l_i departs from the stations s_k after than train l_j, then $yd_{ijk} = 0$ and in this case the equation (6) is active and the equation (5) is reasonable. Similarly the minimum arrival headway can be ensured by the constrain (7) and constrain (8).

Objective: minimizing the operation time of trains

For a given timetable, the objective of compressing timetable is to minimize the train occupation time of all the trains involved in the chosen time window T. The objective can

be calculated in equation (9), in which $d_{s_1}^{l_1}$ is the departure time of the first train from the first station of the section, and $a_{s_k}^{l_i}$ is arrival time of the last train at the last station in section. A detailed graph is shown in Figure 2.

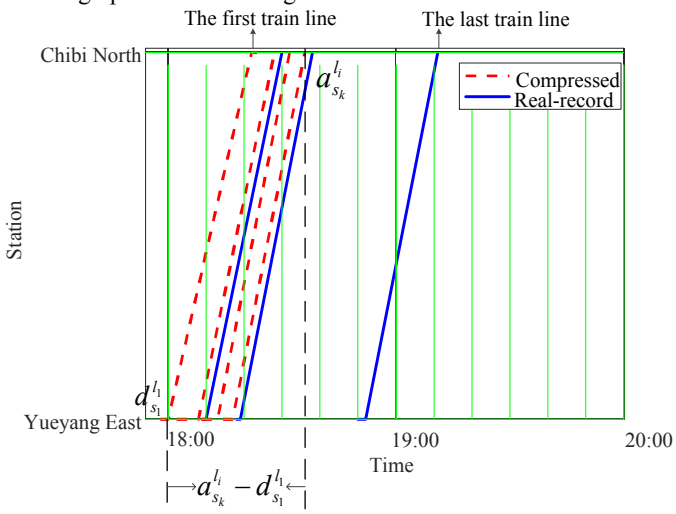


Figure 2: A detailed description of the compressed timetable

$$\min \quad T = \max\left(a_{s_k}^{l_i}\right) - \min\left(d_{s_1}^{l_1}\right) \tag{9}$$

However, on one hand, the objective function has a low constraint on other train lines since just two decision variables of two trains are involved in the objective function, thus the solver to the optimal problem may not be unique. That is there may be several compressed timetable corresponding to one train occupation time. On the other hand, the objective function may reduce the convergence speed due to the large scope space for the optimal solution.

In order to prevent the problems motioned above, the objective function has been promoted, which is shown in equation (10). The total arrival time and departure time of all the involved trains are adopted as the evaluation index to qualify the compressed timetables. In this way the train occupation time in the compressed timetable can be minimized, and the total travel time of trains can be reduced as well.

$$\min \quad Z = \sum_{l_i \in L} \sum_{s_k \in S} \left(a_{s_k}^{l_i} + d_{s_k}^{l_i}\right) \tag{10}$$

The constrains from equation (3) to equation (8) are formulated in a liner way, as well as the objective function. The CPLEX solver is employed to solve the model.

2.2 The steps of calculate capacity consumption

The proposed optimal model based on UIC 406 method is applied to calculate the capacity consumption of each section on HSR in China, based on the real-record timetable. Then the capacity bottleneck can be identified according to the capacity calculation results. The steps of calculate capacity consumption are shown as follows.

Step 1: It is necessary to divide the HSR line into smaller line sections, which can be

handled separately by the UIC 406 capacity method.

Step 2: The compression of the timetable graph has to be done with respect to train orders, overtaking and crossing which have been defined on the timetable. This means that neither the running times, running time supplement or block occupation times are allowed to be changed. As to one section, the capacity utilization can be calculated by the UIC406 method after the train order, overtaking and crossing in the section has been declared.

Step 3: The UIC 406 capacity calculation model is applied to the divided section, thus the capacity utilization of each section in the line can be calculated and the bottleneck section thereby might be excluded from the line.

3 Numerical experiments

To demonstrate the effectiveness and efficiency of the proposed optimal model for compressing timetable, the Wuhan-Guangzhou HSR corridor in China is taken as a case study in the numerical experiments.

In southern China, the 1069-km Wuhan-Guangzhou HSR directly connects Wuhan with Guangzhou. There are eighteen stations on the line and seventeen of them operate for passenger service. Figure 3 lists the 17 operational stations located along the Wuhan-Guangzhou HSR.

With the developing of Chinese HSR network, Wuhan-Guangzhou HSR and Kunming-Shanghai HSR intersect at Changsha South station while Wuhan-Guangzhou HSR and Hu-Han-Rong HSR intersect at Wuhan station. Hengyang-Liuzhou HSR joins into Wuhan-Guangzhou HSR at Hengyang station. The analysis on capacity performance of Wuhan-Guangzhou HSR is a typical case to learn about the capacity utilization of HSR in China.

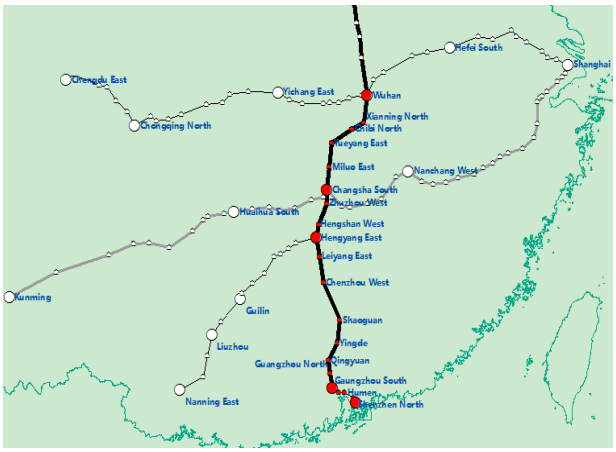


Figure 3: The layout of Wuhan-Guangzhou HSR

The daily train operation records of the Wuhan-Guangzhou HSR line were collected. Only the train data related to 15 stations and 14 sections from Guangzhou North station to Chibi North station are obtained from the Railway Company. The data gathered from 24th, February, 2015 to 30th, November, 2016, includes 29662 HSR train records for up-direction and 29662 HSR train records for down-direction. Table 1 shows a sample of

train operation record.

Table 1: Train running records in a database

Train No.	Date	Station	Arrival time	Departure time	Scheduled arrival time	Scheduled departure time
G1138	2015/3/24	Qingyuan	18:13:00	18:13:00	18:14:00	18:14:00
G1140	2015/3/24	Yingde	19:09:00	19:09:00	19:10:00	19:10:00
G1302	2015/3/24	Shaoguan	12:12:00	12:20:00	12:14:00	12:22:00

In addition, train running records contain the follow information.

- Train number, including train types distinguished by G and D,
- Name of stations,
- Arrival times, departure times, planned arrival times, and planned departure times in the "year/month/day and hour: minute: second" format,
- The interval between train events at stations, including the interval between the successively arriving trains and interval between the successively departing trains at each station.

The scheduled railway timetable in China is adjusted occasionally, especially as new lines start to operate. As we know the scheduled timetable was adjusted on 2015/05/20 and 2015/07/01.The real-record timetable data on 2015/04/20, 2015/06/20 and 2015/08/01 are extracted from the dataset. Timetable compressed method are applied on timetable data to evaluate the capacity performance.

The HSR from Guangzhou North station to Chibi North station is divided into several sections due to the class of stations and the passenger distribution. The information of each section is shown in Table 2.

Table 2: The divided section of Wuhan-Guangzhou HSR for compressing the Timetable

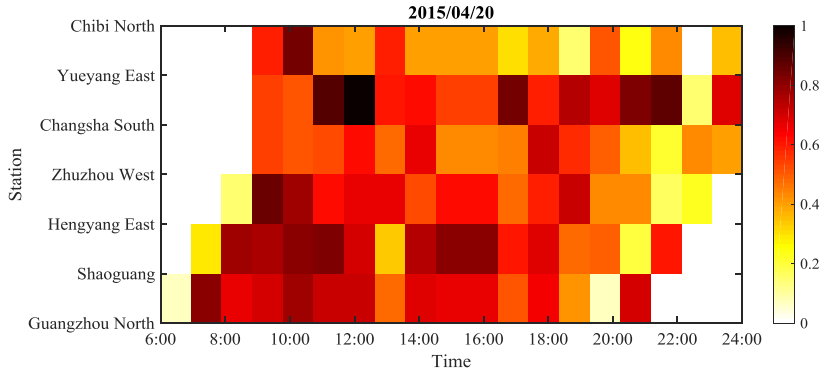
Section No.	Origin station	Destination Station	Length
1	Guangzhou North	Shaoguang	180km
2	Shaoguan	Hengyang East	303km
3	Hengyang East	Zhuzhou West	125km
4	Zhuzhou West	Changsha South	95km
5	Changsha South	Yueyang East	147km
6	Yueyang East	Chibi North	87km

The timetable compressed model based on UIC 406 is applied on each divided section; the capacity consumption of each section in one hour from 6:00 to 23:00 can be calculated according to the compressed timetable results. The capacity consumption results on 1st, August, 2015 are shown in Table 3.

Table 3: The capacity consumption of each section on 1st, August,2015

Time	Section No.1	Section No.2	Section No.3	Section No.4	Section No.5	Section No.6
6:00-7:00	6.67%	0.00%	0.00%	0.00%	0.00%	0.00%
7:00-8:00	63.33%	23.15%	0.00%	0.00%	0.00%	0.00%
8:00-9:00	53.33%	64.69%	13.33%	15.00%	0.00%	0.00%
9:00-10:00	70.00%	60.00%	70.00%	51.67%	75.00%	0.00%
10:00-11:00	65.00%	73.33%	56.40%	56.67%	64.72%	78.33%
11:00-12:00	61.67%	63.33%	66.32%	63.33%	75.30%	45.00%
12:00-13:00	61.67%	73.33%	58.33%	56.67%	72.73%	35.00%
13:00-14:00	59.00%	43.13%	65.00%	40.00%	47.83%	51.25%
14:00-15:00	72.68%	63.02%	57.50%	53.33%	85.00%	34.99%
15:00-16:00	51.67%	59.93%	59.20%	56.67%	81.14%	37.03%
16:00-17:00	68.33%	53.99%	59.57%	51.67%	75.18%	40.00%
17:00-18:00	58.33%	51.67%	53.33%	56.67%	75.00%	40.00%
18:00-19:00	36.67%	51.65%	66.67%	50.00%	56.00%	14.83%
19:00-20:00	16.58%	43.33%	38.33%	56.67%	70.00%	33.33%
20:00-21:00	63.33%	53.33%	40.00%	32.30%	60.00%	25.30%
21:00-22:00	20.00%	62.98%	28.33%	26.67%	50.00%	27.18%
22:00-23:00	0.00%	0.00%	67.47%	58.33%	34.00%	0.00%
23:00-24:00	0.00%	0.00%	0.00%	28.33%	80.00%	41.67%

Heat maps in Figure 4 are used to show a spatial-temporal uneven distribution of capacity consumption on the HSR, respectively for 2015/04/20, 2015/06/20 and 2015/08/01. In the heat map, the horizontal axis stands for the time during one day while the vertical axis is the section. The capacity consumption in each section during different time is measured by the colour area in the figure.



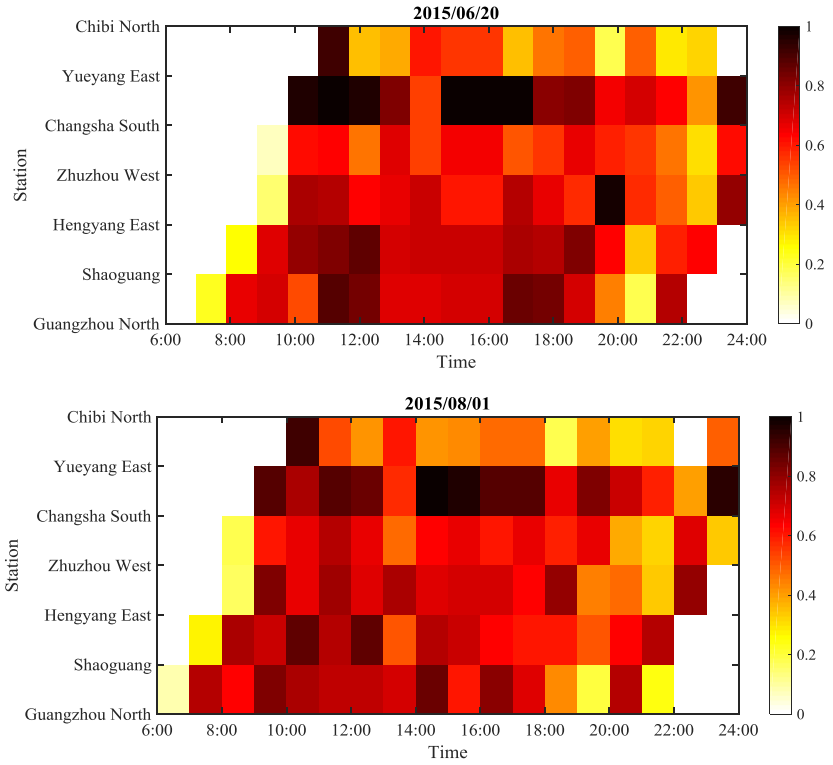


Figure 4: The spatial-temporal distribution of capacity consumption from Guangzhou North station to Chibi North station of each day

The spatial-temporal uneven distribution of capacity consumption is obvious. In terms of spatial uneven distribution, the capacity consumption in the sections from Zhuzhou West station to Changsha North station, from Yueyang East station to Chibi North station is lower, affected by the train stopping plan and the length of sections.

Conversely, the capacity consumption in the section from Changsha South station to Yueyang East station is much higher than other stations, since Shanghai-Chengdu HSR meets Wuhan-Guangzhou HSR at Changsha station and some trains run into the Wuhan-Guangzhou HSR from Shanghai-Chengdu HSR at Changsha South station. Thus, the train operation in the sections from Changsha South station to Yueyang East station is busier, which may lead to a capacity bottleneck.

In terms of temporal uneven distribution, there are peak hours during which trains arrived intensively, leading high capacity consumption. From the view of rail network, the propagation characteristic of peak hours (trains squeeze) at different station might congregate in some blocks which cause a bottleneck in capacity consumption. The peak hour spreads over time at different stations. For each day, the capacity consumption from 8:00 to 13:00 is relatively higher than that from 17:00 to 22:00, which means the train operation in the morning is much busier than that in the afternoon. It should be noticed that there are three peak hours of capacity consumption in the section from Changsha

South station to Yueyang east station, around 11:00, 16:00 and 23:00.

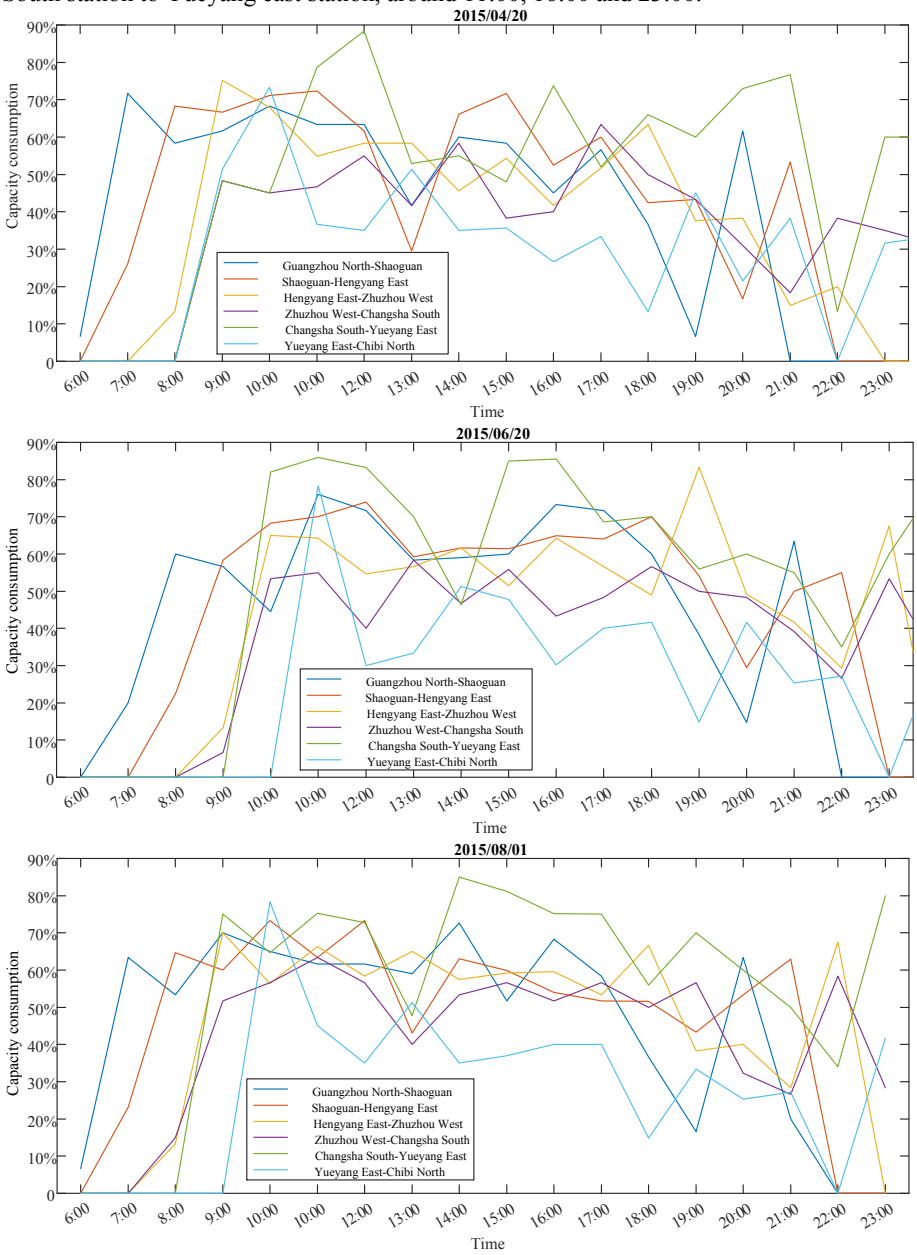


Figure 5: The capacity consumption of each section during various time periods

To examine the temporal-spatial distribution of capacity utilization of Wuhan-Guangzhou HSR, the capacity consumption in different segment and time period are

detailed investigated, as depicted in Figure 5. In Figure 5, the horizontal axis stands for the time and the vertical axis is the capacity consumption. The capacity consumption in each section is calculated every hour, from 6:00 to 23:00.

From an overall perspective, the capacity utilization on each day is similar. In terms of temporal uneven distribution, the capacity consumption in one day shows an increase trend from 6:00 to 9:00, and then kept steady from 9:00 to 18:00, after 18:00 the capacity consumption decreased.

Specifically, the capacity consumption during the time period from 9:00 to 17:00, about 55%, is higher than other time period. More trains are scheduled to operate during these time period to satisfy the passengers demand. The capacity consumption from 18:00 to 20:00 shows a medium level, about 35%. During the early hours every day, from 7:00 to 8:00, it is obvious to see a decrease trend on the capacity consumption in the section from Guangzhou North station to Hengyang East station since most of the up-direction trains originated from Guangzhou North station and the departure interval in the morning is short. Similarly, there is an increasing trend on capacity consumption in the section from Changsha South station to Yueyang East station and the section from Yueyang East station to Chibi North station during the time period from 22:00 to 23:00 since there are many trains arriving or passing by these stations at this time.

When it comes to the spatially uneven distribution of capacity consumption, the capacity consumption in the segment from Changsha South station to Yueyang East station is relatively higher than other sections during most of the time period, and segment is easy to be a bottleneck of the line due to the high capacity consumption. And the capacity consumption in the segment from Yueyang East station to Chibi North station is lower and more trains can be scheduled in this segment

Table 4: The capacity consumption of each divided section on 1st,August,2015

Time	Shaoguan-Hengyang East	Shaoguan-Chenzhou West	Chenzhou West-Hengyang East
6:00-7:00	0.00%	0.00%	0.00%
7:00-8:00	23.15%	20.00%	6.67%
8:00-9:00	64.69%	49.44%	58.33%
9:00-10:00	60.00%	55.00%	64.72%
10:00-11:00	73.33%	66.67%	58.99%
11:00-12:00	63.33%	63.33%	73.33%
12:00-13:00	73.33%	66.67%	58.07%
13:00-14:00	43.13%	41.67%	56.67%
14:00-15:00	63.02%	56.67%	58.08%
15:00-16:00	59.93%	71.62%	73.33%
16:00-17:00	53.99%	54.75%	53.33%
17:00-18:00	51.67%	51.67%	58.33%
18:00-19:00	51.65%	47.05%	44.66%
19:00-20:00	43.33%	41.67%	42.74%
20:00-21:00	53.33%	35.00%	13.33%
21:00-22:00	62.98%	61.35%	58.33%
22:00-23:00	0.00%	0.00%	20.00%
23:00-24:00	0.00%	0.00%	0.00%

The UIC 406 method is only able to calculate capacity consumption for line sections,

and not for either the entire railway network or railway lines. It is necessary to divide the network into smaller line sections, which can be handled separately by the UIC 406 capacity method. The division of the lines into sections is of major importance for the results of capacity consumption, which is specially analysed below.

In the paper above, the timetable compressed model is applied on the section from Shaoguan station to Hengyang east station. In this part, this section is divided into two sections, one is from Shaoguan station to Chenzhou West station and another is from Chenzhou West station to Hengyang East station. The capacity consumptions of the two sections are calculated, shown in the Table 4.The capacity consumption of the three sections varies greatly. For a better understand of the difference, the real-record train lines and the compressed train lines in several hours of different stations are shown in Figure 5. The train occupied time in the section responsible to the origin and destination stations of the section, as well the train stop plan and overtaking in the media stations. For instance, the trains involved in the section from Shaoguan station to Hengyang East station and the section from Shaoguan station to Chenzhou West station are the same as compressing the timetable. However, the length of the section and the train operation in the sections are not the same, the capacity consumptions are different of the two sections.

Therefore, the division of the lines into sections is of major importance for the results of capacity consumption.

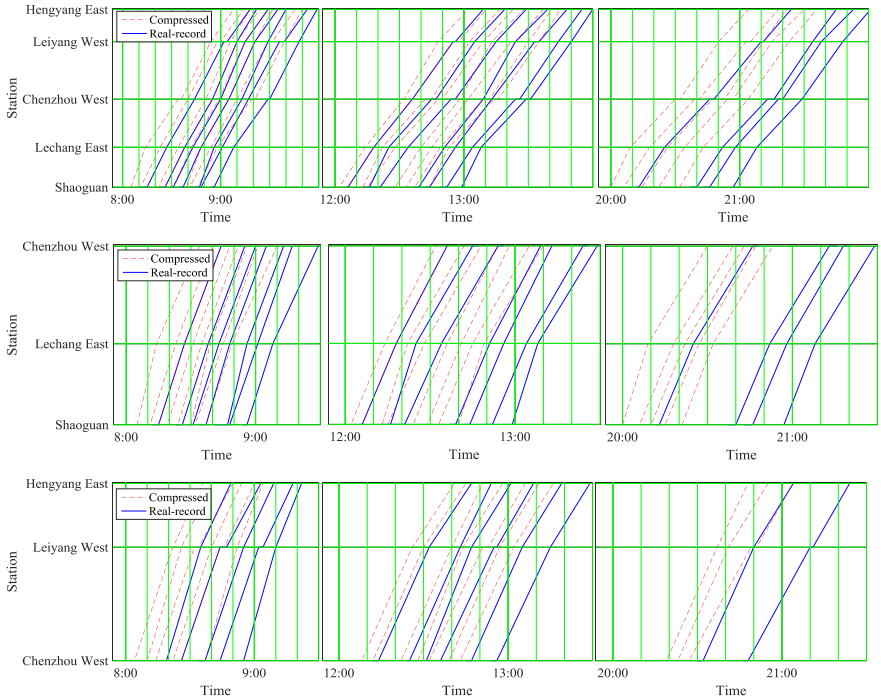


Figure 5: The compressed train line of each divided section on 1st, August, 2015

4 Conclusions

This paper analyses how the UIC 406 method is expounded in China. The results show

that it is possible to use UIC 406 method and real-record timetable and train operation data to calculate the capacity consumption and identify the bottleneck in a line.

A optimal method based on UIC 406 are proposed to compress the timetable for a practical capacity consumption, with respect to train orders, overtaking and crossing have been defined on the given timetable. The method was applied on the Wuhan-Guangzhou HSR. The capacity consumption of each section on Wuhan-Guangzhou HSR is calculated based on the real-record timetable. Then the capacity consumption and bottlenecks are analysed and identified. It can be concluded that the temporal-spatial uneven of capacity consumption is obvious. The capacity consumption in the early times during one day is higher; the section from Guangzhou South station to Yueyang East station is easy to be a bottleneck due to the layout of the HSR.

Besides, the analysis has shown that the capacity consumption on railway lines is very responsive to the line and section examined. Therefore, the capacity consumption should only be compared relatively. Apart from that, the division of the lines into sections is of major importance for the results of capacity consumption. Statements of the degree of capacity consumption in a line need to be based on a consistent division into line sections.

It seems the train stop plan, dwell time and cross plan has an influence of capacity consumption, which will be discussed detailed in the further study.

Acknowledgements

This work was supported by the National Nature Science Foundation of China [grant number 71871188 and U1834209] and National Key R&D Program of China [grant number 2017YFB1200700]. We acknowledge the support of SAPIENZA Università di Roma and the China Scholarship Council. We are grateful for the contributions made by our project partners.

References

Landex, A., Kaas, A. H., Schittenhelm, B., et al. 2006. “Practical use of the UIC 406 Capacity Leaflet by including Timetable Tools in the Investigations”, In *Practical use of the UIC 406 capacity leaflet by including timetable tools in the investigations*,pp.643-652.WIT Press, Southampton.

Landex, A., Schittenhelm B., Kaas A. H., et al. 2008. “Capacity measurement with the UIC 406 capacity method”, In: *Proceedings of The 7th International Conference on Computer System Design and Operation in the Railway and Other Transit Systems (COMPRAIL2008)*, Toledo, Spain

Landex, A., 2009. “Evaluation of railway networks with single track operation using the uic 406 capacity method”. *Networks and Spatial Economics*, vol.9, pp.7-23.

Landex, A., 2008.Methods to estimate railway capacity and passenger delays. Ph.D.Thesis, Department of Transport, Technical University of Denmark.

Leaflet UIC 406-Capacity, 2004, International Union of railways, Paris.

Lindner, T., 2011. “Applicability of the analytical uic code 406 compression method for evaluating line and station capacity”. *Journal of Rail Transport Planning and Management*, vol.1,pp.49-57.

Pavlides, A.,Chow, A. 2016. “Cost functions for mainline train operations and their application to timetable optimization”. In: *Proceedings of The Transportation Research Board 95th Annual Meeting(TRB2016)*, Washington, D.C, United States

Whalborg, M. 2004. Banverket experience of capacity calculations according to the uic capacity leaflet. *Publication of Wit Press*.
Yang, L., Qi. J., Li, S., et al. 2015. “Collaborative optimization for train scheduling and train stop planning on high-speed railways” . *Omega*, vol.64, pp.57-76.

Optimal Train Service Design in Urban Rail Transit Line with Considerations of Short-Turn Service and Train Size

Zhengyang Li [a,b], Jun Zhao [a,b,1], Qiyuan Peng [a,b]
[a] School of Transportation and Logistics, Southwest Jiaotong University, Chengdu, Sichuan 611756, China
[b] National United Engineering Laboratory of Integrated and Intelligent Transportation, Southwest Jiaotong University, Chengdu, Sichuan 611756, China

[1] Corresponding author, E-mail: junzhao@swjtu.edu.cn

Abstract
The train service scheme of an urban rail transit line specifies information such as the total number of train services operated in the line, and the associated turn-back stations, train size and frequency of each service. A reasonable train service scheme can provide satisfactory services for passengers and reduce the operational cost for operators. This paper focuses on the optimal train service design problem in an urban transit line, where both the short-turn services and the train size of each service are considered. A service network based on a given pool of candidate train services with provided turn-back stations is constructed. The optimal strategy is used to assign passenger flows on the service network so as to describe the service choice behaviour of passengers between different train services. Considering many operational and capacity constraints, a mixed integer nonlinear programming model minimizing the sum of the operators' cost and passengers' waiting time cost is developed to identify train services from the service pool and determine the train size and frequency of each chosen service. The nonlinear model is transformed into a linear one, and two simplification methods named service network simplification and OD pair aggregation are proposed to improve further the computational efficiency of the model. Finally, realistic instances from Chongqing Rapid Rail Transit Line 26 in China are used to test the proposed approaches. The results show that our approach can effectively reduce the operators' cost and the passengers' waiting time cost compared with the empirical method frequently used in practice.

Keywords
Urban rail transit, train service design, short-turn service, train size, service network, optimal strategy

1 Introduction

The train service scheme plays an important role in the operation of urban rail transit. It contains information such as the total number of train services in a line, the turn-back stations, train size and frequency of each designed service. The train service scheme affects the waiting time and transfer time of passengers, and it also determines the number of rolling stocks and crews that required to operate an urban rail transit line. A reasonable train service scheme can match the transport capacity with the passenger demand of a line, leading to a reduction of the passengers' waiting time and the operation cost of the line.

At present, using a single full-length train service and determining the associated train size and train frequency according to the maximum passenger load of sections is still a

frequently used method to obtain the train service scheme in an urban rail transit line. In this way, the train frequency is the same in every section of a line. However, this empirical method does not take the possible unbalanced spatial distribution of passenger demand in urban rail transit lines into account, resulting in waste of transport capacity in some sections and congestion in other sections. A commonly used strategy to overcome this problem is to insert short-turn services into the train service scheme. Short-turn services enable different sections to have different train frequencies, where the transport capacity can match the passenger demand better. However, the setting of short-turn services will cause some passengers to transfer, and passengers' service choice behaviour is complex in the case of multiple train services, which complicates the train service scheme design. Therefore, it is necessary to analyse the impact of train services on passengers and operators, and optimize the train service scheme with multiple services to improve the service quality and reduce the operation cost of urban rail transit lines.

This paper tries to optimize the train service scheme in an urban rail transit line where multiple train services consisting of either full-length or short-turn ones can be operated, and the train size and frequency of each service have to be determined. In order to accurately describe the service choice behaviour of passengers between different train services, a service network based on a given pool of candidate train services with specified turn-back stations is constructed, and the optimal strategy proposed by Spiess (1989) is used to assign passenger flows on the service network. A mixed integer nonlinear programming model minimizing the sum of the operators' cost and passengers' waiting time cost is developed to identify train services from the service pool and determine the train size and frequency of each chosen service. Then, the nonlinear model is transformed into a linear one which is further simplified by two methods, enabling the linear model to be solved quickly by commercial optimization solver CPLEX. Finally, instances based on the Chongqing Rapid Rail Transit Line 26 in China are constructed to test the proposed approaches.

The remainder of this paper is structured as follows. Section 2 gives an overview of the related literature. In Section 3, we present the problem description and assumptions. In Section 4, a mixed integer nonlinear programming model is formulated to represent the train service design problem in an urban rail transit line. Section 5 presents a linearization method and two simplification techniques to obtain a simplified mixed integer linear programming model. Section 6 provides our computational experiments on Chongqing Rapid Rail Transit Line 26 in China. Conclusions and future research works are discussed in Section 7.

2 Literature Review

To the best of our knowledge, our problem has not been completely investigated in the literature. Related works mainly focus on the timetable design with short-turn services or multiple vehicle sizes, while few works study the service design in public transport systems especially in urban rail transit lines.

There are few works focusing on obtaining the timetable of public transport systems with short-turn services or multiple vehicle sizes. Furth (1987) considered that the schedule coordination between the full-length trip and short-turn trip is necessary, and proposed an offset schedule algorithm to minimize the bus fleet size and to save the operation cost. Ceder (1989) proposed a two-stage optimization method to obtain the location of turn-back stations and the bus fleet size. Zhang (2018) developed a mixed integer linear programming model to optimize the timetable of an urban rail transit line with short-turn strategy and multiple depots. With the consideration of multiple vehicle sizes, Ceder and Hassold (2011) proposed a multi-objective methodology to create even load and even headway bus

timetables by operating different bus sizes. Hassold and Ceder (2012) presented approaches to use multiple vehicle sizes to improve the matching between bus timetable and passenger demand. Chen (2019) considered the design of headway and vehicle capacity simultaneously with the usage of modular vehicles, aiming to better match the dynamic passenger demand with transit services.

Other works mainly studied the transit service design problem with short-turn services in either bus corridors or urban rail transit lines. Delle Site and Filippi (1998) focused on the short-turn strategies with different bus sizes within multiple operating periods, and developed a net benefit maximization model for optimizing the bus sizes, service frequencies and fares. Tirachini (2011) developed a model to optimize the short-turn trip in a single period by analytical expressions, aiming at increasing the bus frequency in congestion sections. Cancela (2015) considered the interests of both operators and users, and used the optimal strategy proposed by Spiess (1989) to develop a mixed integer linear model to solve the bus routes design problem. Ji (2016) used a Markov model to describe the seats searching behaviour of passengers during their trip, and proposed a model to optimize the schedule coordination between full-length and short-turn bus services. Sun (2016) relaxed the assumption that a full-length service must be operated, and proposed a flexible short-turn service design model to minimize the operators' cost and passengers' waiting time in subways. Yang (2017) developed a bilevel model to design the short-turn strategy on a bus route. Ding (2018) relaxed the constraints of turn-back stations in metro systems, and proposed a nonlinear programming model to design the short-turn services.

A review of existing studies can be summarized as follows:

(1) Most of the previous studies focus on the timetable optimization in both bus and rail systems, while few works focus on the service design problem in transit lines, especially in urban rail transit lines. Compared with the bus systems, urban rail transit lines have more restrictions on setting train services, especially the capacity limitation of turn-back stations.

(2) Due to the difficulty of depicting the services setting on either bus or rail lines and the complexity of analysing the service choice behaviour of passengers in the case of multiple services. Almost all the former works follow an assumption that a full-length service must exist, and the problem is to obtain the optimal parameters of setting one short-turn service. But, when additional services are considered, many existing models become intractable. At present, there are already some urban rail transit lines which have more than two train services, such as Shanghai Metro Line 2 (3 services) and Chongqing Rail Transit Line 3 (3 services) in China. Therefore, it is necessary to continue to study the optimization method for multiple train services design in urban rail transit lines.

(3) Under the assumption that a full-length service exists, most of the initial works assume that passengers can always take a direct service to their destination when designing the service scheme. Few studies have been done to describe passengers' service choice behaviour during their trip, especially in the case of multiple services.

(4) Relative works mainly consider the usage of multiple vehicle sizes in timetable optimization. But in the aspect of service design, most of the existing studies only design short-turn services whose vehicle size is pre-determined. Actually, the train size interacted with the frequency is an important parameter of services which affects the capacity and operation cost of urban rail transit lines. Joint design of the train size and frequency of services could lead to better solutions.

3 Problem Description and Assumptions

3.1 Problem Description

Without loss of generality, we consider an urban rail transit line with several turn-back stations, multiple services and multiple train sizes. Taking an urban rail transit line with 7 stations and 3 turn-back stations shown in Figure 1 as an example. The stations are denoted as v_1 to v_7 in the upward direction, while the sections are denoted as e_1 to e_6 in the upward direction. Stations v_1, v_4 and v_7 are turn-back stations on the line that can reverse the running direction of trains. Restricted by the layout of the turn-back tracks, stations v_1 and v_7 can only reverse the running direction for trains from one direction (v_1 can only switch the running direction of trains from downward to upward, and v_7 can only switch the running direction of trains from upward to downward), while station v_4 can reverse the running direction for trains from both directions.

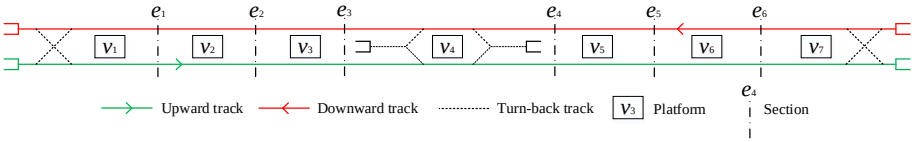


Figure 1: Sketch map of an urban rail transit line

According to the given location of turn-back stations on the line and the feasible reversing directions of each turn-back station, a pool of candidate train services can be generated. The line shown in Figure 1 can operates 3 train services shown in Figure 2. Service 1 has stations v_1 and v_7 as its turn-back stations. Service 2 has stations v_1 and v_4 as its turn-back stations. Service 3 has stations v_4 and v_7 as its turn-back stations.

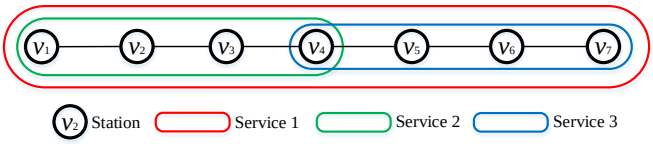


Figure 2: Illustration of train services

In an urban rail transit line, after the construction of candidate train services, the train service design problem is to determine the train services operated on the line, and the train size and frequency of each selected train service. To design the train service scheme, we need to take into account the operation cost changes between different schemes as well as the impacts of schemes on the travel process of passengers. In addition, a feasible train service scheme must consider several operational and capacity constraints, e.g. coverage of stations and sections, passenger load in sections, capacity of turn-back stations, operational rules of train services, etc. Therefore, the train service design problem of an urban rail transit line is essentially to find the optimal combination of train services, and the train size and frequency of each selected service under the constraints of line condition and passenger demand.

Formally, in an urban rail transit line, given the layout of the line, capacity of turn-back stations, minimal and maximum frequency in sections, passenger load in sections and other operational rules of train services, the train service design problem on an urban rail transit line is to determine the train service set on the line, the train size of each selected service

and the corresponding train frequency such that all operational and capacity constraints are respected, while both the operators' cost and passengers' waiting time cost are minimized.

3.2 Assumptions

To simplify the model formulation, the following assumptions are introduced.

(1) Passengers' behaviour accords with the principle of optimal strategy, that is, at each station, each passenger boards the first train passing the station, which can transport him/her close to his/her destination station.

(2) At each station, passengers arrive uniformly and trains arrive on timetable.

(3) Trains are used independently in each service. Only one train composition is allowed in each service. Different train compositions can be arranged to different services.

(4) Trains of each service are assumed to stop at each station of the service route.

4 Model Formulation

4.1 Service Network Construction

According to the given layout especially the turn-back stations of an urban rail transit line, all candidate train services that are allowed to be operated on the line can be generated in advance. Based on which, a directed service network is introduced to design the train services and describe the travel process of passengers in the studied line. Taking the line in Figure 1 as an example, as there are 3 candidate train services in the line, a directed service network shown in Figure 3 can be formed.

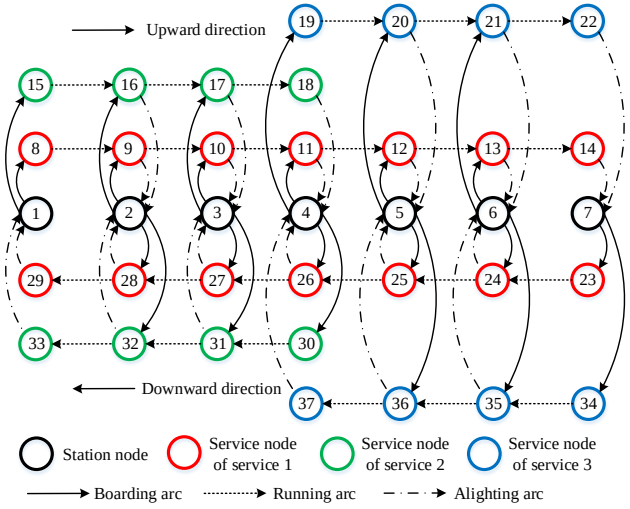


Figure 3: Representation of directed service network

The node set of the service network consists of two parts, including station nodes 1 to 7 and train service nodes 8 to 37. Station nodes 1 to 7 represent the stations v_1 to v_7 in the line, while train service nodes indicate the stations covered by all candidate services. Service 1 covers stations v_1 to v_7. The red nodes 8-14 and 23-29 with respect to Service

1 cover station nodes 1 to 7 in parallel. Service 2 covers stations v_1 to v_4. The green nodes 15-18 and 30-33 with regard to Service 2 cover station nodes 1 to 4 in parallel. Service 3 covers stations v_4 to v_7. The blue nodes 19-22 and 34-37 corresponding to Service 3 cover station nodes 4 to 7 in parallel. Among the train service nodes, nodes 8 to 22 are in the upward direction, and nodes 23 to 37 are in the downward direction.

The arc set is composed of three parts including the boarding arcs, running arcs and alighting arcs. Boarding arcs connect the station nodes to its corresponding train service nodes, indicating the boarding process of passengers from stations to the trains of each service in each direction. In both directions, running arcs sequentially link the train service nodes of each service, expressing the running process of trains through sections in each direction of each service. Alighting arcs are from the train service nodes to the corresponding station nodes, representing the alighting process of passengers from the trains of each service in each direction to the corresponding stations. Note that in both directions, boarding arcs do not connect the station node to the last node of each service, and alighting arcs do not connect the first node of each service to the corresponding station node. For instance, in the upward direction, there is no a boarding arc between node 4 and node 18, because the trains of Service 2 cannot take the passengers from station v_4 to the upward direction. Also, there is no an alighting arc between node 19 and node 4, because station v_4 is the origin station of Service 3 to the upward direction and the trains of Service 3 are empty when originating from station v_4 due to that no passengers need to alight.

After abstracting the considered urban rail transit line into a directed service network, the travel process (including boarding, in-vehicle running, transferring and alighting) of each passengers OD pair on the line can be expressed by a path from its origin station node to its destination station node in the service network. For example, the journey of passengers in OD pair $(1, 6)$ from station v_1 to station v_6 on the line in Figure 1 can be represented by a path from station node 1 to station node 6 in the service network of Figure 3 as follows:

(1) If passengers take the trains of Service 1 directly from station v_1 to station v_6, their travel process can be expressed by the path 1-8-9-10-11-12-13-6 in the service network.

(2) If passengers firstly take the trains of Service 2 from station v_1 to station v_4, and then transfer to trains of Service 3 until arriving at station v_6, the travel process is represented by the path 1-15-16-17-18-4-19-20-21-6 in the service network.

(3) If passengers from station v_1 to station v_6 sequentially take the trains of Service 2 and Service 1 with a transfer at station v_4, their travel process can be indicated by the path 1-15-16-17-18-4-11-12-13-6 in the service network.

In order to analyse the impact of different train service schemes on the service choice behaviour of passengers, not only the travel process of passengers on the line but also the distribution of passenger flows on each arc of the service network should also be determined. Here we adopt the optimal strategy proposed by Spiess (1989) to assign the passengers of each OD pair on the service network. Under the optimal strategy, passengers take the first train which can transport them close to their destination station. Due to that, at any station node, the probability that a passenger chooses a boarding arc originated from this station node to the travel direction of the passenger is the ratio of the frequency of the service corresponding to the boarding arc to the total frequency of all boarding arcs originating from this station node to the travel direction of this passenger. With the optimal strategy, the distribution of each OD pair on the service network is obtained. Thus, the impact of different train service schemes on the trip decisions of passengers can be quantitatively analysed.

4.2 Notation

The notation to be used in the model is provided in Table 1.

Table 1: Definition of sets, parameters and decision variables

Notation	Description
V	Set of stations of an urban rail transit line with index v.
E	Set of sections with index e.
L	Set of candidate train services with index l.
T	Set of optional train sizes with index t.
D	Set of running directions with index d.
F	Set of optional trains with index f.
N	Set of nodes in the service network with index n, $N = \{N_1, N_2\}$.
N_1	Set of station nodes.
N_2	Set of train service nodes.
A	Set of arcs in the service network with index a, $A = \{A_1, A_2\}$.
A_1	Set of boarding arcs, $A_1 = \{A_{1d} \mid d \in D\}$.
A_{1d}	Set of boarding arcs in direction d.
A_2	Set of running and alighting arcs.
A_{n+}	Set of outgoing arcs from node n.
A_{n-}	Set of incoming arcs to node n.
λ_n	Corresponding station of node n in the service network.
τ_a	Corresponding train service of arc a in the service network.
h	Time span of the study period, unit: minute.
c_t^1	Fixed cost of a train with size t within the study period.
c_t^2	Operating cost of a train per kilometre with train size t.
σ	Passenger's waiting time cost per hour.
m_l	Round-trip time of train service l.
g_l	Round-trip distance of train service l.
α_{lv}	0-1 parameters, if service l covers station v, α_{lv}=1, 0 otherwise.
β_{le}	0-1 parameters, if service l covers section e, β_{le}=1, 0 otherwise.
γ_{lvd}	0-1 parameters, if trains in direction d of service l turn back at station v, γ_{lvd}=1, 0 otherwise.
q_{ij}	Volume of OD pair (i,j) from station i to station j within the study period.
θ_{ij}	Travel direction of OD pair (i,j), $\theta_{ij} \in D$.
p_{ed}	Passenger load at section e in direction d within the study period, which can be obtained from the passenger OD demand.
r_t	Capacity of a train with size t.
δ	Required surplus of transport capacity in sections, unit: %.
Ω	Maximum allowable number of train services on the studied line.
φ	Minimum train frequency of each service.
s_{vd}	Turn-back capacity of station v in direction d within the study period.
b_e^{min}	Minimal train frequency requirement of section e within the study period.
b_e^{max}	Maximal train frequency limitation of section e within the study period.
x_{lt}	0-1 variables, if train size t is adopted on service l, x_{lt}=1, 0 otherwise.
y_{ltf}	0-1 variables, if the fth train with size t on service l is operated, y_{ltf}=1, 0 otherwise. Thus, the frequency of trains with size t on service l is $\sum_{f \in F} y_{ltf}$.
k_a^{ij}	Continuous variables, represents the volume (number of passengers) of OD pair (i,j) on arc a of the service network.
w_n^{ij}	Continuous variables, represents the total passenger waiting time of OD pair (i,j) at station node n of the service network.

4.3 Operator's Cost

The operator's cost of an urban rail transit line contains the fixed cost and operation cost. Fixed cost refers to the cost of purchasing rolling stocks used in the line. The rolling stocks of an urban rail transit line are not only operated in the study period, and the life cycle of a rolling stock is typically 30 years. Therefore, to define comparable cost items, the fixed cost of operators in purchasing rolling stocks is apportioned by the time span of the study period. Meanwhile, operation cost refers to the total operating cost of all trains running on the line during the study period, which depends on the round-trip distance of all chosen train services. The fixed cost and operation cost of operators are formulated as follows:

$$Z_1 = \sum_{l \in L} \sum_{t \in T} \sum_{f \in F} \frac{c_t^1 m_l y_{ltf}}{h} \tag{1}$$

$$Z_2 = \sum_{l \in L} \sum_{t \in T} \sum_{f \in F} c_t^2 g_l y_{ltf} \tag{2}$$

Objective (1) represents the fixed cost of operators. Note that the number of rolling stocks required by each train service equals to the round-trip time of the service multiplied by the frequency of the service. Objective (2) indicates the operation cost of operators.

4.4 Passengers' Waiting Time Cost

The travel time of passengers in an urban rail transit line is mainly composed of four parts including the access time, origin/transfer waiting time, in-vehicle time and egress time. The train service scheme rarely impacts the access time and egress time of passengers. Meanwhile, trains with different sizes always have the same running time in each section of the line, due to that different train service schemes make no difference to the in-vehicle time of passengers. However, under multiple train services, the origin, destination and frequency of operated services are different, which could lead to different waiting time of passengers. Therefore, we focus on minimizing the waiting time cost of passengers.

In theory, the waiting time of passengers at each station to each direction is the average waiting time of passengers multiplied by the number of passengers at the station to the direction. Following the optimal strategy in Spiess (1989), assume that passengers arrive at stations uniformly and each passenger boards the first train which can transport them close to their destination station, i.e. passengers show no preference for the choice of services. Moreover, in the practical operation of urban rail transit lines, the arrival of the trains regulated by timetable at each station to each direction is deterministic and even. Therefore, the expected average waiting time of passengers at a station node to a direction is half of the combined arrival interval of the services corresponding to all boarding arcs of the station to the direction.

The expected average waiting time of passengers at station node n to direction d is:

$$E(wt) = \frac{h}{2 \sum_{a \in A_{n+} \cap A_{1d}} \sum_{t \in T} \sum_{f \in F} y_{\tau_a tf}} \tag{3}$$

Meanwhile, the number of passengers at a station node to a direction is the sum of volume of OD pairs on all boarding arcs of the station to the direction, which can be obtained by the passenger flow distribution on the service network. Thus, the waiting time cost of passengers on the whole line is:

$$Z_3 = \frac{\sigma}{60}\sum_{i\in V}\sum_{j\in V}\sum_{n\in N_1}\frac{h\sum_{a\in A_{n+}\cap A_{1\theta_{ij}}} k_a^{ij}}{2\sum_{a\in A_{n+}\cap A_{1\theta_{ij}}}\sum_{t\in T}\sum_{f\in F} y_{\tau_a tf}} \tag{4}$$

4.5 Basic Model

Based on the above modelling, the original compound objective function to minimize the operators' cost and passengers' waiting time cost is as follows:

$$\begin{aligned}\min \sum_{l\in L}\sum_{t\in T}\sum_{f\in F}\frac{c_t^1 m_l y_{ltf}}{h} + \sum_{l\in L}\sum_{t\in T}\sum_{f\in F} c_t^2 g_l y_{ltf} \\ + \frac{\sigma}{60}\sum_{i\in V}\sum_{j\in V}\sum_{n\in N_1}\frac{h\sum_{a\in A_{n+}\cap A_{1\theta_{ij}}} k_a^{ij}}{2\sum_{a\in A_{n+}\cap A_{1\theta_{ij}}}\sum_{t\in T}\sum_{f\in F} y_{\tau_a tf}}\end{aligned} \tag{4}$$

Note that the third part of Objective (4) which represents the waiting time cost of passengers is a nonlinear representation which will cause difficulties in solving the model.

We utilize the optimal strategy proposed in Spiess (1989) to assign passengers in the service network such that a linear representation of the passengers' waiting time can be obtained. Then, we formulate our problem as the following basic and incomplete model:

$$\min \sum_{l\in L}\sum_{t\in T}\sum_{f\in F}\frac{c_t^1 m_l y_{ltf}}{h} + \sum_{l\in L}\sum_{t\in T}\sum_{f\in F} c_t^2 g_l y_{ltf} + \frac{\sigma}{60}\sum_{i\in V}\sum_{j\in V}\sum_{n\in N_1} w_n^{ij} \tag{6}$$

$$\text{s.t.} \quad \sum_{a\in A_{n+}} k_a^{ij} - \sum_{a\in A_{n-}} k_a^{ij} = \begin{cases} q_{ij}, & \lambda_n = i, n\in N_1, \forall i\in V, j\in V \\ -q_{ij}, & \lambda_n = j, n\in N_1, \forall i\in V, j\in V \\ 0, & n\in N_2, \forall i\in V, j\in V \end{cases} \tag{7}$$

$$k_a^{ij} \le \frac{2}{h} w_n^{ij}\sum_{t\in T}\sum_{f\in F} y_{\tau_a tf}, \quad \forall a\in A_{n+}\bigcap A_{1\theta_{ij}}, \forall n\in N_1, \forall i\in V, j\in V \tag{8}$$

$$k_a^{ij} = 0, \quad \forall a\in A_{n+}\bigcap A_{1\theta_{ji}}, \forall n\in N_1, \forall i\in V, j\in V \tag{9}$$

$$k_a^{ij} \ge 0, \quad \forall a\in A, \forall i\in V, j\in V \tag{10}$$

$$w_n^{ij} \ge 0, \quad \forall n\in N_1, \forall i\in V, j\in V \tag{11}$$

Objective function (6) minimizes the operators' cost and the passengers' waiting time cost. Constraints (7) are the flow conservation of OD pairs, which assure that each OD pair can be routed from its origin to its destination in the service network. Constraints (8) are the optimal strategy requirements imposed on decision variables y_{ltf}, k_a^{ij} and w_n^{ij}. These constraints mean that for each OD pair (i,j), the number of passengers of this pair k_a^{ij} on each boarding arc a originating from each station node n is not greater than the quotient between the waiting time of all passengers of this pair w_n^{ij} at station node n and 1/2 of the arrival interval of the service $h/\sum_{t\in T}\sum_{f\in F} y_{\tau_a tf}$ in boarding arc a. Constraints (8) are specially proposed to ensure that the assignment of OD pairs satisfy the optimal strategy, and it can help to obtain a linear representation of the waiting time cost of passengers in the objective function. Constraints (9) indicate that passengers do not take trains which would carry them away from their destination. In other words, there are no flows on the boarding arcs to the opposite travel direction of each OD pair. Constraints (10) and (11) are the non-negative constraints of decision variables.

The basic model (6) to (11) mainly conducts the passenger assignment in the service network under the optimal strategy, aiming to obtain the distribution of each OD pair on the

service network and compute the total waiting time cost of passengers. The operator's fixed cost and operation cost are computed and other operational and capacity requirements are respected by incorporating the following additional constraints.

4.6 Additional Constraints

4.6.1 Covering of stations and sections

The train services operated on an urban rail transit line need to cover each station and each section of the line. These requirements are as follows:

$$\sum_{l\in L}\sum_{t\in T}\alpha_{lv}\,x_{lt}\geq 1,\qquad \forall v\in V \tag{12}$$

$$\sum_{l\in L}\sum_{t\in T}\beta_{le}\,x_{lt}\geq 1,\qquad \forall e\in E \tag{13}$$

Constraints (12) assure that each station is covered by at least one train service. Constraints (13) indicate that each section is covered by at least one train service.

4.6.2 Passenger load in sections

For each section of the line, the transport capacity should not be less than the passenger load of the section in both directions. That is, the passenger load of each section in each direction should be satisfied by the combined transport capacity supplied by all selected train services. In addition, in order to ensure the comfortableness of passengers and deal with the occasional large passenger flow on the line, the operator of an urban rail transit line usually holds a transport capacity surplus δ when designing the train service scheme. We have:

$$(1-\delta)\sum_{l\in L}\sum_{t\in T}\sum_{f\in F}\beta_{le}r_t y_{ltf}\geq \max_{d\in D}\{p_{ed}\},\qquad \forall e\in E \tag{14}$$

4.6.3 Operational rules on train services

In the daily operation of an urban rail transit line, too many train services operated on the line, a service with different train sizes and a service with a very low frequency will not only increase the operation complexity of the line, but also not be conducive to the travel experience of passengers. Hence, the number of operated train services, the number of train sizes on a service and the minimum frequency of each service are restricted as follows:

$$\sum_{l\in L}\sum_{t\in T}x_{lt}\leq \Omega \tag{15}$$

$$\sum_{t\in T}x_{lt}\leq 1,\qquad \forall l\in L \tag{16}$$

$$\sum_{t\in T}\sum_{f\in F}y_{ltf}\geq \varphi\sum_{t\in T}x_{lt},\qquad \forall\, l\in L \tag{17}$$

$$\sum_{f\in F}y_{ltf}\leq |F|x_{lt},\qquad \forall\, l\in L,\forall t\in T \tag{18}$$

Constraints (15) ensure that the number of train services operated on the line does not exceed the upper limit Ω. Constraints (16) require that each service can choose at most one type of train size. Constraints (17) assure that the frequency of a train service is not less than the minimum value φ if this service is operated. Constraints (18) are the relationship between variables x_{lt} and y_{ltf}, indicating that the trains with size t can be operated on

service l only if this train size is adopted on the service.

4.6.4 Capacity of stations and sections

Generally, only a few stations on an urban rail transit line have turn-back facilities, and there may also be direction limit on the switch operation at each turn-back station. In addition, due to the track layout and other infrastructure facilities at turn-back stations, the number of trains that can be switched at each turn-back station in each direction within the study period is restricted by an upper limit s_{vd}. Thus, the capacity constraint of turn-back stations in the line can be formulated as follows:

$$\sum_{l\in L}\sum_{t\in T}\sum_{f\in F}\gamma_{lvd}\,y_{ltf} \le s_{vd}, \qquad \forall v \in V, \forall d \in D \tag{19}$$

Influenced by the train control system, on an urban rail transit line, the tracking headway between two adjacent trains to the same direction cannot be lower than a specified minimum value. Thus, during the study period, the frequency of trains in each section of the line should not be greater than the maximum frequency of the section, i.e. b_e^{max}. Besides, in order to avoid passengers from waiting too long at some stations, the maximum headway in each section are limited when designing the train service scheme for an urban rail transit line. That is, during the study period, the frequency in each section of the line should not be less than the minimum frequency of the section, i.e. b_e^{min}. Thus, the minimum and maximum frequency in sections of the line are satisfied by:

$$b_e^{min} \le \sum_{l\in L}\sum_{t\in T}\sum_{f\in F}\beta_{le}y_{ltf} \le b_e^{max}, \qquad \forall e \in E \tag{20}$$

4.6.5 Valid inequalities

In this paper, we introduce 0-1 variable y_{ltf} to indicate whether the fth train with size t on service l is operated. Thus, the frequency of trains with size t on service l is equal to the sum of a group of y_{ltf}, i.e. $\sum_{f\in F} y_{ltf}$. This will lead to a large quantity of feasible combinations of y_{ltf} under the same frequency for any service l and any train size t, heavily aggravating the symmetry in the model. For instance, if in total 10 trains of size t is operated on service l (i.e. $\sum_{f\in F} y_{ltf} = 10$) and the maximum frequency of trains is 20 (i.e. $|F| = 20$), the number of feasible combinations of y_{ltf} is C_{20}^{10}.

The symmetry of variable y_{ltf} can be easily broken by requiring that the $(f+1)$th train with size t on service l can be operated only if the corresponding fth train with the same size is operated. These valid inequalities are formulated as follows:

$$y_{ltf} \ge y_{lt,f+1}, \qquad \forall f \in [1, |F|-1], \forall l \in L, \forall t \in T \tag{21}$$

4.6.6 Domain of variables

The integrity requirement of decision variables x_{lt} and y_{ltf} are provided by:

$$x_{lt} = 0 \text{ or } 1, \qquad \forall l \in L, \forall t \in T \tag{22}$$

$$y_{ltf} = 0 \text{ or } 1, \qquad \forall l \in L, \forall t \in T, \forall f \in F \tag{23}$$

Now, we can completely formulate the train service design problem (TSD) in urban rail transit lines as a mixed integer nonlinear programming model M1 to minimize the objective function (6) with Constraints (7) to (23).

5 Linearization and Simplification of Model

5.1 Model Linearization

In model M1, there is only one nonlinear item $w_n^{ij} \sum_{t\in T} \sum_{f\in F} y_{\tau_a tf}$ in Constraints (8). Indeed, it is a continuous variable w_n^{ij} multiplied by a 0-1 variable $y_{\tau_a tf}$ which can be easily linearized by many existing techniques. Here, we present a novel linearization technique for this nonlinear item by utilizing the characteristics of the model. Recall that in our model the frequency of each train service is discretized from 1 to $|F|$, and variable y_{ltf} denotes whether the fth train with size t is operated on service l. That is, the frequency of the fth train with size t on service l is just 1. Due to that, each boarding arc a originating from station node n in the service network could be duplicated by $|T| \times |F|$ times. The frequency of the duplicated arc corresponding to the fth train with size t on boarding arc a is only 1. A new non-negative contiguous variable κ_{atf}^{ij} is defined to represent the number of passengers of OD pair (i,j) on the arc with respect to the fth train with size t on boarding arc a. Thus, Constraints (8) can be linearized by the following constraints.

$$\kappa_{atf}^{ij} \le \frac{2}{h} w_n^{ij}, \qquad \forall t \in T, \forall f \in F, \forall a \in A_{n+} \bigcap A_{1\theta_{ij}}, \forall n \in N_1, \forall i \in V, j \in V \tag{24}$$

$$\kappa_{atf}^{ij} \le M_1 y_{\tau_a tf}, \qquad \forall t \in T, \forall f \in F, \forall a \in A_{n+} \bigcap A_{1\theta_{ij}}, \forall n \in N_1, \forall i \in V, j \in V \tag{25}$$

$$k_a^{ij} = \sum_{t\in T} \sum_{f\in F} \kappa_{atf}^{ij}, \qquad \forall a \in A_{n+} \bigcap A_{1\theta_{ij}}, \forall n \in N_1, \forall i \in V, j \in V \tag{26}$$

$$\kappa_{atf}^{ij} \ge 0, \qquad \forall t \in T, \forall f \in F, \forall a \in A_{n+} \bigcap A_{1\theta_{ij}}, \forall n \in N_1, \forall i \in V, j \in V \tag{27}$$

Constraints (24) are the disaggregation representation of the optimal strategy requirement. These constraints take the same effects as Constraints (8). However, they are purely linear as the corresponding frequency is 1. Constraints (25) specify the relationship between variables κ_{atf}^{ij} and $y_{\tau_a tf}$. There, M_1 is a large positive constant. For each OD pair (i,j), it can take the volume q_{ij} of the pair. Constraints (26) compute the number of passengers of OD pair (i,j) on boarding arc a by summing that on all the associated duplicated arcs. Constraints (27) are the non-negative requirements of variables κ_{atf}^{ij}.

Through the above model linearization, the nonlinear train service design model M1 can be transformed into a mixed integer linear programming model M2 to minimize the objective function (6) and satisfy Constraints (7) and (9) to (27).

5.2 Model Simplification

5.2.1 Service network simplification

In our computational experiments, the model size and computation time are strongly influenced by the size of the service network. Observe that train services can be operated only between turn-back stations. Meanwhile, in each direction, the average waiting time of passengers at an intermediate station between two adjacent turn-back stations denoted as tv_1 and tv_2 (tv_2 is in front of tv_1 in the corresponding direction) is the same as that at turn-back station tv_1. Thus, we develop a method to reduce the size of the service network without losing of the solution accuracy of the model. The service network simplification

method works as follows:

Step1: In the service network, the station nodes and service nodes with respect to the stations without turn-back facilities are removed. Only the station nodes and service nodes related to turn-back stations are remained. Following the rules, the service network in Figure 3 is simplified as a smaller network shown in Figure 4.

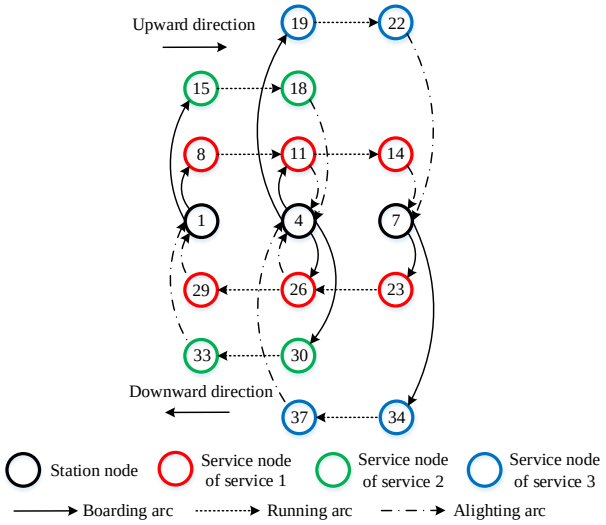


Figure 4: Simplified service network

Step2: The origin, destination and volume of OD pairs in the original service network are adjusted and aggregated in the simplified service network as follows:

(i) If the origin and destination of an OD pair are both turn-back stations, the origin and destination station node of this pair in the simplified network remain unchanged.

(ii) If the origin of an OD pair is not a turn-back station, the origin station node of this pair is set as the nearest turn-back station node behind the travel direction of this pair in the simplified network.

(iii) If the destination of an OD pair is not a turn-back station, the destination station node of this pair is set as the nearest turn-back station node in front of the travel direction of this pair in the simplified network.

(iv) After the adjustment of origin and destination, the volume of aggregated OD pairs in the simplified network is determined according to the volume of OD pairs in the original network. For instance, in the upward direction of Figure 4, the relationships between the volume of the aggregated OD pairs $\{q'_{ij} | i \in N'_1, j \in N'_1\}$ and that of the original OD pairs $\{q_{ij} | \forall i \in V, j \in V\}$ are as follows:

$$q'_{14} = q_{12} + q_{13} + q_{14} + q_{23} + q_{24} + q_{34}$$
$$q'_{47} = q_{45} + q_{46} + q_{47} + q_{56} + q_{57} + q_{67}$$
$$q'_{17} = q_{15} + q_{16} + q_{17} + q_{25} + q_{26} + q_{27} + q_{35} + q_{36} + q_{37}$$

To distinguish the simplified service network from the original one, additional notation shown in Table 2 is introduced.

Table 2: Notation of the simplified service network

Notation	Description
G'	Simplified service network, $G' = \{N', A'\}$.
N'	Set of nodes in the simplified service network G', $N' = \{N'_1, N'_2\}$.
N'_1	Set of station nodes in G'.
N'_2	Set of train service nodes in G'.
A'	Set of arcs in G', $A' = \{A'_1, A'_2\}$.
A'_1	Set of boarding arcs in G', $A'_1 = \{A'_{1d} \mid d \in D\}$.
A'_{1d}	Set of boarding arcs in direction d of G'.
A'_2	Set of running and alighting arcs in G'.
A'_{n+}	Set of outgoing arcs from node n in G'.
A'_{n-}	Set of incoming arcs to node n in G'.

After the service network simplification, the linear train service design model M2 can be reduced as a smaller size linear model M3 as follows:

$$\min \sum_{l\in L}\sum_{t\in T}\sum_{f\in F}\frac{c_t^1 m_l y_{ltf}}{h} + \sum_{l\in L}\sum_{t\in T}\sum_{f\in F} c_t^2 g_l y_{ltf} + \frac{\sigma}{60}\sum_{i\in N'_1}\sum_{j\in N'_1}\sum_{n\in N'_1} w_n^{ij} \tag{28}$$

s.t. Constraints (11)-(22)

$$\sum_{a\in A'_{n+}} k_a^{ij} - \sum_{a\in A'_{n-}} k_a^{ij} = \begin{cases} q'_{ij}, & \lambda_n = i, \quad n \in N'_1, \forall i \in N'_1, j \in N'_1 \\ -q'_{ij}, & \lambda_n = j, \quad n \in N'_1, \forall i \in N'_1, j \in N'_1 \\ 0, & n \in N'_2, \forall i \in N'_1, j \in N'_1 \end{cases} \tag{29}$$

$$\kappa_{atf}^{ij} \le \frac{2}{h} w_n^{ij}, \quad \forall t \in T, \forall f \in F, \forall a \in A'_{n+} \bigcap A'_{1\theta_{ij}}, \forall n \in N'_1, \forall i \in N'_1, j \in N'_1 \tag{30}$$

$$\kappa_{atf}^{ij} \le M_2 y_{\tau_a tf}, \quad \forall t \in T, \forall f \in F, \forall a \in A'_{n+} \bigcap A'_{1\theta_{ij}}, \forall n \in N'_1, \forall i \in N'_1, j \in N'_1 \tag{31}$$

$$k_a^{ij} = \sum_{t\in T}\sum_{f\in F} \kappa_{atf}^{ij}, \quad \forall a \in A'_{n+} \bigcap A'_{1\theta_{ij}}, \forall n \in N'_1, \forall i \in N'_1, j \in N'_1 \tag{32}$$

$$k_a^{ij} = 0, \forall a \in A'_{n+} \bigcap A'_{1\theta_{ji}}, \quad \forall n \in N'_1, \forall i \in N'_1, j \in N'_1 \tag{33}$$

$$\kappa_{atf}^{ij} \ge 0, \quad \forall t \in T, \forall f \in F, \forall a \in A'_{n+} \bigcap A'_{1\theta_{ij}}, \forall n \in N'_1, \forall i \in N'_1, j \in N'_1 \tag{34}$$

$$k_a^{ij} \ge 0, \quad \forall a \in A', \forall i \in N'_1, \forall j \in N'_1 \tag{35}$$

$$w_n^{ij} \ge 0, \quad \forall n \in N'_1, \forall i \in N'_1, \forall j \in N'_1 \tag{36}$$

In Constraints (31), the big-M parameter M_2 can be valued as q'_{ij} for each OD pair.

5.2.2 OD pair aggregation

In model M3, variables w_n^{ij}, k_a^{ij}, κ_{atf}^{ij} and constraints with respect to passenger assignment are generated for each OD pair in the service network. The number of OD pairs and passenger assignment constraints increases at a square speed with the number of turn-back stations in the service network, leading to a rapid increase in the scale of the model. Consider that only the waiting time of passengers is influenced by train services. Meanwhile, under the optimal strategy-based passenger assignment, the trip of single OD pair does not impact the computation of the total waiting time of passengers. To further simplify the model, we refer to the OD pair aggregation method of Spiess (1989) to process the OD pairs in the service network. There, variables w_n^{ij}, k_a^{ij}, κ_{atf}^{ij} and passenger assignment constraints only need to be generated for each group of OD pairs to each destination station node in the service network. The OD pair aggregation method is implemented as follows:

Step1: Aggregate all OD pairs in the service network into groups of OD pairs such that

the OD pairs in each group go to the same destination station node. For destination station node i and station node n in the service network, let μ_n^i either be the total number of passengers from all other station nodes to i if $n = i$, or the number of passengers from station node n to i if $n \neq i$. Thus, the relationship between μ_n^i and q'_{ij} is as follows:

$$\mu_n^i = \begin{cases} -\sum_{j \in N'_1, j \neq i} q'_{ji}, & n = i, \forall i \in N'_1, \forall n \in N'_1 \\ q'_{ni}, & n \neq i, \forall i \in N'_1, \forall n \in N'_1 \end{cases}$$

Step2: Replace variables w_n^{ij}, k_a^{ij} and κ_{atf}^{ij} with w_n^i, k_a^i and κ_{atf}^i, respectively. Here, w_n^i indicates the waiting time of all the passengers to destination station node i at station node n. k_a^i represents the number of all the passengers to destination station node i on boarding arc a. κ_{atf}^i denotes the number of all the passengers to destination station node i on the arc with respect to the fth train with size t on boarding arc a.

Through the OD pair aggregation, model M3 can be further simplified as follows:

$$\min \sum_{l \in L} \sum_{t \in T} \sum_{f \in F} \frac{c_t^1 m_l y_{ltf}}{h} + \sum_{l \in L} \sum_{t \in T} \sum_{f \in F} c_t^2 g_l y_{ltf} + \frac{\sigma}{60} \sum_{i \in N'_1} \sum_{n \in N'_1} w_n^i \tag{37}$$

s.t. Constraints (11)-(22)

$$\sum_{a \in A'_{n+}} k_a^i - \sum_{a \in A'_{n-}} k_a^i = \begin{cases} \mu_n^i, & \forall n \in N'_1, \forall i \in N'_1 \\ 0, & \forall n \in N'_2, \forall i \in N'_1 \end{cases} \tag{38}$$

$$\kappa_{atf}^i \leq \frac{2}{h} w_n^i, \quad \forall t \in T, \forall f \in F, \forall a \in A'_{n+} \bigcap A'_{1\theta_{ni}}, \forall n \in N'_1, \forall i \in N'_1 \tag{39}$$

$$\kappa_{atf}^i \leq M_3 y_{\tau_a tf}, \quad \forall t \in T, \forall f \in F, \forall a \in A'_{n+} \bigcap A'_{1\theta_{ni}}, \forall n \in N'_1, \forall i \in N'_1 \tag{40}$$

$$k_a^i = \sum_{t \in T} \sum_{f \in F} \kappa_{atf}^i, \quad \forall a \in A'_{n+} \bigcap A'_{1\theta_{ni}}, \forall n \in N'_1, \forall i \in N'_1 \tag{41}$$

$$k_a^i = 0, \quad \forall a \in A'_{n+} \bigcap A'_{1\theta_{in}}, \forall n \in N'_1, \forall i \in N'_1 \tag{42}$$

$$\kappa_{atf}^i \geq 0, \quad \forall t \in T, \forall f \in F, \forall a \in A'_{n+} \bigcap A'_{1\theta_{ni}}, \forall n \in N'_1, \forall i \in N'_1 \tag{43}$$

$$k_a^i \geq 0, \quad \forall a \in A', \forall i \in N'_1 \tag{44}$$

$$w_n^i \geq 0, \quad \forall n \in N'_1, \forall i \in N'_1 \tag{45}$$

The big-M parameter M_3 of Constraints (40) can take the value of $\sum_{j \in N'_1, j \neq i,} q'_{ji}$ for each destination station node i.

It is worth noting that the two model simplification techniques including the service network simplification and OD pair aggregation are independent. They can be used either separately or unitedly to simplify model M2. The order of the two simplification processes are not fixed. For comparison, we call model M2 simplified only by the OD pair aggregation as model M4, and model M2 simplified by both techniques as model M5.

6 Computational Experiments

In this section we describe our computational experiments on the (planned) Chongqing Rapid Rail Transit Line 26 in China. The proposed approaches are coded by MATLAB R2016a and CPLEX 12.8 is invoked to solve the optimization models. We run all experiments on a PC with Inter Core i7-7700 3.6 GHz CPU and 16 GB RAM.

6.1 Test Line and Parameter Setting

The total length of the test line with 20 stations and 19 sections is 121.7 km as shown in Figure 5. Stations are numbered from v_1 to v_{20} along the upward direction. There are four turn-back stations namely station v_1, v_6, v_{17} and v_{20} on the line. Station v_1 can only switch the running direction of trains from downward to upward. Station v_{20} can only switch the running direction of trains from upward to downward. Different from station v_1 and v_{20}, station v_6 and v_{17} can reverse the running direction for trains from both directions. The capacity of turn-back stations in each direction are listed in Table 3. It should be mentioned that in addition to the turn-back track, the depot entrance and exit track can also reverse the running direction for trains during the operation period.

Six candidate train services on the line can be generated according to the layout of turn-back stations. The turn-back stations and round-trip time of each candidate train service are provided in Table 4. In the Table, the number in each cell represents the round-trip time of the train service formed by the turn-back stations in the row and column of the cell.

There are three types of train sizes that can be operated on the line, including 4-car trains, 6-car trains and 8-car trains. The relevant parameters of train sizes are described in Table 5. Other parameters of the test line are in Table 6.

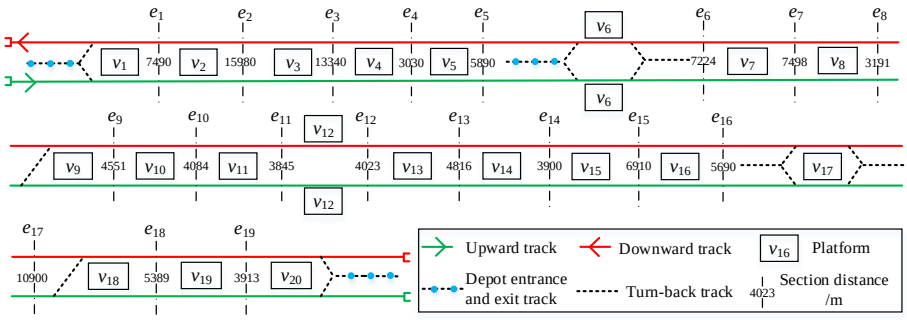


Figure 5: Layout of the test line

Table 3: Capacity of turn-back stations

Station	Capacity to upward/ trains	Capacity to downward/ trains	Station	Capacity to upward/ trains	Capacity to downward/ trains
v_1	20	0	v_{17}	20	20
v_6	20	20	v_{20}	0	20

Table 4: Round-trip time of train services (min)

Turn-back stations	v_6	v_{17}	v_{20}
v_1	60.8	146.8	172.8
v_6	-	93.1	119.2
v_{17}	-	-	33.4

Table 5: Parameters of train sizes

Train size	c_t^1 (¥/train)	c_t^2 (¥/train·km)	r_t (passengers)
4-car train	230	100	896
6-car train	340	150	1376
8-car train	450	200	1856

Table 6: Other input parameters

Parameter	Value	Unit	Parameter	Value	Unit
h	60	min	b_e^{min}	6	trains
σ	28	¥/h	φ	6	trains
b_e^{max}	20	trains	δ	10	%

6.2 Instances Generation

In order to analyse the performance of the proposed model and the effectiveness of the two model simplification methods, 15 realistic instances based on the test line are constructed to test the performance of model M2, M3, M4 and M5.

We use three different scales of passenger demand at the morning rush hour from 8:00 to 9:00 in the initial, immediate and long-term planning horizon of the test line, as displayed in Figure 6. Among them, the total number of passengers of Figure 6(a), 6(b) and 6(c) is 87973, 98240 and 111136, respectively. For convenience, let $OD = 1,2,3$ represent the three scenarios of passenger demand in Figure 6. Besides, for each scenario of passenger demand, the maximum allowable number of train services Ω is set from 1 to 5 (i.e. $\Omega = \{1, \dots, 5\}$). Thus, in total 15 instances are obtained to test the proposed approaches.

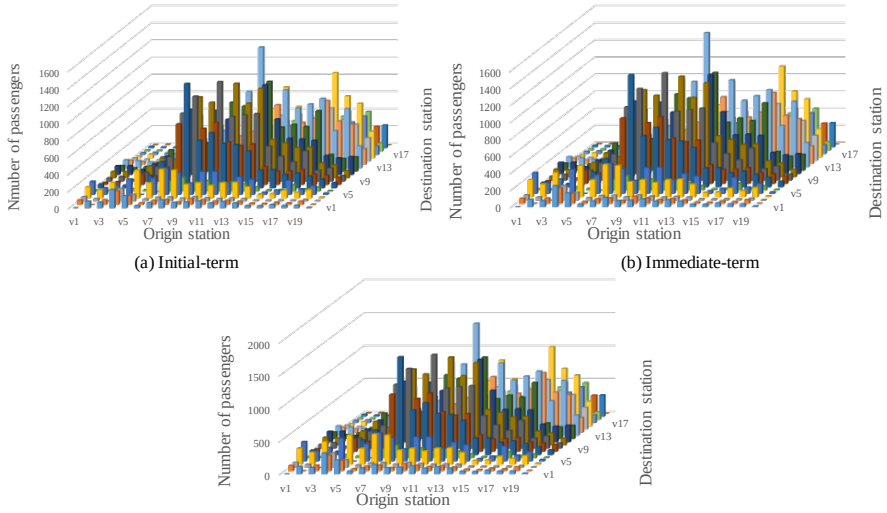


Figure 6: Three scenarios of passenger demand

6.3 Results

6.3.1 Effectiveness of model simplification methods

We first compare the scale of the four configured models listed in Table 7. There, column 2 and 3 are the number of nodes and arcs in the associated service network, respectively. The last three columns are the number of 0-1 variables, continuous variables and constraints in the model, respectively. As shown, the scale of service network and scale of model in M2 are the largest. It has millions of variables and constraints. The size of service network in model M3 which is simplified by the layout of turn-back stations is obviously reduced. The number of continuous variables and constraints decrease significantly too. Model M4

has the same service network size as model M2. However, the OD pair aggregation method provides a smaller size to model M4 compared with model M2. The size of model M4 is between that of model M2 and M3. Model M5 simplified by the two methods has the same size of service network as model M3. But there are only near 5000 continuous variables and 8000 constraints in model M5. Thus, in terms of model scale, we have M2 > M4 > M3 > M5. In addition, all models have the same number of 0-1 variables. Because both of the two simplification methods only reduce the number of continuous variables w_n^{ij}, k_a^{ij} and κ_{atf}^{ij} and the constraints which contain these continuous variables. The 0-1 variables x_{lt} and y_{ltf} which determine the train service scheme are not simplified.

Table 7: Scale of models

Model	# Nodes	# Arcs	# 0-1 variables	# Continuous variables	# Constraints
M2	168	408	378	3435200	4941789
M3	36	60	378	20224	23837
M4	168	408	378	171760	259813
M5	36	60	378	5056	8057

We then analyse the computational effectiveness of the configured models. The maximum running time of each model is limited to 4 h. The computational results are summarized in Table 8. In the Table, Columns 3-6 are the objective function value of the models. Columns 7-10 are the optimality gap between the best lower bound and upper bound. The last four columns are the computation time. As observed, model M2 can only find the optimal solution for three instances ($OD = \{1,2,3\}, \Omega = \{1\}$). For other instances, only feasible solutions with large gaps (39% to 56%) are obtained in 4 h. Contrarily, Model M3, M4 and M5 can obtain the optimal solution for all instances within the limited time. The computation time of the four models is consistent with the scale of the models. The solving speed of model M3 and M5 is quite fast with an average computation time less than 10 s. Due to the model scale, Model M4 is relatively difficult to solve and its average computation time is near 20 min.

Table 8: Computational results of models

Instance		Objective/ ¥				Gap/ %				Time/ s			
OD	Ω	M2	M3	M4	M5	M2	M3	M4	M5	M2	M3	M4	M5
1	1	644286	644286	644286	644286	0	0	0	0	26	0	2	0
	2	881654	491088	491088	491088	56	0	0	0	14400	4	1001	3
	3	881654	476368	476368	476368	56	0	0	0	14400	7	1661	6
	4	881654	476368	476368	476368	56	0	0	0	14400	7	1583	6
	5	881654	476368	476368	476368	56	0	0	0	14400	14	1494	6
2	1	718038	718038	718038	718038	0	0	0	0	24	0	2	0
	2	890171	536576	536576	536576	52	0	0	0	14400	9	1233	2
	3	890171	526337	526337	526337	52	0	0	0	14400	8	1354	5
	4	890171	526337	526337	526337	52	0	0	0	14400	13	1677	5
	5	890171	526337	526337	526337	52	0	0	0	14400	14	2064	7
3	1	793981	793981	793981	793981	0	0	0	0	24	0	2	0
	2	901251	610022	610022	610022	46	0	0	0	14400	4	727	2
	3	901251	600232	600232	600232	46	0	0	0	14400	7	1594	4
	4	803371	595174	595174	595174	39	0	0	0	14400	16	1512	5
	5	901251	595174	595174	595174	46	0	0	0	14400	18	1483	5
Ave		850049	572846	572846	572846	41	0	0	0	11525	8	1159	4

Based on the above comparison, we can conclude that the two model simplification methods including the service network simplification and OD pair aggregation can both

effectively reduce the scale of the original model and improve the solution quality. When the original model is simplified by one of the two methods alone, the service network simplification has better effects in reducing the model scale and improving the solving speed. Note that the OD pair aggregation also has a notable simplification effect. When the two model simplification methods work together, the simplified model M5 has the smallest model scale and the shortest computation time, which enables us to solve practical-sized train service design problems of urban rail transit lines in extremely short time.

6.3.2 Comparison with single train service scheme

To testify the quality of the train service scheme we proposed, we compare the multiple train services scheme obtained by model M5 with the single train service scheme frequently designed by experience in practice. For simplicity, we only design train services in the long-term planning horizon of the test line. The corresponding passenger demand is shown in Figure 6(c). The two schemes are obtained as follows:

(i) Single train service scheme (STSS). In practice, the train service scheme of an urban rail transit line is usually determined by using a full-length train service and a single size according to the maximum passenger load of sections in the line. Under this rule, the best single train service scheme of the test line can be obtained by setting $OD = 3$, $\Omega = 1$ and $T = \{8\}$ in model M5. The resulting train size and frequency of the single full-length service are the 8-car train and 14 pairs of trains, respectively.

(ii) Multiple train services scheme (MTSS). To obtain the optimal train service scheme of the test line, we can set $OD = 3$, $\Omega = 5$ and $T = \{4,6,8\}$ in model M5. The obtained multiple train services scheme is depicted in Figure 7. As seen, three are 4 train services where 4-car trains and 8-car trains are used.

We first analyse the transfer of passengers under different train service schemes. For scheme STSS, as a full-length train service is operated, passengers do not need to transfer when traveling on the line. Regarding to scheme MTSS, some passengers need to transfer at most twice when they travel. For example, the travel process of partial passengers from station v_{20} to station v_1 is v_{20}-Service 4-v_{17}-Service 3-v_6-Service 1 or Service 2-v_1, and hence the number of transfers is two. However, only the passengers from stations v_{18}, v_{19} and v_{20} to stations v_1, v_2, v_3, v_4 and v_5 need to transfer twice during their journey. The number of these passengers is only 457, accounting for 0.41% of the total number of passengers. We will indicate that the operator's cost and passengers' cost can be reduced significantly while a tiny proportion of passengers has an inconvenient journey.

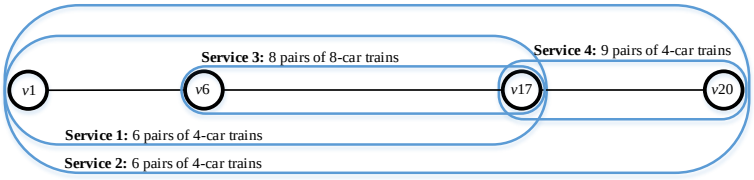


Figure 7: Proposed multiple train services scheme

Then we compare the frequency of trains in sections for different schemes are shown in Figure 8. As indicated, scheme STSS has a single full-length service covering all stations and sections of the line, i.e. 14 pairs of 8-car trains are operated on the whole line. In scheme MTSS, 12 pairs of 4-car trains are operated in sections e_1 to e_5, 12 pairs of 4-car trains and 8 pairs of 8-car trains serve sections e_6 to e_{16}, and 15 pairs of 4-car trains run in

sections e_{17} to e_{19}. The frequency of trains in scheme MTSS is higher than that of scheme STSS in sections e_6 to e_{19}, where exist most of the passenger flows in the line, thus leading to a shorter waiting time for most of the passengers.

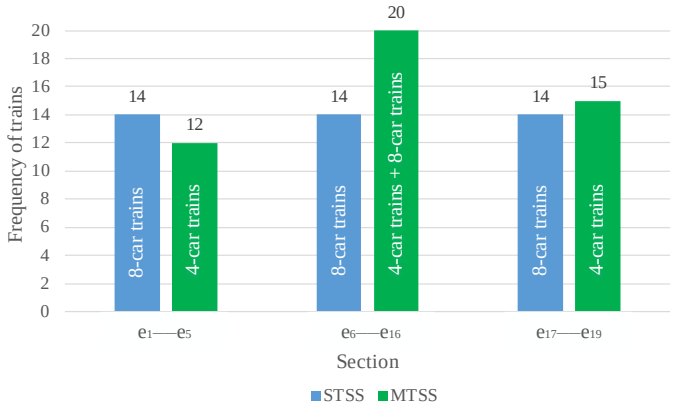


Figure 8: Frequency of trains in sections

We further analyse the match between capacity and demand in sections for both schemes. The capacity, demand and load factor in sections under the two schemes are displayed in Figure 9. As known, in scheme STSS, only 8-car trains are used and the frequency of trains is the same in each section of the line. Therefore, the capacity in each section is equal. However, the passenger load is obviously unbalanced in sections. The passenger load in the middle sections of the line is large while that in the two ends is small. This leads to the waste of capacity in the two ends of the line, which is not economical for the operator. On the contrary, for scheme MTSS, 4 services and 2 types of train sizes are used on the lines, such that different sections of the line can be more flexibly equipped with capacity. The capacity in sections e_1 to e_5 and e_{17} to e_{19} are smaller but match the passenger load better than scheme STSS, which can help to reduce the cost of operating the line. Note that lower capacity in sections does not necessarily mean lower frequency of trains in sections. Because in scheme MTSS, many small size trains are operated to increase the frequency of trains in sections so as to reduce the waiting time of passengers. As shown in Figure 8, the frequency of trains under MTSS in sections e_1 to e_5 is lower than that of STSS. But in sections e_{17} to e_{19}, the frequency of trains in MTSS is higher than that of STSS.

Finally, we compare the objective function value under different schemes as summarized in Table 9. It can be seen from Table 9 that compared with scheme STSS, scheme MTSS reduces the total cost by 26.58%. Meanwhile, the total fixed cost of operators, the total operation cost of operators and the total waiting time cost of passengers are all decreased. Besides, through a close look at the composition of the total cost in scheme MTSS, we can find that the total operation cost Z_2 accounts for most of the total cost with a rate of 81.06%, while the total fixed cost Z_1 with a rate of 2.37% is the smallest part of the total cost. This is different from our empirical understanding that the fixed cost may account for the majority of the total cost in urban rail transit systems. Nevertheless, this difference reflects that we should save the operation cost as much as possible so as to control the total cost of operating multiple train services in urban rail transit lines.

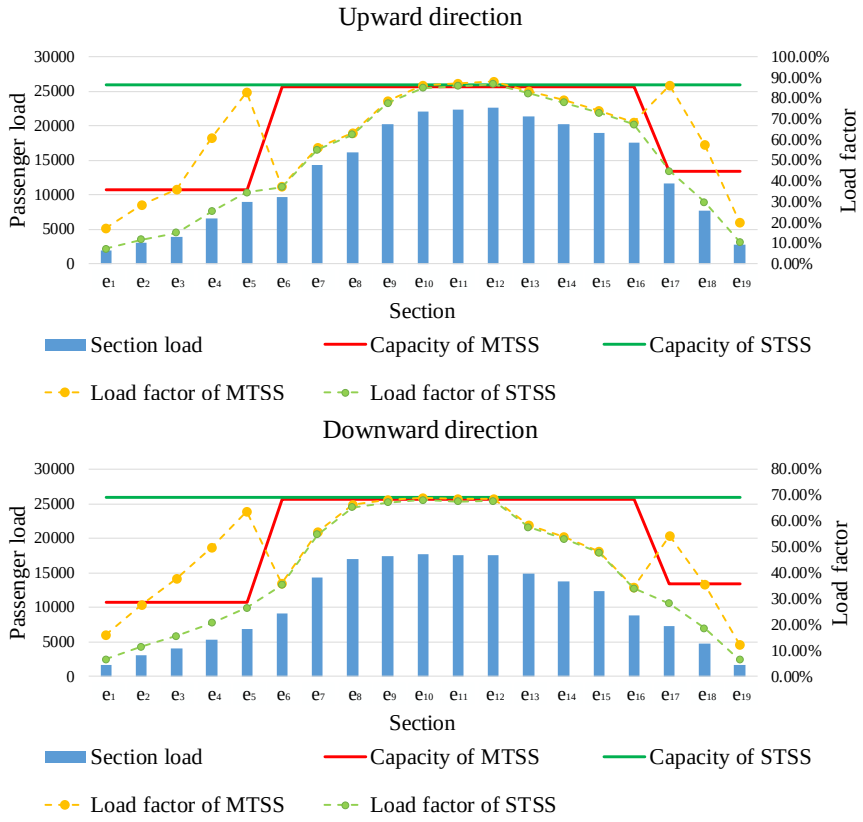


Figure 9: Capacity, demand and load factor in sections

Table 9: Objective function value of schemes

Objective	STSS	MTSS
Total cost/ ¥	810600.2	595173.9
Total fixed cost of operators Z_1/ ¥	18145.8	14087.3
Total operation cost of operators Z_2/ ¥	681318.4	482457.2
Total waiting time cost of passengers Z_3/ ¥	111136.0	98629.4

In summary, in light of the above comparisons, we conclude that our approach which can flexibly design a train service scheme with multiple services and multiple train sizes is better than the empirical method frequently used in practice to design a single train service scheme. With our approach, although some passengers may experience a limited number of transfers, the capacity in sections can match the passenger load better. Meanwhile, the frequency of trains in sections can be increased as much as possible, reducing significantly of the operators' cost and passengers' waiting time cost. Thus, the proposed approach can be used to design practically acceptable train services in urban rail transit systems.

7 Conclusions

This study proposes an optimization approach for the train service design in an urban rail transit line considering short-turn services and multiple train sizes. A service network is constructed based on a pool of candidate services generated in advance. By considering a series of operational and capacity constraints, the problem is formulated as a mixed integer nonlinear programming model to minimize the operators' cost and passengers' waiting time cost. After a model linearization, two simplification methods namely service network simplification and OD pair aggregation are used to simplify the linear model. Finally, the Chongqing Rapid Rail Transit Line 26 in China is used to test our approach.

Compared with the existing studies, the model we proposed is more flexible in the sense that it can be applied to the train service design of an urban rail transit line with multiple train services including either full-length or short-turn ones and multiple train sizes. Next, seldom former studies consider the travel process of passengers in case of multiple train services in urban rail transit systems. In this work, we propose a service network construction method based on a pool of candidate services generated in advance, and use the optimal strategy to assign passenger flows in the service network so as to well describe the travel process of passengers, such that the impact of multiple train services on the trip choices of passengers can be accurately analysed. Furthermore, a service network simplification and an OD pair aggregation are developed to simplify our model efficiently. Computational experiments show that the model with the two simplification techniques can be quickly solved to optimality by commercial solvers for typical urban rail transit lines. The obtained multiple train services scheme effectively reduces the operator's cost and passengers' waiting time cost compared to the single train service scheme frequently designed by experience in practice.

Acknowledgments

We are partially supported by the National Natural Science Foundation of China (No. 61603318, U1834209), and the National Key Research and Development Program of China (No. 2017YFB1200701).

References

Furth, P. G., 1987. "Short Turning on Transit Routes", *Transportation Research Record*, vol. 1108, pp. 42–52.

Ceder, A., 1989. "Optimal Design of Transit Short-Turn Trips", *Transportation Research Record*, vol. 1221, pp. 8–22.

Zhang, M., Wang, Y., Su, S., Tang, T., & Ning, B., 2018. "A Short Turning Strategy for Train Scheduling Optimization in an Urban Rail Transit Line: The Case of Beijing Subway Line 4", *Journal of Advanced Transportation*, 2018.

Ceder, A. A., Hassold, S., & Dano, B., 2013. "Approaching even-load and even-headway transit timetables using different bus sizes", *Public Transport*, vol. 5, no. 3, pp. 193-217.

Hassold, S., & Ceder, A., 2012. "Multiobjective approach to creating bus timetables with multiple vehicle types". *Transportation Research Record*, vol. 2276, pp. 56-62.

Chen, Z., Li, X., & Zhou, X., 2019. "Operational design for shuttle systems with modular vehicles under oversaturated traffic: Discrete modeling method". *Transportation Research Part B: Methodological*, vol. 122, pp. 1-19.

Delle Site, P., & Filippi, F., 1998. “Service Optimization for Bus Corridors with Short-Turn Strategies and Variable Vehicle Size”, *Transportation Research Part A: Policy and Practice*, vol. 32, no. 1, pp. 19-38.

Tirachini, A., Cortés, C. E., & Jara-Díaz, S. R., 2011. “Optimal design and benefits of a short turning strategy for a bus corridor”, *Transportation*, vol. 38, no. 1, pp. 169-189.

Cancela, H., Mauttone, A., & Urquhart, M. E., 2015. “Mathematical programming formulations for transit network design”, *Transportation Research Part B*, vol. 77, pp. 17-37.

Ji, Y., Yang, X., & Du, Y., 2016. “Optimal design of a short-turning strategy considering seat availability”, *Journal of Advanced Transportation*, vol. 50, no. 7, pp. 1554-1571.

Sun, Y., Schonfeld, P. M., Lu, Y., & Zhou, M., 2016. “Redesigning rail transit short-turn operations: case study of line 2 of the shanghai metro in China”, *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2540, pp. 46-55.

Yang, X., Ji, Y., Du, Y., & Zhang, H. M., 2017. “Bi-Level Model for Design of Transit Short-Turning Service Considering Bus Crowding”, *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2649, pp. 52-60.

Ding, X., Guan, S., Sun, D. J., & Jia, L., 2018. “Short turning pattern for relieving metro congestion during peak hours: the substance coherence of Shanghai, China”, *European Transport Research Review*, vol. 10, no. 2, pp. 28.

Spiess, H., & Florian, M., 1989. “Optimal strategies: a new assignment model for transit networks”, *Transportation Research Part B*, vol. 23, no. 2, pp. 83-102.

Modelling the Influences of Primary Delays Based on High-speed Train Operation Records

ZHONGCAN LI [a], PING HUANG [a,b], CHAO WEN [a,b,1], YIXIONG TANG [a]

[a] School of Transportation and Logistics, Southwest Jiaotong University
Nr.111, North 1st Section of Second Ring Road, 610031, Chengdu, China
[b] High-speed Railway Research Center, University of Waterloo
Waterloo, N2L3G1, Canada
[1] E-mail: c9wen@uwaterloo.ca, Phone: 1-2269788096

Abstract
Primary delays (PDs) are the driving force of delay propagation. Hence, accurate predictions of the number of affected trains (NATs) and the total time of affected trains (TTATs) due to PDs can provide a theoretical background for the dispatch of trains in real time. Train operation data were obtained from Wuhan-Guangzhou High-Speed Railway (HSR) station from 2015 to 2016, and the NAT and TTAT influence factors were determined after analyzing the PD propagation mechanism. The NAT predictive model was established using eXtreme Gradient Boosting (XGBOOST) algorithm which was more efficient than other machine learning methods after comparison. Furthermore, the TTAT predictive model was established based on the NAT model using the support vector regression (SVR) algorithm. The results indicate that the XGBOOST algorithm has good performance on the NAT predictive model, whereas SVR is the best method for the TTAT model using Lessthan5 variable, which is the ratio of the difference between the sample size of actual and the predicted values in less than 5 min and the total sample size. In addition, 2018 data were used to evaluate the application of NAT and TTAT models over time. The results indicate that NAT and TTAT models have a good application over time.

Keywords:
High-speed railway, Primary delay, Number of affected trains, Total time of affected trains, Machine learning

1. INTRODUCTION

High-speed railway (HSR) transportation is becoming more popular than other modes of transportation worldwide owing to their high speed, safety, and density. In China, HSR trains have become one of the major means of transportation. High punctuality of these trains is an important factor considered by railway companies in attracting passengers (Yuan et al., 2002). However, they are influenced by bad weather, mechanical failure of the systems, and organization strategies during operation, which could lead to delays. These delays disrupt railway operation and transportation, increase travel time of passengers, and reduce the passenger travel experience, thereby making HSR trains less reliable.

Delays are categorized as primary delays (PDs) and secondary (knock-on) delays. PDs are the driving force of delay propagation. They occur when some uncertain events directly disrupt the train operations. However, secondary delays are attributed to the delay propagation caused by PDs. When a PD occurs, the operation adjustment mainly depends on the experience of train dispatchers. However, there are no scientific theories and methods that support the strategies used. Meanwhile, the number of affected trains (NATs) and the total time of affected trains (TTATs) due to a PD can be used to estimate the influence of PD and accurately determine the severity of the delay. Therefore, NAT and TTAT predictive models can assist the train dispatcher in estimating the train operation state, provide the theoretical basis for the rescheduling strategy, facilitate more scientific and reliable rescheduling decisions and adjustment based on the station work plan (Wen et al., 2018). Furthermore, NAT and TTAT predictive models are vital in the automatic operation of trains and the intelligent dispatch of HSRs.

The impact of the NAT and TTAT predictive models on PD propagation is determined in this study. The models were built based on the data obtained from Wuhan-Guangzhou HSR station (Guangzhou Railway Bureau, China) from March 2015 to November 2016, and evaluated using common machine learning classification and regression algorithms. The results indicated that eXtreme Gradient Boosting (XGBOOST) and support vector regression (SVR) algorithms had the best predictive results for NAT and TTAT models, respectively. Furthermore, the models were evaluated using 2018 data in order to test their effectiveness over time. The results show that the models have good predictive abilities and can be used for a long time.

This paper is structured as follows: Section 1 introduces the background and significance of the research. Section 2 reviews some studies conducted on delay propagation while Section 3 presents the problem to be solved and also describes the data used. The NAT and TTAT predictive models are established and tested in Section 4 while the conclusions are discussed in Section 5.

2. LITERATURE REVIEW

PDs may be caused by exogenous events such as irregularities in the natural environment or vehicle faults, accidents, facility failures, etc., in internal systems (Goverde, 2005). The severity of the delay is measured using a delay probability distribution model when the delay distribution corresponds to an exponential distribution and secondary delays are induced in different traffic scenarios (Huisman and Boucherie, 2001). (Meester and Muns, 2007) obtained the knock-on delay distribution from PD distributions using a phase-type distribution. However, (Goverde et al., 2013) found that Weibull distributions can be fitted on the PD distribution using empirical data. Meanwhile, (Wen et al., 2017) indicated that PD distributions could be well approximated by log-normal distributions while line regression models can be used to approximate NAT distributions. However, studies on predictive models of delay propagations are mostly based on mathematical optimization methods. (Huisman et al., 2002) and (Milinković et al., 2013) estimated train delays using Queuing and Petri net models, respectively. Meanwhile, (Hansen et al., 2010) proposed an online model for the prediction of running time and arrival time using timed event graphs. In addition, (Kecman and Goverde, 2015) proposed a timed event graph approach for the accurate prediction of train event times using dynamic arc weights model. Furthermore, (Goverde, 2007) established a delay propagation model using the max-plus algebra theory.

Data-driven studies are increasingly used in delay/disruption management. (Goverde,

2005) studied the systematic delay propagation in trains and employed a robust linear regression model to investigate the correlation among arrival delays using data obtained from Eindhoven Railway Station, Netherlands. Meanwhile, (Kecman et al., 2015) discussed the dynamics of train delays over time and space, and modeled the uncertainty of train delays based on a Markov stochastic process. (Şahin, 2017) also described the train operation process as a Markov chain and concluded that the train states at certain event timesteps could be determined by transition probability matrices. Furthermore, (Corman and Kecman, 2018) proposed an online Bayesian network to predict train delay over time using historical data in Sweden, while (Lessan et al., 2018) established a hybrid Bayesian network to estimate train arrival and departure delays based on real data in China. Artificial neural networks (ANNs) have been widely used to predict the delays in passenger trains (Chapuis, 2017; Pongnumkul et al., 2014; Yaghini et al., 2013). However, (Marković et al., 2015) indicated that SVR is more accurate for predicting train arrival delays in comparison with ANN algorithms based on Serbian Railways data. Meanwhile, (Tang et al., 2018) discovered the relationship between the causes of PD and the duration based on NAT and TTAT models using SVR. However, the NAT is unknown when a PD occurs that will lead to the model cannot predict online.

3. PROBLEM STATEMENT AND DATA DESCRIPTION

3.1 Problem statement

The headway between two trains in a station comprises the minimum interval time and the timetable supplement time. If a train is delayed before it arrives the station while the preceding train is not delayed, the delayed train is considered as a PD train. In other words, a delayed train is regarded as a PD train if a minimum threshold (e.g., 5 min in Wuhan-Guangzhou HSR station) exists between the arrival time (or scheduled arrival time) of the delayed train and the actual arrival time of the preceding train. The PD train greatly influences the motion of the subsequent train, thereby leading to the secondary delay. This process occurs for all successive trains. However, the PD train has less influence on the subsequent trains when the delay duration is less than 5 min such that the rescheduling of the trains is not necessary. Hence, only PD durations of more than 5 min are considered in this paper. Meanwhile, the delays are reduced by timetable supplement time until they are eliminated. Hence, there is a sequence in the PD influence where the number of PD and knock-on trains is classified by NAT and TTAT, which is the sum of the PD and knock-on delay time.

Figure 1 shows the process of PD propagation at two stations (Station A and Station B) in Wuhan-Guangzhou HSR station. The red and black lines are actual train lines and scheduled train lines, respectively. A minimum time interval exists between Train 1 and the preceding train, such that Train 1 is a PD train having a delay duration of t_1. Meanwhile, Train 2 is delayed as the interval between the actual arrival time of Train 1 and the scheduled arrival time of Train 2 is less than 5 min, thereby leading to a delay in Trains 3 and 4. The PD stops at Train 4 due to the supplement time t_{sup}^{i} such that Train 5 returns to normal operation. The delayed trains (Trains 1–4) form a PD sequence where NAT is 4 and TTAT is $\sum_{i=1}^{4} t_i$.

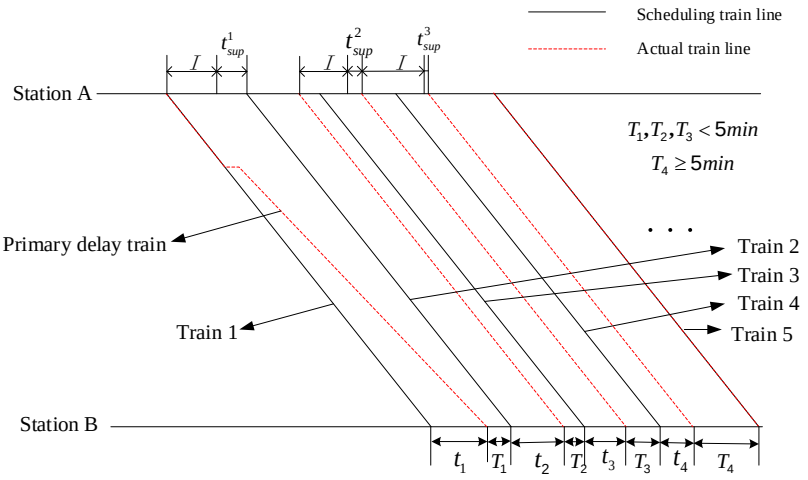


Figure 1: PD propagation process at two stations

The trains overtake one another if the actual arrival sequence is different from the scheduled arrival sequence. This sequence is due to rescheduling. However, the propagation process is complicated as many influence factors need to be considered. Hence, these sequences are not considered in this paper.

3.2 Data description

The data used in this study were obtained from the station operation records of Wuhan-Guangzhou HSR station (Guangzhou Railway Bureau, China) for Guangzhou North (GZN), Qingyuan (QY), Yingde West (YDW), Shaoguan (SG), Lechang East (LCE), Chenzhou West (CZW), Leiyang West (LYW), Hengyang East (HYE), Heangshan West (HSW), Zhuzhou West (ZZW), and Changsha South (CSS). Table 1 summarizes a portion of the data.

Table 1: Raw data from Guangzhou Station

Train NO	Date	Station	Scheduled arrive time	Scheduled departure time	Actual arrive time	Actual departure time
G280	2015/3/24	GuangzhouNorth	7:00:00	7:00:00	7:01:00	7:01:00
G636	2015/3/24	GuangzhouNorth	7:07:00	7:07:00	7:07:00	7:07:00
G1102	2015/3/24	GuangzhouNorth	7:13:00	7:13:00	7:14:00	7:14:00
G6102	2015/3/24	GuangzhouNorth	7:20:00	7:20:00	7:20:00	7:20:00

The primary influence predictive model was established by preprocessing the data in a series of steps summarized as follows:

- Step 1: Gather the data from the database and eliminate abnormal entries such as duplicate entries, errors and invalid entries.
- Step 2: Sort the data by actual arrival time in the station.
- Step 3: Select the PD train and obtain the train sequences which do not overtake based on PD influence.

- Step 4: Extract the features of the influence factors and calculate NAT and TTAT based on the PD influence sequences.

Thus, the feature sets of the influence factors of NAT and TTAT were obtained by analyzing the mechanism of the PD propagation. These influence factors are described as follows:

D: Primary delay duration of PD,

I: Scheduled interval between the PD train and the subsequent adjacent train,

B: 0-1 variable, which is 0 when the PD train does not stop at the station and 1 otherwise,

T: Period of a PD occurrence, and classify the period by hour

N: The number of affected trains if supplement times are fully utilized.

Table 2 summarizes a sample data after pre-processing:

Table 2: A sample of modeling data

D	***I***	***B***	***T***	***N***	**NAT**	**TTAT**
5	6	0	8:00-9:00	2	2	9
6	7	0	16:00-17:00	3	3	12
5	8	1	8:00-9:00	3	2	7
6	6	0	9:00-10:00	3	5	28
6	7	0	17:00-18:00	2	2	11

In this study, *D* presents the primary delay train delay duration; *I* record the scheduled headway between the PD train and the first train subsequently; *B* is a 0-1 variable, and it equals to 0 when the PD train does not stop at the station. Otherwise it equals to 1; Classify the period by hour and marked *T* as the period of PD occurs. *N* indicates the number of affected trains when the supplement times were fully utilized. All the factors above are obtained when PD occurs based on a real-time timetable. Hence, real-time rescheduling is possible if NAT and TTAT predictive models are investigated using these factors.

The predictive models were established using the data obtained from March 2015 to November 2016. Seventy percent of the data was used as the training data while 30% was used as the validation data for the model in order to prevent overfitting. Finally, the models were evaluated by using data obtained in 2018 as the test data.

4. PREDICTIVE MODEL OF NAT AND TTAT

4.1 The predictive model of NAT

Figure 2 shows the heatmap and 3D histogram of the intensity distribution of PD influence over time, which can assist train dispatchers in carrying out risk warnings. The PD duration and the period of PD occurrence for GZN station were plotted on the horizontal and vertical coordinates, respectively.

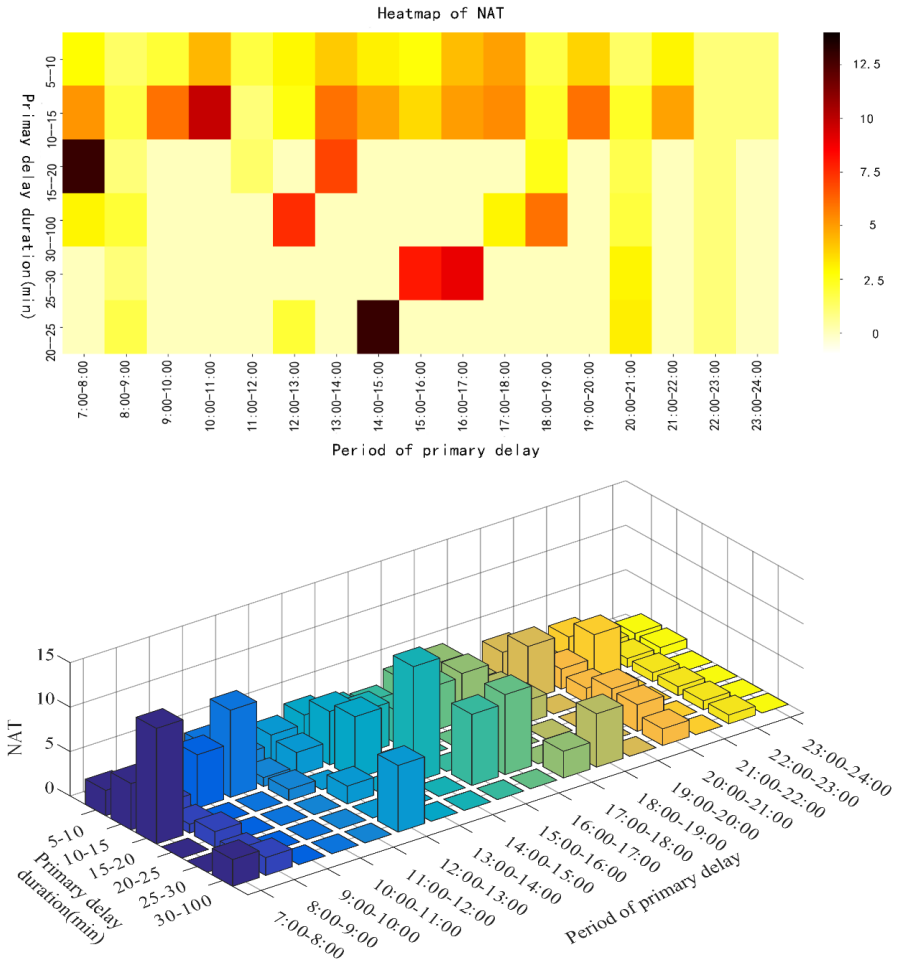


Figure 2: The NAT heatmap and 3D histogram of GZN station

The influence factors of NAT (G) are *D*, *B*, *I*, *T*, and *N*. NAT is a discrete random variable whose prediction is a classification problem. The output of the model is set to *S* while the feature set of the influence factors is the input such that the relationship between *S* and *G* is

$$S = \Phi(D, B, T, I, N) \tag{1}$$

where Φ is the classification algorithm. When NAT > 5, the sample size corresponding to each value is small, and the distribution is discrete. Thus, the NAT values that were greater than 5 were classified as 6 and more. Finally, NAT was divided into six categories (1 / 2 / 3 / 4 / 5 / 6 and more).

Meanwhile, XGBOOST was used as the classification algorithm. It is an improved algorithm based on gradient boosting decision tree which is highly efficient and flexible

and can be used for solving regression and classification problems. For a given dataset with *n* ensembles and *m* features, the result $\hat{y}_i$ is given by an ensemble represented by the model as follows:

$$D = \left\{ (x_i, y_i) : i = 1, 2, ...n, x_i \in R^m, y_i \in R \right\} \tag{2}$$

$$\hat{y}_i = \phi(x_i) = \sum_{k=1}^{K} f_k(x_i), f_k \in F \tag{3}$$

$$F = \left\{ f(X) = w_{q(x)} \right\} \left(q : R^m \longrightarrow T, \quad w \in R^T \right) \tag{4}$$

where f_k is a regression tree (also known as CART), $f_k(x_i)$ represents the score given by the *k*-th tree to the *i*-th sample in the data, *q* represents the structure of each tree that maps an example to the corresponding leaf index, and *T* is the number of leaves in the tree. Each f_k corresponds to an independent tree structure *q* and leaf weight *w*.

Minimizing the regularized function to give the objective function：

$$\ell(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \tag{5}$$

where *l* is the loss function and Ω is the penalty term to prevent overfitting and complexity of the model, given as:

$$\Omega(f) = \gamma T + \frac{1}{2} \lambda \|w\|^2 \tag{6}$$

where γ and λ control the penalty based on *T* and *w*, respectively.

Furthermore, an iterative method was used to minimize the objective function. The objective function which is minimized at *t*-th iteration is

$$\ell^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) \tag{7}$$

Using Taylor expansion, Eqn. (7) can be derived for loss reduction after the tree splits from the given node as

$$\ell_{split} = \frac{1}{2} \left[\frac{(\sum_{i \in I_L} g_i)^2}{\sum_{i \in I_L} h_i + \lambda} + \frac{(\sum_{i \in I_R} g_i)^2}{\sum_{i \in I_R} h_i + \lambda} - \frac{(\sum_{i \in I} g_i)^2}{\sum_{i \in I} h_i + \lambda} \right] - \gamma \tag{8}$$

where

$$g_i = \frac{\partial L(y_i, \hat{y}_i^{(t-1)})}{\partial \hat{y}_i^{(t-1)}}, h_i = \frac{\partial^2 L(y_i, \hat{y}_i^{(t-1)})}{\partial \hat{y}_i^{(t-1)}} \tag{9}$$

where I is a subset of the available observations in the current node, and I_R and I_L are subsets of the available observations in the left and right node after the split, respectively. The best split can be found using Eqn. (8) at any given node, which is based on the regularization parameter (λ) and the loss function.

The detailed derivation is presented by (Chen and Guestrin, 2016).

To evaluate the predictive accuracy of the XGBOOST algorithm, other classification algorithms such as random forest (RF), support vector machine (SVM), Logistic Regression (LR) and K-nearest neighbor (K-NN) were used as the evaluation criteria. The optimal parameter value of each algorithm was calculated using hyperparametric search. Accuracy was then used as the standard measure to assess the predictive precision of the model, which is calculated as follows：

$$ACCURACY = \frac{N_c}{N_a}$$

where N_c : Sample size of correct classification, and

N_a : Total sample size.

The accuracy of each classification algorithm using validation data at different stations is shown in Table 3 and Figure 3.

Table 3: NAT predictive accuracy using different classification algorithms

	RF	**XGBOOST**	**SVM**	**LR**	**KNN**
GZN	0.7711	**0.7766***	0.7520	0.6676	0.7084
QY	0.7105	**0.8005***	0.6972	0.5642	0.7864
YDW	0.7200	**0.7200***	0.7200	0.6400	0.6933
SG	0.6453	**0.6816***	0.6065	0.5375	0.6271
LCE	0.7573	**0.7908***	0.7414	0.6837	0.7774
CZW	0.7239	**0.7692***	0.6916	0.6099	0.7658
LYW	0.7173	**0.7589***	0.6922	0.6182	0.7543
HYE	0.7544	**0.7424***	0.6393	0.5773	0.7246
HSW	0.7316	**0.7677***	0.6677	0.6098	0.7231
ZZW	0.6799	**0.7266***	0.6173	0.6072	0.7165
CSS	0.6805	**0.7427***	0.6473	0.6017	0.6390

* indicate the best predictive accuracy

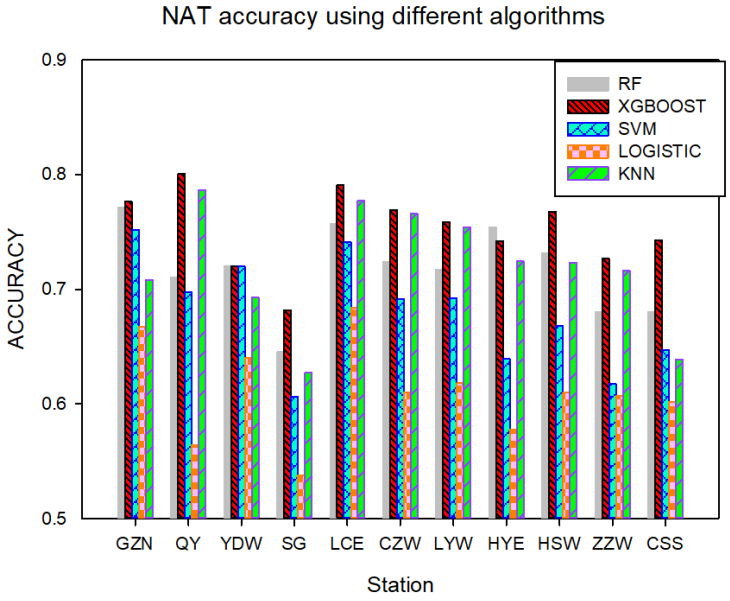


Figure 3: NAT predictive accuracy using different classification algorithms

The results show that (1) the XGBOOST algorithm has the highest accuracy at all stations in comparison with other algorithms; (2) the accuracy value of XGBOOST algorithm maintained high levels (up to 0.7) at all stations except at SG. This proves that the NAT predictive model based on the XGBOOST algorithm has good precision.

The timetable and infrastructure of the Wuhan-Guangzhou HSR station from 2015 to 2016 do not change significantly in comparison with 2018 data. Hence, the train operation data can be used as validation data to evaluate the precision of the model based on the data obtained from 2015 to 2016. Meanwhile, the data obtained from March to July 2018 were used as test data to evaluate the application of the model over time. The results of the predictive accuracy at different stations are shown in Figure 4.

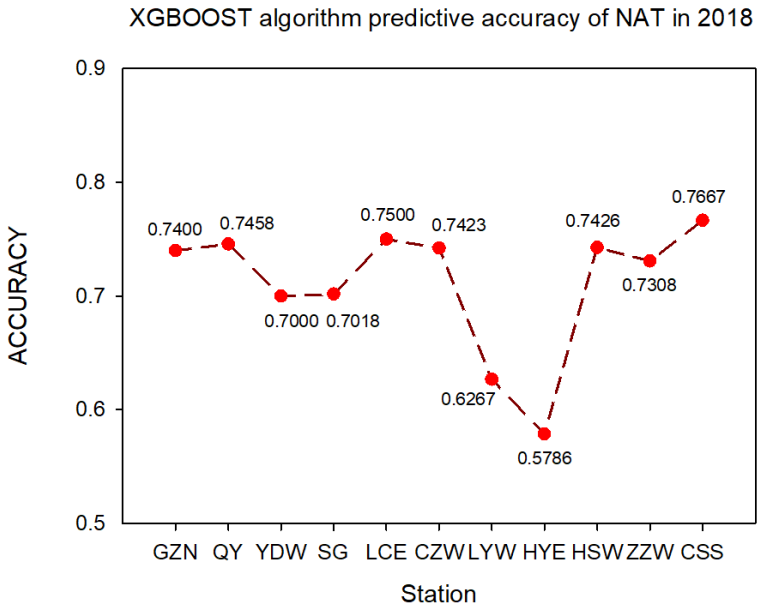


Figure 4: XGBOOST algorithm predictive accuracy of NAT in 2018

The model has a good precision and high accuracy (up to 0.7) in the stations at Wuhan-Guangzhou HSR station except for LYW and HYE. When this is combined with the accuracy values of the validation data, the results indicate that the model based on XGBOOST algorithm can accurately predict the number of affected trains by PD at Wuhan-Guangzhou HSR station.

4.2 The predictive model of TTAT

TTAT is another indicator that measures the severity of the PD influence. The overall scope of influence can be determined by combining TTAT and NAT results. The specific derivation process is described below:

Given a PD influence sequence, the TTAT and NAT are given as T_{td} and N_1, respectively, while the delay duration of i-th train is $T_{at}^{\ i}$. The discriminant relationship is obtained as follows:

IF $i=1$; THEN, the TTAT of the PD sequence is T_{td}, while NAT is N_1,

IF $1<i\le N_1$; THEN, the subsequent TTAT of the PD sequence is $T_{td}-\sum_{i=1}^{N_1} t_{at}^{\ i}$, while NAT is N_1-i.

The heatmap and 3D histogram of the TTAT for GZN station are shown in Figure 5.

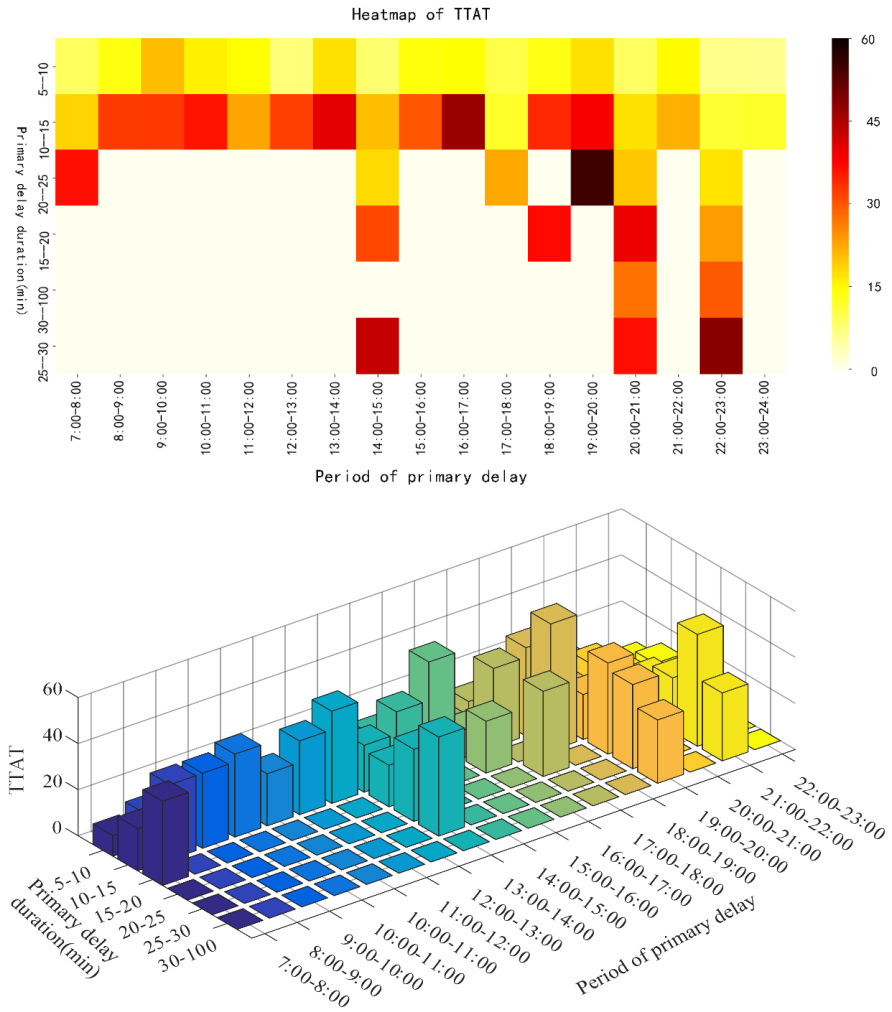


Figure 5: The TTAT heatmap and 3D histogram of GZN station

Because TTAT strongly depends on NAT, the predictive model is established based on the NAT model. Thus, the prediction is set as S' and Y for NAT and TTAT, respectively. Hence, TTAT predictive model is expressed as

$$Y = \varphi(D, B, T, I, N, S') \tag{10}$$

φ is a regression algorithm as TTAT is a continuous variable. The TTAT model was established using SVR, and compared with several algorithms such as RF, XGBOOST, Ridge regression (Ridge), and Lasso regression (LASSO)

Given a data set $D = \{(x_i, y_i) : i = 1, 2, ...n, x_i \in R^m, y_i \in R\}$, where x_i denotes the i

nput and y_i the output of the sample. The goal of SVR is to find a function $f(\mathbf{x})$ t hat has the most deviation (ε) from the actual and predicted values. $f(\mathbf{x})$ is defined as $f(\mathbf{x}) = w^{\mathrm{T}}\mathbf{x}+b$, where w is a hyperplane direction and b is an offset scalar.

The objective function is expressed as follows：

$$\begin{aligned} &\min_{w,b,\xi_i,\xi_i^*} = \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{m}\left(\xi_i + \xi_i^*\right) \\ &\textbf{s.t.} \begin{cases} -\varepsilon - \xi_i^* \le f\left(\mathbf{x}_i\right) - y_i \le \varepsilon + \xi_i \\ \xi_i^*, \xi_i \ge 0, i = 1, 2, ..., m \end{cases} \end{aligned} \tag{11}$$

where *C* is a penalty factor which determines the trade-off between the flatness of *f* and the values to which deviations larger than ε are tolerated. The ε-insensitive loss function $|\xi|_\varepsilon$ is given as

$$|\xi|_\varepsilon := \begin{cases} 0, & if\ |\xi| \le \varepsilon; \\ |\xi| - \varepsilon, & otherwise. \end{cases} \tag{12}$$

Using Lagrange multipliers, Eqn. (11) can be expressed as

$$\begin{aligned} &L(w, b, \alpha, \alpha^*, \xi_i, \xi_i^*, u, u^*) = \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{m}\left(\xi_i + \xi_i^*\right) - \sum_{i=1}^{m} u_i\xi_i - \sum_{i=1}^{m} u_i^*\xi_i^* \\ &+\sum_{i=1}^{m}\alpha_i\left(f(\mathbf{x}_i) - y_i - \varepsilon - \xi_i\right) + \sum_{i=1}^{m}\alpha_i^*\left(f(\mathbf{x}_i) - y_i - \varepsilon - \xi_i^*\right) \end{aligned} \tag{13}$$

The optimal solution can be obtained by solving Eqn. (13) to yield

$$\begin{aligned} &f(\mathbf{x}) = \sum_{i=1}^{m}(\alpha_i^* - \alpha_i)\mathbf{x}_i^{\mathrm{T}}\mathbf{x} + b \\ &b = y_i + \varepsilon - \sum_{i=1}^{m}(\alpha_i^* - \alpha_i)\mathbf{x}_i^{\mathrm{T}}\mathbf{x} + b \end{aligned} \tag{14}$$

The detailed derivation is presented by (Smola and Schölkopf, 2004). To evaluate the model, Lessthan5 variable was defined which is given as

$$Lessthan5 = \frac{N_d}{N_a}.$$

where N_d : The sample size of the absolute value of the difference between the actual and predicted values in less than 5 min.

N_a : Total sample size.

The optimal parameter value of each algorithm was calculated using hyperparametric search. The Lessthan5 value of each algorithm is shown in Table 4 and Figure 6.

Table 4: TTAT Lessthan5 value using different algorithms

	RF	XGBOOST	SVR	Ridge	LASSO
GZN	81.638	81.638	**85.311***	79.096	80.508
QY	77.526	77.526	**78.739***	76.395	77.850
YDW	70.000	70.000	**74.286***	70.000	71.429
SG	74.444	74.444	**76.173***	73.827	74.321
LCE	83.761	83.761	**84.444***	78.291	79.915
CZW	76.590	76.590	**77.009***	72.676	72.467
LYW	76.410	76.410	**77.098***	72.765	73.040
HYE	73.829	73.829	**74.582***	72.324	71.739
HSW	74.917	74.917	**75.116***	69.927	69.661
ZZW	76.362	76.362	**76.510***	72.680	71.355
CSS	80.090	80.090	**81.900***	78.281	78.281

* indicate the maximum Lessthan5 value in different regression algorithms

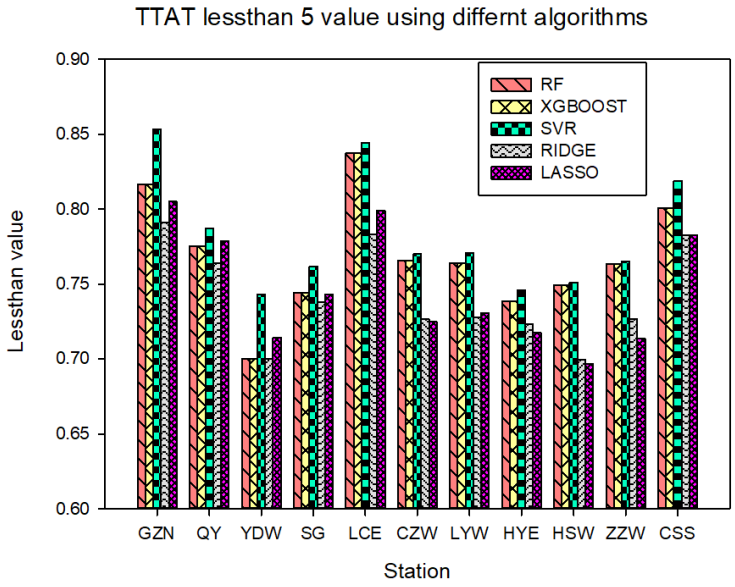


Figure 6: Lessthan5 value of TTAT using different algorithms

The results indicate that (1) the SVR algorithm has the highest Lessthan5 value at the stations in comparison with other algorithms. This proves that the SVR algorithm is the best algorithm for the TTAT predictive model. (2) The TTAT predictive accuracy of SVR algorithm at all stations was ~0.74, which proves that the SVR algorithm has good predictive accuracy.

Furthermore, 2018 data were used as the validation data to evaluate the application of TTAT model over time. The Lessthan5 values of the validation data for Wuhan-Guangzhou HSR stations are shown in Figure 7.

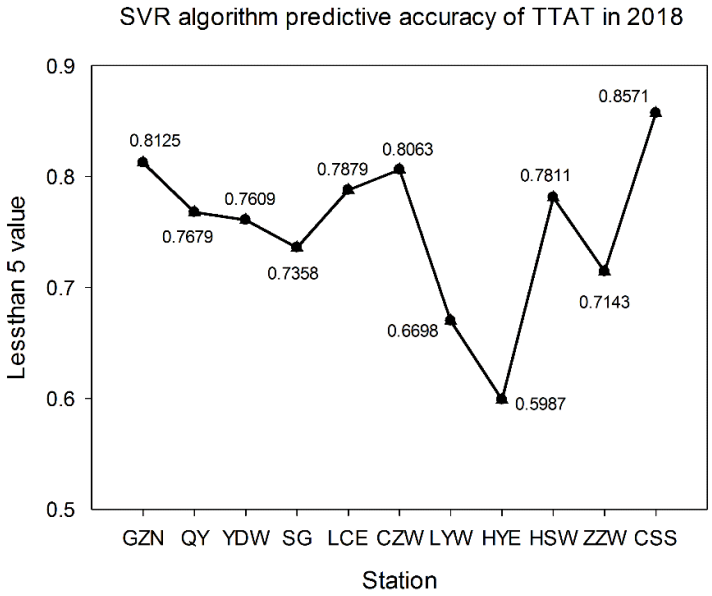


Figure 7: SVR algorithm predictive accuracy of TTAT in 2018

The TTAT predictive model has good predictive accuracy (~0.71) in most of the stations except LYW and HYE. The low precision of TTAT model for LYW and HYE is due to the low precision of NAT predictive model for these stations.

5. CONCLUSION

Prediction of the severity of PD influence in a station can assist the train dispatcher to develop rescheduling strategies and adjust the work plan of the station accordingly. The NAT and TTAT influence factors were determined by analyzing the mechanism of the PD propagation process. Moreover, the NAT and TAT predictive models were established and compared with several algorithms using the influence factors as model input. Data obtained from March 2015 to November 2016 were used to establish the models while the application of the models over time were evaluated using 2018 data. The main conclusions are as follows:

(1) NAT predictive model has a good predictive accuracy at Wuhan-Guangzhou HSR station based on the XGBOOST algorithm. When 2018 data were used as the test data, the results showed the NAT predictive model had a good application over time.

(2) NAT prediction results were used as the input values of the TTAT predictive model. The TTAT model was established using the SVR algorithm and compared with other regression algorithms. Furthermore, 2018 data were used as test data to test the application of TTAT model over time. The results indicate that the TTAT predictive model also has a good predictive accuracy over time.

(3) When a PD occurs, the influence scope can be obtained accurately using the NAT

and TTAT predictive models at each station. This provides a theoretical background needed by the dispatcher to develop rescheduling strategies and adjust the station work plan accordingly.

ACKNOWLEDGMENT

This work was supported by the National Nature Science Foundation of China [grant number 71871188]; the Science & Technology Department of Sichuan Province [grant number 2018JY0567]; We are grateful for the contributions made by our project partners.

REFERENCE

Chapuis, X., 2017. "Arrival Time Prediction Using Neural Networks", In: *7th International Conference on Railway Operations Modelling and Analysis. Lille (France): International Association of Railway Operations Research*, pp. 1500-1510.

Chen, T., Guestrin, C., 2016. Xgboost: "A scalable tree boosting system",In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. ACM, pp. 785-794.

Corman, F., Kecman, P., 2018. "Stochastic prediction of train delays in real-time using Bayesian networks". *Transportation Research Part C: Emerging Technologies* 95, 599-615.

Goverde R M P. Punctuality of railway operations and timetable stability analysis[D]. TU Delft, Delft University of Technology, 2005.

Goverde, R.M., 2007. "Railway timetable stability analysis using max-plus system theory". *Transportation Research Part B: Methodological* 41(2), 179-201.

Goverde, R.M., Corman, F., D'Ariano, A., 2013. "Railway line capacity consumption of different railway signalling systems under scheduled and disturbed conditions". *Journal of Rail Transport Planning & Management* 3(3), 78-94.

Hansen, I.A., Goverde, R.M., van der Meer, D.J., 2010. "Online train delay recognition and running time prediction",In: *Intelligent Transportation Systems (ITSC), 2010 13th International IEEE Conference on*. IEEE, pp. 1783-1788.

Huisman, T., Boucherie, R.J., 2001. "Running times on railway sections with heterogeneous train traffic". *Transportation Research Part B: Methodological* 35(3), 271-292.

Huisman, T., Boucherie, R.J., Van Dijk, N.M., 2002. "A solvable queueing network model for railway networks and its validation and applications for the Netherlands". *European Journal of Operational Research* 142(1), 30-51.

Kecman, P., Corman, F., Meng, L., 2015. "Train delay evolution as a stochastic process",In: *6th International Conference on Railway Operations Modelling and Analysis-RailTokyo2015*.

Kecman, P., Goverde, R.M., 2015. "Online data-driven adaptive prediction of train event times". *IEEE Transactions on Intelligent Transportation Systems* 16(1), 465-474.

Lessan, J., Fu, L., Wen, C., 2018. "A hybrid Bayesian network model for predicting delays in train operations". *Computers & Industrial Engineering*.

Marković, N., Milinković, S., Tikhonov, K.S., Schonfeld, P., 2015. "Analyzing passenger train arrival delays with support vector regression". *Transportation Research Part C:*

Emerging Technologies 56, 251-262.
Meester, L.E., Muns, S., 2007. “Stochastic delay propagation in railway networks and phase-type distributions”. *Transportation Research Part B: Methodological* 41(2), 218-230.
Milinković, S., Marković, M., Vesković, S., Ivić, M., Pavlović, N., 2013. “A fuzzy Petri net model to estimate train delays”. *Simulation Modelling Practice and Theory* 33, 144-157.
Pongnumkul, S., Pechprasarn, T., Kunaseth, N., Chaipah, K., 2014. “Improving arrival time prediction of Thailand's passenger trains using historical travel times”, In: *Computer Science and Software Engineering (JCSSE), 2014 11th International Joint Conference on*. IEEE, pp. 307-312.
Şahin, İ., 2017. “Markov chain model for delay distribution in train schedules: Assessing the effectiveness of time allowances”. *Journal of Rail Transport Planning & Management* 7(3), 101-113.
Smola, A.J., Schölkopf, B., 2004. “A tutorial on support vector regression”. *Statistics and computing* 14(3), 199-222.
Tang, Y., Wen, C., Huang, P., Li, Z., Li, J., Yang, Y., 2018. Support Vector Regression Models for Influenced Time Prediction in High-Speed Rail System. In: *Transportation Research Board 97th Annual Meeting*, Washington DC, United States.
Wen, C., Li, Z., Lessan, J., Fu, L., Huang, P., Jiang, C., 2017. “Statistical investigation on train primary delay based on real records: evidence from Wuhan–Guangzhou HSR”. *International Journal of Rail Transportation* 5(3), 1-20.
Wen, C., Li, Z., Lessan, J., Fu, L., Huang, P., Jiang, C., Muresan, M.I., 2018. Analysis of Causes and Effects of Primary Delays in a High-Speed Rail System. In: *Transportation Research Board 97th Annual Meeting*, Washington DC, United States.
Yaghini, M., Khoshraftar, M.M., Seyedabadi, M., 2013. “Railway passenger train delay prediction via neural network model”. *Journal of advanced transportation* 47(3), 355-368.
Yuan, J., Goverde, R., Hansen, I., 2002. “Propagation of train delays in stations”. WIT Transactions on The Built Environment 61.

Passenger Flow Control with Multi-station Coordination on an Oversaturated Urban Rail Transit Line: A Multi-objective Integer Linear Programming Approach

Denghui Li [a,1], Qiyuan Peng [a,2], Gongyuan Lu [a,3]
[a] School of Transportation and Logistics, Southwest Jiaotong University, Chengdu, China
[1] E-mail: lidenghuildh1993@163.com
[2] E-mail: qiyuan-peng@home.swjtu.edu.cn
[3] E-mail: lugongyuan@home.swjtu.edu.cn, Phone: +0086-13880609100

Abstract
With the booming travel demands in megacities, the limited transportation capacity hasn't satisfied them in urban rail transit. Passenger congestion problem become increasingly serious, which causes the potential accident risks on platforms. To further efficiently improve the conditions, this paper proposes an effective collaborative optimization method for the accurate passenger flow control strategies on an oversaturated urban rail transit line by simultaneously adjusting the number of inbound passengers entering multiple stations on the line. Through considering the space-time dynamic characteristics of passenger flow, a multi-objective integer linear programming model is formulated to minimize the number of passengers who are limited to enter stations, minimize the total passenger waiting time on platforms at all of involved stations where the optimal passenger flow control is imposed and maximize the passenger person-kilometres. And it is solved by CPLEX solver efficiently. Moreover, the passenger flow demands are time-variant, so it's very necessary for the accurate and easy-to-implement passenger flow control strategies to determinate the control time intervals. Hence, this paper develops a method based on Fisher optimal division to get an optimal determination of the control time intervals before solving the model. Finally, two sets of numerical experiments, including a small-scale case and a real-world instance with operation data of Chengdu metro system, are implemented to demonstrate the performance and effectiveness of the proposed approach.

Keywords
Urban rail transport, Passenger flow control, Multi-station coordination, Fisher optimal division method

1 Introduction

1.1 Background

With the acceleration of urbanization process and the drastic increase of urban population in China, urban rail transit (URT) transportation capacity has been unable to satisfy the booming travel demands in some cities (i.e. Beijing, Shanghai, etc.), especially in peak hours. Passenger congestion problem is becoming more and more serious, and it gradually affects the operation safety and reliability of URT. At the same time, due to the limitation of infrastructures, transportation capacity cannot be improved in the short term. Therefore, under the condition of the limited transportation capacity, it is urgent to develop a management strategy to relieve congestion and further improve the operation efficiency.

So passenger flow control becomes a better choice with the limited transportation

capacity at present. In fact, the measures for passenger flow control have been taken widely in some cities (i.e. Beijing, Shanghai, et al.) in China. For example, passenger flow controls will be imposed when the ratio of the number of the passengers entering certain station to the maximum number that the station holds gets to 70% (Xu et al., 2016). In the daily operation, passenger flow control works by setting railings outside metro station, reducing the number of the used gates and slowing down the speed of the escalators to limit the number of passengers entering the platforms and relieve pressure on platforms. However, the control strategies are implemented at each station without coordination, respectively, and mainly depend on the staffs' subjective work experience currently, which is lack of mathematical programming and scientific method (Jiang et al., 2018).

1.2 Literature Review

On the level of passenger flow control, many researches have ever investigated it from different perspectives in recent years, including pedestrian boarding/alighting management, station capacity and station pedestrian management.

In pedestrian boarding/alighting management, Baee et al. (2012) proposed different boarding/alighting strategies for Tehran subway system to increase satisfaction level and service success rate while reducing travel time by simulation; Fernández et al. (2015) demonstrated the existence of pedestrian saturation flows in public transport doors and showed various capacities of train doors under different conditions by real-scale experiments.

In station capacity, Chen et al. (2012) proposed a M/G/c/c-based capacity model of staircases and corridors for passenger evacuation in consideration of space facility in metro stations through analysis of passenger movements; Xu et al. (2014) developed a SSC optimization model of station capacity according to the gathering and scattering process and the analytical queuing network to identify bottleneck facilities to improve capacity; Xu et al. (2016) proposed an approach to measure a transfer metro station capacity for different ratio of inbound, outbound and transfer passenger volumes to the total passenger volume, according to the passenger routes in the station.

In station pedestrian management, Hoogendoorn and Bovy (2004) put forward a new theory of pedestrian behaviors under uncertainty based on the concept of utility maximization to simultaneously optimize route choice, activity area choice, and activity scheduling using dynamic programming for different traffic conditions and uncertainty levels based on this normative theory; Davidich et al. (2013) evaluated the impact of waiting pedestrians and proposed a cellular automata model for waiting pedestrians to analysis and prediction waiting zone capacity in critical situations.

The above studies are in-depth in theory, and provide a theoretical basis and reference on the practical application. However, they focus on passenger flow control at a single station or several stations, the complex characteristics of passenger flow Origin and Destination (OD) and the passenger flow at other stations in the network aren't taken into consideration, which tends to result in the section capacity not being fully utilized and is bad to improve the service quality and economic benefits in the entire network.

In view of the above deficiencies, some researchers turn their attentions to the passenger flow control with multi-station coordination on an oversaturated line (Wang et al., 2015; Li et al., 2017; Shi et al., 2018; Jiang el at., 2018). For example, Wang et al. (2015) took average passenger delay as the objective to develop an integer programming model based on the analysis of passenger delay and the processes by which passengers alight and board, which aims to disperse the pressure of oversaturated stations into others and achieve the

optimal state for the entire line, and model is verified by a case study; Jiang et al. (2018) proposed a method based on reinforcement learning to optimize the inflow volume during a certain period of time at each station with the aim of minimizing the safety risks imposed on passengers at the metro stations, and the performance of the approach was tested by the simulation experiment carried out on a real-world metro line in Shanghai.

In addition, some studies mainly proposed the stop-skipping strategy to enhance the operation efficiency and indirectly relieve the pressure caused by the huge passenger volume (Wang et al., 2014; Niu et al., 2015). Nevertheless, trains always run with relatively fixed all-stop patterns from the start station to the terminal station in URT.

To our knowledge, the majority of existing studies focus on passenger flow control at a single station or several stations while the researches on the passenger flow control at multi-station coordination on the line is relatively few. Also, there are still deficiencies in the existing studies on the passenger flow collaborative control. For example, there is no scientific method to determine the control time intervals while the proposed approach in this paper addresses precisely these gaps.

.

1.3 Contributions

The proposed approach contributes to the state-of-the-art related passenger flow control research in three ways.

(1) In order to obtain accurate and easy-to-implement passenger flow control strategies, the method , Fisher optimal division ,to determine the control time intervals is proposed according to the historical passenger flow data.

(2)To take the interests of both passengers and operators into consideration simultaneously, based on characteristic of passenger flow OD and dynamic passenger demands, a multi-objective integer linear programming model is proposed to minimize the number of passengers limited to enter stations, minimize the total passenger waiting time on platforms at all of involved stations and maximize the passenger person-kilometres.

(3) Train dwelling time has an important influence on the boarding/alighting behaviors at stations, so it is taken into consideration and is taken as an important constraint in the proposed model.

The rest of the paper is structured as follows. Section 2 provides the concrete definition of passenger inflow control with multi-station coordination problem. Section 3 describes the integer linear programming models taking the characters of passenger flow into account. In Section 4, solution approaches, including the methods to determinate the control time intervals and solve the model, are described. Two sets of numerical experiments, including a small-scale case and a real-world instance with operation data of Chengdu metro system, are carried out to verify the effectiveness of these models in Section 5. Finally, conclusions and further studies are presented in Section 6.

2 Problem Description

2.1 Descriptions of the Passenger Flow Control System

This study considers a single-direction oversaturated urban rail transit line with n stations,

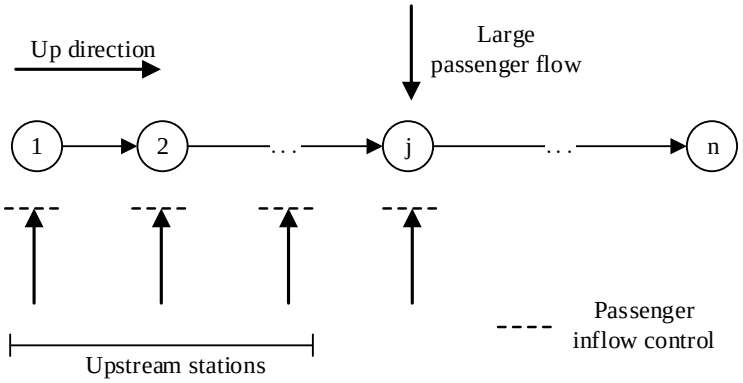


Figure 1: The illustration of urban rail transit line

high-frequency services and lack of train capacity in peak hours. The illustration of the line is shown in Figure 1. The line consists of stations and sections between these adjacent stations. Along certain operational direction (such as up direction.), the stations are numbered as 1, 2, ..., n consecutively, in which stations 1 and n represent the start and terminal stations respectively of in-service trains. For clarity, we use a set $N = \{1,2, \dots, n\}$ to identify stations in this direction. And trains run along the direction on the line according to fixed schedules.

Passenger flow collaborative control is related to the remaining loading capacity, which is dependent on the boarding and alighting passengers at upstream stations. As shown in Figure 1, it is assumed that there is large passenger flow at station *j* in peak hours. In order to rapidly evacuate passengers at this station, and guarantee operational safety and high service quality, the enough remaining loading capacity is necessary for this station. Therefore, the upstream stations, besides this station, start to limit the number of the inbound passengers according to the passenger demands at station *j* and other stations. For each station where passenger flow control is imposed, the inbound passengers are limited to queue at station entrances or gates, which reduce the congestion on platforms.

It is worth pointing out that the decrease of total demands will not be obvious after passenger flow control strategy is implemented. The ultimate objective for passenger flow control is to balance or reassign the limited transportation capacity among different stations (or sections) on the line, to achieve greater operation efficiency.

2.2 Definition of the Passenger Flow Control Problem

From the perspective of supply and demands, the participants are mainly operators and passengers in URT. When the passenger flow control strategies are formulated, the interests of both operators and passengers should be considered. For the passengers, they expect to quickly get to the destination from the departure, which can be reflected indirectly by the service quality of the operators. For the operators, increasing their revenues is an important goal. Since the revenues of URT are correlated with the travel distance and volumes of passengers, the passenger person-kilometres is the best indicator to measure their economic benefits.

As mentioned above, the passenger flow coordinated control is defined as a control

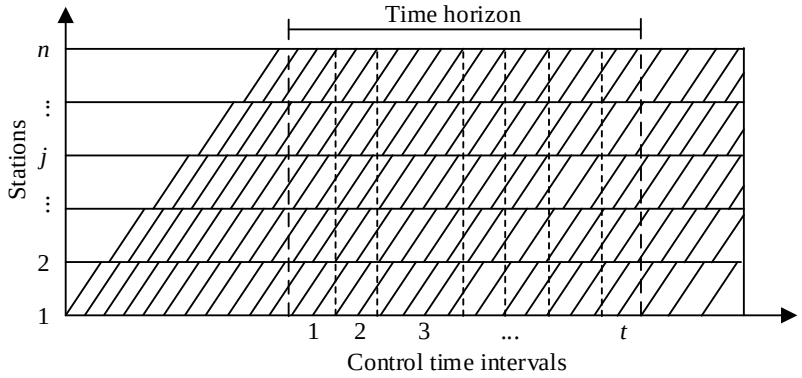


Figure 2: The illustration of the determination of control time intervals

congestion problem on the URT line to minimize the number of the passengers limited to enter the stations, minimize the total waiting time of stranded passengers and maximize the passenger person-kilometres in this paper. Note that the stranded passengers refer to passengers who cannot board the previous trains due to the limited train loading capacity and stay on platforms to wait the following trains.

Meanwhile, in order to guarantee the feasibility of the coordinative control strategies, some systematic constraints should be formulated, such as train capacity, train dwelling time, station design passing capacity, platform capacity and passenger demands, etc.

2.3 Descriptions of Determination of the Control Time Intervals

Since the passenger demands are time-variant, in order to obtain the accurate and easy-to-implement passenger flow control strategies, it is necessary to discretize the considered time horizon into several control time intervals, so that the control time intervals and the time-variant characteristics of the passenger demands are as close as possible. Hence, we disperse the continues time horizon into a finite number of control time intervals. And these intervals will be denoted by a set $T = \{1,2, \dots, t\}$. The illustration of the determination of control time intervals is shown in Figure 2.

For the determination of control time intervals, it is a great ideal to take the train departure interval as control time intervals. However, in the daily operation, it is more difficult to formulate and implement the passenger flow control strategies in such a short interval, in which there is also no statistical laws for passenger flow. In addition, if each control time interval is too long, the passenger flow control strategy cannot be flexibly adjusted as the passenger demands change, which leads to losing the meaning of passenger flow control. Therefore, it is necessary to develop a scientific method to scientifically determine the number of control time intervals and the length of each interval. And the specific method to determinate the control time intervals is described in Section 4.

In summary, the problem of the passenger flow coordinative control on an urban rail transit line can be summarized as:

On an over-saturated urban rail transit line, according to the space-time distribution dynamic characteristics of passenger flow, all stations on the line are associated in the space and time dimensions; meanwhile, according to the historical data of passenger flow, the

considered time horizon is dispersed into several control time intervals; Under the constraints of train capacity, train dwelling time, station design passing capacity, platform capacity and passenger demands, each station adjusts the number of the inbound passengers by the collaborative control strategies in each control time interval on the line, to make urban rail transit systems achieve greater transportation efficiency.

Table 1: Sets, subscripts, input parameters and decision variables

Symbols	Descriptions
Sets and subscripts	
N	Set of stations.
T	Set of control time intervals.
i, j, k	Index of stations, $\forall i, j, k \in N$.
t	Index of control time intervals, $\forall t \in T$.
Input parameters	
$\alpha_{i,t}$	The ratio of the passenger volumes choosing certain operating direction to the total passenger volumes entering station i in interval t.
$Q_{i,t}$	The maximum number of passengers loaded by trains at station i in interval t.
L	The maximum loading capacity per train.
Z	The maximum number of passengers who can enter stations within an hour.
$\beta_{i,j,t}$	The ratio of the number of passengers from station i to j to the number of boarding passengers at station i in interval t.
P	The maximum number of passengers that the platform can hold.
$A_{i,t}$	The maximum number of passengers who can be accumulated on the platform at station i in interval t.
t^{s}	The time spent boarding the trains per person.
t^{x}	The time spent alighting from the trains per person.
$t_{i,t}^{\mathrm{dwell}}$	The average dwelling time of train at station i in interval t.
t_{o}	The time it takes for the trains to open and close the doors.
B	The number of carriages per train.
m	The number of doors of each train carriage.
$n_{i,t}$	The number of trains passed station i in interval t.
$C_{i,t}$	The maximum number of passengers allowed to enter station i in interval t.
$D_{i,t}$	The passenger demands at station i in interval t.
d_i	The length of section i.
θ	The minimum ratio of the number of the inbound passengers to the total passenger demands at each station in any interval.
ΔT_t	The length of control time interval t.
Decision variables	
$x_{i,t}$	The total number of passengers entering station i in interval t, including up and down direction.
$x_{i,t}^{\mathrm{s}}$	The number of boarding passengers at station i in interval t.
$x_{i,t}^{\mathrm{x}}$	The number of passengers alighting from trains at station i in interval t.
$S_{i,t}$	The number of passengers stranded on the platform at station i in interval t.
$l_{i,t}$	The number of passengers in trains when leaving the station i in interval t.
$\omega_{i,t}$	The rate of passenger flow control at station i in interval t.

3 Mathematic Model

3.1 Assumptions and Notations

The following assumptions are made in this paper and notations of the model and their descriptions are shown in Table 1.

(1) Trains run well on the line with fixed schedules.

(2) Travel demands and the characteristics of passenger flow OD are known in each control time interval, and passenger flow OD can be obtained from historical data.

(3) Passengers follow the principle of "alighting first, boarding later" in the whole boarding/alighting activities.

(4) There is no passengers stranded on platforms at all stations before the first control time interval.

(5) All passengers board trains without the passengers stranded on the platform when the remaining train capacity is more than the number of passengers accumulated on the platform at each station in each control time interval during train operation.

(6) This paper studies the passenger flow control on an urban rail transit line rather than a single station, so the influences of the capacity of station gates, staircases, escalators and elevators are weakened. And the indicator, the maximum number of passengers who can enter the station within an hour, is adopted to replace them.

(7)The process from the entrance facilities to the platform will not be considered. That is, passengers allowed to enter the station can arrival at the platform immediately. This similar assumption can also be found in by Shi et al. (2018).

(8)The passengers are evenly distributed on platforms with the staffs' guidance before trains arrive at stations in peak hours.

3.2 Decision Variables and the Related Expressions

The decision variables in this paper are defined in Table 1, and the related expressions are shown in formula (1)-(5). Note that all the decisions variables are non-negative integers.

(1) The number of passengers alighting from trains

The number of alighting passengers at the station i equals to the sum of the products of the number of boarding passengers at each upstream station and the alighting rate from it to i during interval t. And it can be expressed as:

$$x_{i,t}^{\mathrm{x}} = \sum_{k=1}^{i-1} x_{k,t}^{\mathrm{s}} \cdot \beta_{k,i,t}. \tag{1}$$

Where $\beta_{i,j,t}$ is known and given according to the historical data from Automatic Fare Collection (AFC).

(2) The number of passengers stranded on platforms

According to the principle of flow conservation, the number of passengers stranded on the platform at station i during interval t can be expressed as:

$$S_{i,t} = \begin{cases} x_{i,t} \cdot \alpha_{i,t} - x_{i,t}^{\mathrm{s}}, & t = 1 \\ S_{i,t-1} + x_{i,t} \cdot \alpha_{i,t} - x_{i,t}^{\mathrm{s}}, & t > 1 \end{cases}. \tag{2}$$

Where $\alpha_{i,t}$ is known and given according to the historical data from AFC.

(3) The number of passengers in trains

According to the principle of flow conservation, the number of passengers in trains when leaving station i during interval t can be expressed as:

$$l_{i,t} = \begin{cases} x_{i,t}^{\mathrm{s}}, & i = 1 \\ l_{i-1,t} + x_{i,t}^{\mathrm{s}} - x_{i,t}^{\mathrm{x}}, & i > 1 \end{cases}. \tag{3}$$

(4) The number of boarding passengers

The number of boarding passengers is the minimum of both the remaining loading capacity and the number of passengers accumulated on the platform at station i during interval t. For the start station, there is no alighting passengers. And there is no passengers in trains before trains arrival at the start station. So it can be given as:

$$x_{i,t}^{\mathrm{s}} = \min\{Q_t - l_{i-1,t} + x_{i,t}^{\mathrm{x}}, S_{i,t-1} + x_{i,t} \cdot \alpha_{i,t}\}. \tag{4}$$

(5) The rate of passenger flow control at stations

During the control time interval t, the rate of passenger flow control at station i is defined as the ratio of the number of passengers limited to enter this station to total passenger demands at this station during this control time interval.

$$\omega_{i,t} = \frac{D_{i,t} - x_{i,t}}{D_{i,t}}, i \neq n. \tag{5}$$

Where $D_{i,t}$ is known and given according to the historical data from AFC.

3.3 Constraints

The descriptions and the specific expressions the systematic constraints, including train capacity, platform capacity, train dwelling time, station design passing capacity, passenger demands, etc., are given in this subsection.

(1) Train capacity constraint

The number of passengers in trains should not exceed their maximum capacity in any control time interval. Note that the maximum capacity in any control time interval equals to the product of the number of trains passing station i in the interval and the maximum loading capacity of each train.

$$\begin{cases} l_{i,t} \leq Q_{i,t} \\ Q_{i,t} = n_{i,t} \cdot L \end{cases}. \tag{6}$$

(2) Platform capacity constraint

The number of passengers stranded on the platform should not exceed the platform capacity under the safe level at the end of the interval t. Meanwhile, the number of passengers accumulated on the platform will reach the maximum after passengers alight from trains. To ensure the safety, all the number of the inbound passengers, the stranded passengers and the alighting passengers combined should not exceed the maximum number of passengers who can be accumulated on the platform at station i in interval t.

$$\begin{cases} S_{i,t} \le P \\ x_{i,t} \cdot \alpha_{i,t} + S_{i,t-1} + x_{i,t}^{\mathrm{x}} \le A_{i,t} \\ A_{i,t} = n_{i,t} \cdot P \end{cases}. \tag{7}$$

(3) Train dwelling time constraint

When studying the influence of the large passenger flow on stations in peak hours, the passengers' boarding/alighting activities cannot be ignored. The trains will stay at each station for a period of time in order to complete the passengers' boarding/alighting service during their operation. According to the actual operation, passengers' boarding and alighting service is completed from the time when the doors are fully open to the time when the doors start to be closed after the train arrives at the station.

$$\frac{x_{i,t}^{\mathrm{s}} \cdot t^{\mathrm{s}} + x_{i,t}^{\mathrm{x}} \cdot t^{\mathrm{x}}}{n_{i,t} \cdot m \cdot B} \le t_{i,t}^{\mathrm{dwell}} - t_{\mathrm{o}}. \tag{8}$$

Where $t^{\mathrm{s}} = 0.76\mathrm{s/p}$, $t^{\mathrm{x}} = 0.55\mathrm{s/p}$, according to Cao (2009); Li (2011) pointed out that at the stations with screen doors, it takes $t_{\mathrm{o}} = 15s$ for the trains to open and close the doors; $t_{i,t}^{\mathrm{dwell}}$ can be determined according to the fixed schedules.

(4) Station design passing capacity constraint

The number of passengers entering the station at station i during interval t should not exceed the maximum number of passengers allowed to enter this station during this interval.

$$x_{i,t} \le C_{i,t}. \tag{9}$$

(5) Passenger demands constraint

The number of inbound passengers at station i during interval t should not be more than the realistic passenger demands in this interval. At the same time, there is still certain service ability at each station.

$$\begin{cases} x_{i,t} \ge \theta \cdot D_{i,t} \\ x_{i,t} \le D_{i,t} \end{cases}. \tag{10}$$

(6) Additional constraint

In order to reduce the waiting time of the passengers stranded on platforms as much as possible, it is necessary to ensure that the passengers stranded on platforms during interval t are served during interval $t+1$.

$$S_{i,t} \le x_{i,t+1}^{\mathrm{s}}, t < T. \tag{11}$$

3.4 Objective function

In this paper, the objective functions for the problem of the passenger flow collaborative control on an urban rail transit line are as follows:

$$\begin{cases} \min z_1 = \sum_{i=1}^{n} \sum_{t=1}^{T} \left(D_{i,t} - x_{i,t} \right) \\ \min z_2 = \sum_{i=1}^{n} \sum_{t=1}^{T} S_{i,t} \cdot \Delta T_t \\ \max z_3 = \sum_{i=1}^{n} \sum_{t=1}^{T} l_{i,t} \cdot d_i \end{cases}. \tag{12}$$

Where z_1 is defined to minimize the number of passengers limited to enter the stations; z_2 is defined to minimize the waiting time of the passengers stranded on the platform involved all stations; z_3 is defined to maximize the passenger person-kilometres.

The multi-objective integer linear programming model of the passenger flow collaborative control with multi-station on an URT line in this paper is formulated by combining (1)-(12).

4 Solution Approaches

Note that the control time intervals must be determinated before solving the model to obtain all the input parameters of the model. Therefore, the method to determinate the control time intervals is described firstly,and finally the method to solve the model is depicted.

4.1 Algorithm for Determining Control Time Intervals

In the determination of the control time intervals, the existing studies disperse the time horizon into equal intervals (10min or 15min) as the control time intervals, without giving a scientific method. Therefore, in order to determine the control time intervals that is consistent with the time-variant characteristics of passenger demands and develop the accurate and easy-to-implement passenger flow control strategies, based on Fisher optimal division method (Xiao et al., 2014), the optimal control time intervals can be obtained by cluster analysis on the historical inbound passenger flow time series during the time horizon in this paper. The specific steps are as follows.

Step 1 Constructing inbound passenger flow time series and matrix.

Firstly, the historical data of passenger flow during time horizon is counted at intervals of Δt, then the inbound passenger flow time series is obtained. Let $\boldsymbol{H}_t$ be the inbound passenger flow matrix for the interval t, then the time series of the inbound passenger flow during the time horizon can be expressed as:

$$\boldsymbol{H} = \{\boldsymbol{H}_1, \boldsymbol{H}_2, \dots, \boldsymbol{H}_T\}. \tag{13}$$

Where $\boldsymbol{H}_t$ is given as:

$$\boldsymbol{H}_t = (p_1^t \quad p_2^t \quad \dots \quad p_n^t)^{\mathrm{T}}. \tag{14}$$

Where p_i^t is defined as the number of the inbound passengers at station i in control time interval t.

Step 2 Calculating class diameter.

Let the class G contain samples $\{\boldsymbol{H}_{(i)}, \boldsymbol{H}_{(i+1)}, \dots, \boldsymbol{H}_{(j)}\}$, denoted as $\mathrm{G} = \{\boldsymbol{H}_{(i)}, \boldsymbol{H}_{(i+1)}, \dots, \boldsymbol{H}_{(j)}\}$, where $j > i$, and the mean vector of the class is given as:

$$\boldsymbol{H}_{\mathrm{G}} = \frac{1}{j-i+1}\sum_{t=i}^{j} \boldsymbol{H}_{(t)} . \tag{15}$$

Then the diameter of the class G is:

$$D(i,j) = \sum_{t=i}^{j}\left(\boldsymbol{H}_{(t)} - \boldsymbol{H}_{\mathrm{G}}\right)^{\mathrm{T}}\left(\boldsymbol{H}_{(t)} - \boldsymbol{H}_{\mathrm{G}}\right). \tag{16}$$

The $D(i,j)$ is obtained by Step 2.

Step 3 Calculating the classification loss function.

The classification loss function is calculated by the following recursive formula. When n and k are fixed, the smaller $L[b(n,k)]$ is, the smaller sum of deviation square of all classes is, and the more reasonable the classification is. Therefore, it is necessary to find a classification method to minimize $L[b(n,k)]$. $P(n,k)$ is denoted as the classification method that $L[b(n,k)]$ takes the minimum value.

$$\begin{cases} L[b(n,2)] = \min\{D(1,j-1) + D(j,n)\}, 2 \le j \le n \\ L[b(n,k)] = \min\{L[P(j-1,k-1)] + D(j,n)\}, k \le j \le n \end{cases} \quad (17)$$

The calculation steps are as follows:

- Calculate the optimal two-partition for $\boldsymbol{H} = \{\boldsymbol{H}_1, \boldsymbol{H}_2, \dots, \boldsymbol{H}_T\}$ according to (17);
- Calculate the optimal k-partition for $\boldsymbol{H} = \{\boldsymbol{H}_1, \boldsymbol{H}_2, \dots, \boldsymbol{H}_T\}$ according to (17).

The $L[b(n,k)]$ is obtained by the above steps.

Step 4 Determine the number of classifications-k .

Since the number of classifications- k cannot be predetermined, generally speaking, the inflection points in the k -changing trend diagram of $L[b(n,k)]$, can be used as the basis for determining the number of classifications- k. However, the inflection points maybe not unique. Therefore, In order to further determinate the number of classifications- k, this paper also uses an indicator, slope difference, to help find better k.

This paper calculates the slope difference between adjacent line segments in the graph according to the following formula. Note that $\gamma(k)$ is the slope difference between adjacent line segments at k.

$$\gamma(k) = \left|\frac{L[b(n,k-1)] - L[b(n,k)]}{(k-1)-k}\right| - \left|\frac{L[b(n,k)] - L[b(n,k+1)]}{k-(k+1)}\right|, 3 \le k \le n-1. \quad (18)$$

From the geometric sense, when $\gamma(k)$ reaches its maximum, point corresponding to k in the graph of $L[b(n,k)]\sim k$ divides the line in the graph into two parts. It is steep before this point and it is relatively flat after this point in the graph of $L[b(n,k)]\sim k$, which means there are few differences among classifications after this point and it is meaningless if it continues to be divided. That is, k is the better number of classifications at this moment. And the better k is obtained. In addition, when $\gamma(k)$ is close to 0, it no longer continues to be divided.

4.2 Model Solution

This model is a multi-objective integer linear programming model, which can be effectively handled through linear weighted methods in this paper.

5 Numerical Experiments

In this section, a series of numeric experiments, involving a small-scale case study and a real-world case study in Chengdu Metro System, are implemented to illustrate the applications of the proposed model. And the model is solved by CPLEX solver.

5.1 A Small-scale Case Study

In this part, we consider a single-direction line with 3 stations and 4 control time intervals, to test the performance of the model.

(1)Experiment descriptions and Parameter settings

For the convenience of description, we particularly name the stations A, B, C, and name the control time intervals T1, T2, T3, T4. In the experiments, the data is shown in Table 2. The train capacity is 80 persons, the length of each section is 3 km and the platform capacity is 140 persons.

Table 2: The data of the small-scale case

Origin	Interval number(Length)	Destination	$\beta_{i,j,t}$	$\alpha_{i,t}$	Demands
A	T1（6）	B	0.3	1	300
		C	0.7		
	T2（3）	B	0.4	1	200
		C	0.6		
	T3（6）	B	0.5	1	200
		C	0.5		
	T4（6）	B	0.6	1	100
		C	0.4		
B	T1（6）	C	1	0.8	200
	T2（3）	C	1	0.8	100
	T3（6）	C	1	0.8	100
	T4（6）	C	1	0.8	50

(2)Computational results analysis

The experiment is solved by CPLEX solver, and the optimal solution is obtained: $z_1 = 304$ persons, $z_2 = 2004$min, $z_3 = 3345$person-kilometres. The number of the boarding passengers, the alighting passengers, the passengers stranded on platforms and the passengers limited to enter the stations from the case is shown in Table 3.

Table 3: The results of the numerical example

Stations	Intervals	$x^{s}_{i,t}$	$x^{x}_{i,t}$	The number of the stranded passengers		$\omega_{i,t}$ (%)
				Platforms	Out of stations	
1	T1	155	0	0	145	48.3
	T2	80	0	60	60	30.0
	T3	160	0	100	0	0
	T4	160	0	40	0	0
2	T1	52	47	30	97	48.5
	T2	32	32	76	2	2.0
	T3	80	80	76	0	0
	T4	96	96	20	0	0
3	T1	0	160	0	0	0
	T2	0	80	0	0	0
	T3	0	160	0	0	0
	T4	0	160	0	0	0

The optimal solution is shown in Table 3. As seen in Table 3, the time intervals to implement the flow control measures at stations 1 and 2 is mainly in the first two control time intervals, and passengers entering the stations aren't limited in the last two intervals. During the first two control time intervals, passenger demands are large, and some passengers cannot enter the stations due to the limited train capacity and platform capacity. In the following intervals, passenger demands are decreased at each station, and these passengers can be gradually satisfied by the platform capacity. And the passenger flow control measures are gradually removed. Meanwhile, train capacity gradually satisfies the demands of the stranded passengers on the platform at each station, and the number of the stranded passengers are gradually decreased. During the last interval, there are still the passengers stranded on the platform at stations 1 and 2, and these passengers can be loaded by the following available trains. The experimental results shows that the model can accurately describes the behaviors of the passenger flow control for the problem of passenger flow coordination control on an urban rail transit line.

5.2 Numerical Experiments on Chengdu Metro Line 2

To further demonstrate the performance of the proposed model for large-scale problems, we next consider a real-world case study on the Chengdu metro line 2 with 32 stations. And we only consider the up direction. In the implementations, all the dynamic input parameters are obtained from AFC. In the following, we shall first give the detailed experimental descriptions and parameters.

(1)Experiment descriptions and parameter settings

In the experiments, the considered time horizon is set as 7:30-9:00, which is the morning peak-hours. And the stations are numbered as 1, 2, ..., 32 consecutively along up direction on the line, in which stations 1 and 32 represent the start and terminal stations respectively of in-service trains. Note that the transfer passenger flow is converted into the inbound passenger flow or the outbound passenger flow at the transfer station on the line. The parameters settings are as follows.

• The determination of control time intervals

The considered time horizon (7:30-9:00) is dispersed into 18 statistical intervals at intervals of $\Delta t = 5min$, to construct an inbound passenger flow time series $\boldsymbol{H} = \{\boldsymbol{H}_1, \boldsymbol{H}_2, \dots, \boldsymbol{H}_{18}\}$. Calculate it according to the steps in Subsection 4.1, and the $L[b(n,k)]\sim k$ trend diagram is obtained by using C#, as shown in Figure 3.

According to Fisher optimal division method, the optimal number of classifications is obtained at the inflection point of the curve. However, the inflection points are not unique (such as 4,5,6), as seen in Figure 3. So the slope difference between adjacent line segments is calculated according to the formula (18). And the $\gamma(k)\sim k$ trend diagram is plotted and shown in Figure 4.

As seen Figure 4, when k = 5, $\gamma(k)$ reaches the maximum. According to step 4 in subsection 4.1, the inbound passenger flow time series is divided into 5 clusters, and the results of classification are shown in Table 4.

According to the divided control time intervals, combing the historical AFC data and the schedules, $\beta_{i,j,t}$, $\alpha_{i,t}$, $D_{i,t}$ and $n_{i,t}$ can be obtained.

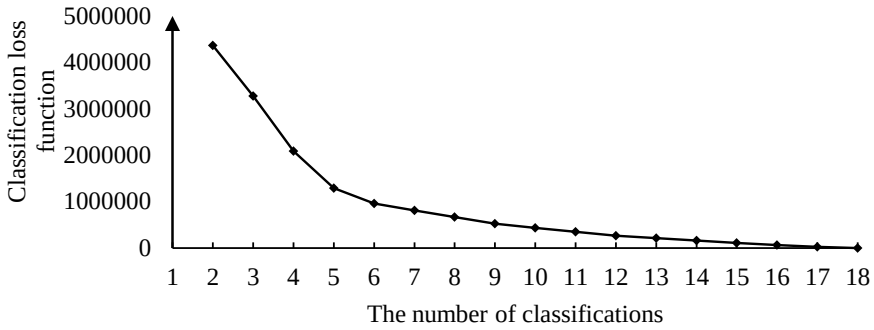


Figure 3: $L[b(n,k)]\sim k$ trend diagram

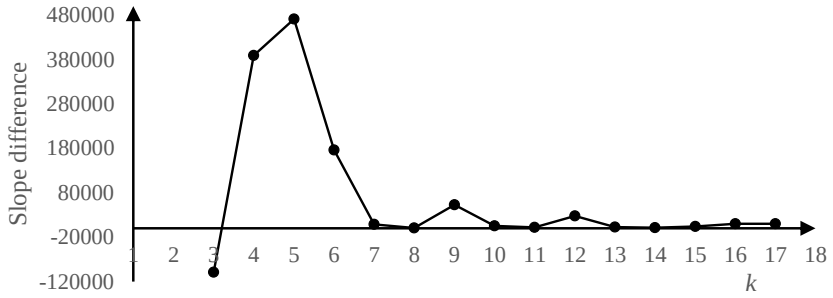


Figure 4: $\gamma(k)\sim k$ trend diagram

Table 4: The determination of the control time intervals

Classification number	Statistical time interval	Control time interval
1	T1-T4	7:30-7:50
2	T5-T7	7:50-8:05
3	T8-T11	8:05-8:25
4	T12-T15	8:25-8:45
5	T16-T18	8:45-9:00

• Model other parameters

The values of model other parameters are shown in Table 5.

Table 5: Model parameters

Parameters	Values
L	1888
Z	21000
P	1100
θ	0.5
Time horizon	7:30-9:00

(2) Computational results analysis between collaborative and non- collaborative control

The input parameters are obtained in (1), and the model is solved by CPLEX solver efficiently. The objective function values, the mean crowding degree on the platform at each station and the average number of passengers limited to enter stations at each station are selected as three indicators to make a comparative analysis between the passenger flow non-collaborative and collaborative control strategies.

The optimal solution, passenger flow collaborative control strategy, is shown in Table 6, and the objective function values are shown in Table 7.

Note that the passenger flow collaborative control strategy is obtained by solving the proposed model while the passenger flow non-collaborative control strategy is based on the idea that passengers at upstream stations board the trains with the priority and obtained by successive recursion method after giving the upstream stations greater weights under the constraints of train transportation capacity, train dwelling time and station design passing capacity constraints. And the objective function values of the passenger flow non-collaborative control strategy are shown in Table 7.

Table 6: Time-dependent and average inflow control rate of stations

Control time intervals	**Stations**							（%）
	8	9	10	11	12	13	15	16
7:30-7:50	49.9	—	34.9	47.6	—	—	20.5	—
7:50-8:05	49.9	48.8	49.3	49.3	7.6	1.2	10.3	—
8:05-8:25	49.9	48.2	48.6	50.0	36.1	0.7	24.1	—
8:25-8:45	—	—	50.0	—	33.7	—	37.6	9.4
8:45-9:00	—	—	—	—	18.0	—	30.5	28.9
The average control rate	29.9	19.4	36.6	29.4	19.1	0.4	24.6	7.7

As seen in Table 6, during the time horizon, the stations with the limited passenger flow are mainly from station 8 to 16, which is consistent with the practical situation. From the perspective of time, the control time interval is mainly 7:50-8:25. In the first four intervals, passenger flow control is imposed at the upstream stations, to ensure the rapid evacuation of the large passenger flow at stations 12, 15 and 16. In the last interval, passenger flow control is mainly implemented at stations 12, 15 and 16, to guarantee the rapid evacuation of the passengers stranded on platforms at the upstream stations. As time goes by, passenger demands are decreased. Some stations can remove passenger flow control measures or reduce passenger flow control intensity. During the last interval, some passengers at stations 12, 15 and 16 are still limited to enter the stations, and these passengers can be served by the following available trains. From the average rate of passenger flow control at the stations, passenger flow control intensity is large at stations 8, 10 and 11, which shows that these stations suffered from greater passenger flow organization pressure.

Table 7: The objective under the uncoordinated and coordinated conditions

Objective function	**Passenger flow non-collaborative control**	**Passenger flow collaborative control**
z1	25587	23190
z2	88650	0
z3	908175	909861

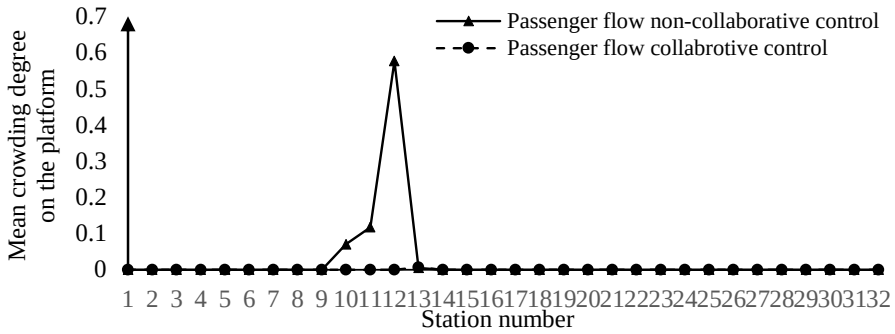


Figure 5: Mean crowding on the platform of stations

As seen in Table 7, the number of passengers limited to enter stations under the collaborative control is reduced by about 9.4% than that under the non-collaborative control. There are no passengers stranded on platforms under the collaborative control. And the passenger person-kilometres is increased by roughly 0.2% than that under the non-collaborative control. Overall, for the operation enterprises, the service quality can be intensely improved and effectively alleviate the contradiction between the limited transportation capacity and passenger demands under the collaborative control. Meanwhile, there is a small increase on the passenger person-kilometres. That is, the economic performance is also enhanced.

Mean crowding degree on the platform under passenger flow non-collaborative and collaborative control is shown in Figure 5. The mean crowding degree on the platform is defined as the ratio of the average number of the passengers stranded on the platform at each station in the five control time intervals to the capacity of the platform at each station in this paper, and this indicator can well reflect the passenger flow on the platform at each station in the whole time horizon.

As shown in Figure 5，the passengers are accumulated with an unbalanced situation at several stations under the non-collaborative strategies. For example, the mean crowding degree of platform at the station 12 is highest, which shows that the congestion situation is very serious on the platform. And stations 10 and 11 are less crowded, which represents that passenger flow pressure is relatively small on the platform at these stations; nevertheless, the mean crowding degree of platform at each station is 0 under passenger flow collaborative control, which shows that the proposed model can effectively reduce the number of the passengers stranded on the platform at each station and avoid potential accident risks caused by it.

The average number of the passengers limited to enter stations at each station under the passenger flow non-collaborative and collaborative control is shown in Figure 6. As seen in Figure 6, under the passenger flow non-collaborative control, the limited inbound passengers are mainly at stations 12, 15 and 16, and there is huge passenger volume at these stations, which shows these stations undertake the huge pressure caused by the large passenger flow. However, under the passenger flow collaborative control, the limited inbound passengers are mainly at stations from 8 to 16. And the number of the limited inbound passengers is reduced at stations 12 and 15 than that under the non-collaborative control, which shows that the proposed model can effectively balance the passenger flow pressure among all the stations and accelerate the evacuation of the passenger flow at the s-

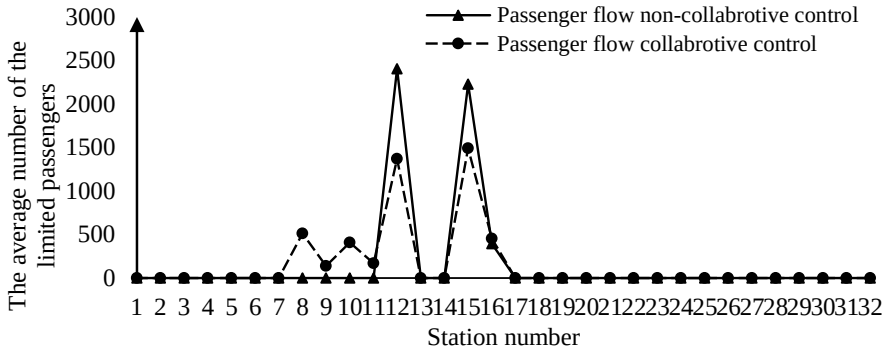


Figure 6: The average number of the passengers limited to enter stations

tations where there is the large passenger flow.

For the transfer passenger flow, since the cross-line passenger flow is converted into the passenger flow of the line in this paper, the transfer passenger flow and the inbound passenger flow are not distinguished at the transfer station during the calculation. Therefore, when the passenger flow control strategies are formulated, appropriate adjustments should be made based on the results according to the practical operation.

(3) Computational results analysis of different type of control time intervals

The control time interval is mainly classified two types, including the control time intervals obtained by Fisher optimal division method and equal control time intervals. This section mainly analyses the influences on the computational results of different type of control time intervals. For the former, the results are presented by (2); for the latter, the length of each control time interval is set as 15min, and the number of the intervals is 6, the determination of the latter is shown in Table 8. And the computational results of different type of control time intervals are as shown in Table 9.

Table 8: The determination of the equal control time intervals

Classification number	Statistical time interval	Control time interval
1	T1-T3	7:30-7:45
2	T4-T6	7:45-8:00
3	T7-T9	8:00-8:15
4	T10-T12	8:15-8:30
5	T13-T15	8:35-8:45
6	T16-T18	8:45-9:00

Table 9: The influences on the results of different type of control time intervals

Objective function	The control time intervals obtained by Fisher optimal division	The equal control time intervals (15min)
z1	23190	22967
z2	0	0
z3	909861	880818

In the following analysis, note that the former refers to the result of the model using control time intervals obtained by Fisher optimal division method; the latter refers to the result the model using equal control time intervals.

As seen in Table 9, the number of passengers limited to enter stations for the latter is reduced by about 1% than that of the former while the passenger person-kilometres for the former is increased by about 3.3% than that of the latter. And there are no passengers stranded on platforms for both the former and latter. Combining the Table 7, the number of passengers limited to enter stations for the latter is reduced by about 10.4% than that under the non-collaborative control while its passenger person-kilometres is reduced by roughly 3% than that under the non-collaborative control.

From the above analysis, it is concluded that the former is better than the latter in the trade-off between the service quality and the profits of management enterprise. That is, the results of the model using the control time intervals obtained by Fisher optimal division is better than that using equal control time intervals, which verifies the performance of the proposed approach.

6 Conclusions

In this paper, the problem of the passenger flow cooperative control on an urban rail transit line in peak hours is studied under the condition of limited transportation capacity. The mixed integer linear programming model of the passenger flow cooperative control on the line is developed. Through analysis of the instance, the objective function value, the mean crowding degree on platforms at all of involved stations and the average number of the passengers limited to enter stations are compared under the strategy obtained by solving the model and the non-collaborative control strategy. The results show that the former is better than the latter in the above three aspects, which tests the performance of the proposed approach and provides a theoretical basis and reference on the practical application.

This paper assumes that the trains run well on the line with the fixed schedules. However, a disturbance or disruption always happens inevitably in the daily operation. For example. There is an extension of dwelling time at a station with large passenger flow on the platform, which will impose influence on the train operation, such as delays. Under the circumstance in which a disturbance or disruption happens, train regulation is typically necessary. Therefore, further considering the disturbance or disruption in the train operation process, extending our work to jointly optimize train regulation and passenger flow control can be our future research directions.

Acknowledgements

The research was supported by National Key Research and Development Program of China (No. 2017YFB1200700-1) and The National Natural Science Foundation of China (No. U1834209).

References

Baee, S., Eshghi, F., Hashemi, S.M., Moienfar, R., 2012. “Passenger boarding/alighting management in urban rail transportation”, In: *2012 Joint Rail Conference. American Society of Mechanical Engineers*, pp. 823–829.

Chen, S., Liu, S., Xiao, X., Hong, J., Mao, B., 2012. “M/G/c/c-based model of passenger

evacuation capacity of stairs and corridors in metro stations", *Journal of the China railway Society*, vol.34, pp. 7–12.

Cao, S., 2009. *Analysis and modeling on passengers traffic characteristics for urban rail transit*. Ph.D. Dissertation, School of Traffic and Transportation, Beijing Jiaotong University, Beijing, China. (in Chinese)

Davidich, M., Geiss, F., Mayer, H.G., Pfaffinger, A., Royer, C., 2013. "Waiting zones for realistic modelling of pedestrian dynamics: a case study using two major German railway stations as examples", *Transportation Research Part C: Emerging Technologies*, vol. 37, pp. 210–222.

Fernández, R., Valencia, A., Seriani, S., 2015. "On passenger saturation flow in public transport doors", *Transportation Research Part A: Policy and Practice*, vol. 78, pp. 102–112.

Jiang, Z., Fan, W., Liu, W., Zhu, B., Gu, J., 2018. "Reinforcement learning approach for coordinated passenger inflow control of urban rail transit in peak hours", *Transportation Research Part C: Emerging Technologies*, vol. 88, pp. 1-16.

Li, S., Dessouky, M., Yang, L., Gao Z., 2017. "Joint optimal train regulation and passenger flow control strategy for high-frequency metro lines", *Transportation Research Part B: Methodological*, vol. 99, PP. 113-137.

Li, Z., 2011. "Research on dwelling time at Tianfu-Square station for Chengdu Metro Line 1", *Journal of Transportation Engineering and Information*, vol. 9, pp. 20-24. (in Chinese)

Hoogendoorn, S.P., Bovy, P.H.L., 2004. "Pedestrian route-choice and activity scheduling theory and models", *Transportation Research Part B: Methodological*, vol. 38, pp. 169-190.

Niu, H., Zhou, X., Gao, R., 2015. "Train scheduling for minimizing passenger waiting time with time-dependent demand and skip-stop patterns: Nonlinear integer programming models with linear constraints", *Transportation Research Part B: Methodological*, vol. 76, pp. 117-135.

Shi, J., Yang, L., Yang, J., Gao, Z., 2018. "Service-oriented train timetabling with collaborative passenger flow control on an oversaturated metro line: An integer linear optimization approach", *Transportation Research Part B: Methodological*, vol. 110, pp. 26-59.

Xiao, C., Gu, S., Gui, W., Gao, L., Li, Z., 2014. "Application of Fisher Optimal Partition Method in Division of Flood Season in Lixianjiang Basin", *Water Resources and Power*, vol. 32, pp.70-74. (in Chinese)

Xu, X., Liu, J., Li, H., Hu, J., 2014. "Analysis of subway station capacity with the use of queueing theory", *Transportation Research Part C: Emerging Technologies*, vol. 38, pp. 28–43.

Xu, X., Liu, J., Li, H., Jiang, M., 2016. "Capacity-oriented passenger flow control under uncertain demands: algorithm development and real-world case study", *Transportation Research Part E: Logistics and Transportation Review*, vol. 87, pp. 130-148

Wang, L., Yan, X., Wang, Y., 2015. "Modeling and Optimization of Collaborative Passenger Control in Urban Rail Stations under Mass Passenger Flow," *Mathematical Problems in Engineering*, vol. 2015, pp. 8.

Wang, Y., De Schutter, B., van den Boom, T. J. J., Ning, B., Tang, T., 2014. "Efficient bi-level approach for urban rail transit operation with stop-skipping", *IEEE Transactions on Intelligent Transportation Systems*, vol. 15, pp. 2658 – 2670.

Passenger Flow Prediction of High Speed Railway Based on LSTM Deep Neural Network

Jie Li [a,b], Ping Huang [a], Yuxiang Yang [a] ,Qiyuan Peng[a,1]
[a] School of Transportation and Logistics, Southwest Jiaotong University
High-tech Zone West Campus, 611756, Chengdu, China
[b] Civil, Buildings and Environmental Engineering Department, SAPIENZA Università di Roma
Via Eudossiana 18, 00184 Roma, Italy
[1] E-mail: qiyuan-peng@swjtu.cn, Phone: +86 13808061287

Abstract
The paper presents the characteristics of the departing passenger flow in different stations based on the real-record passenger flow data of Wuhan-Guangzhou high speed railway, from January, 2010 to December, 2015. The passenger dataset is framed for the long short-term memory (LSTM) model, considering the expectation input format of LSTM layers and the characteristics of the data. The Keras model in Python is used to fit LSTM model with tuning and regulating all the parameters necessary in the model. Then the fitted LSTM model is applied to forecast the short-term departing passenger flow of Wuhan-Guangzhou high speed railway. The influence of important parameters in the LSTM model on the prediction accuracy is analysed, and the comparison with other representative passenger flow forecast models is conducted. The results show that the LSTM model can get the valid information in a long passenger flow time series and achieve a better performance than other models. The passenger flow prediction errors valued by MAPE are 7.36%, 7.33%, 8.03%, respectively for Chenzhou station, Hengyang station and Shaoguan station. The parameters in the LSTM model such as the number of hidden units, the batch size and the input historical data length have a great influence on the prediction accuracy.

Keywords
High speed railway, Passenger flow prediction, Long short-term memory model, Deep learning, Time series

1 Introduction

The HSR has developed significantly in China due to its efficiency in transporting large numbers of passengers within short travel times. The short-term forecasting of high-speed rail passenger flow is one of the most critical issues because passenger demand provides a reference for seat allocation, ticket booking and train routing. The daily-based passenger demand in the near future is essential for the railway revenue management.

Short-term passenger flow prediction has a long history, and many successful models have been developed for this issue. These models can be generally divided into three categories: parametric approaches, nonparametric approaches and hybrid models.

In general, parametric approaches are model-based methods, whose structure is predetermined based on certain theoretical assumptions and the model parameters can be computed with empirical data. A variety of parametric models have been applied on traffic forecasting, such as the grey forecasting model, exponential smoothing model (Kyungdo (1995) and Tan (2009)), Kalman filtering models((Chen (2001), Chien (2003),Wang(2006) and Van (2008)), state space model (Liu (2006)) and so on. The most

widely used parametric method is Autoregressive Moving Average (ARIMA) model, which assumes the traffic condition is a stationary process. ARIMA performs well and is effectively in modelling linear and stationary time series. A number of ARIMA based time series models have been proposed for traffic prediction (Moreira (2013), Williams (2003), Smith (2002), Williams (2001) and Chandra (2009)).However, the parametric approaches cannot work well on stochastic and nonlinear data, thus the nonparametric methods are developed to forecast the traffic flow with stochastic and nonlinear.

In the nonparametric approaches, the parameters and the structure of the nonparametric approaches are uncertain. The non-parametric models used for traffic forecasting include support vector regression (Wu (2004), Zhang (2009), Asif (2014) and Zhang (2007)), neural networks (Çetiner (2010) and Tsai (2009)), Kalman filtering(Van Lint (2008) and Wang (2007)), Gaussian maximum likelihood (Tang (2003)) and so on. SVM is an artificial intelligence method based on the structural risk minimization principle and has the potential to overcome the problems of nonlinearity, small samples, high dimension, local minima and over-fitting. Neural networks are capable of handling multi-dimensional data with flexible model structure, strong learning ability as well as adaptability. The Neural networks has been applied in many researches (Karlaftis (2011), and Ma (2015)).However the neural networks have drawbacks of the potential of over fitting, the requirement of large train samples and the cost of long training time.

Third, hybrid models have been proposed for a better performance in passenger performance. Zhang (2014) proposed a hybrid EEMD-GSVM model and applied the model to forecast the short-term passenger flows of three typical origin–destination pairs in terms of travel distances. Wei (2012) forecasted metro passenger flows with a hybrid of EMD and neural networks that generated higher forecasting accuracy and stability than the seasonal ARIMA. Zhu (2007) presented a hybrid method based on EMD and SVM for short-term electronic load forecasting. Li (2014) proposed an ensemble learning framework to appropriately combine estimation results from multiple macroscopic traffic flow models. Khashei (2012) proposed a new hybrid model of the autoregressive integrated moving average (ARIMA) and probabilistic neural network (PNN) to yield more accurate results than traditional ARIMA models.

Although numerous passenger flows forecast models have been developed, the short-term forecast of HSR passenger flow is still challenging because daily passenger flows are highly oscillated, nonlinear and non-stationary. In addition, most HSR lines in China are still under development, while passenger flows of opened HSR lines can be influenced by unstable demands such as holidays.

Currently, deep learning has been successfully applied in many fields and achieved reasonable results (Srivastava (2015), Donahue (2017) and Polson (2017)). Ma (2017) proposed a deep convolutional neural network for large-scale traffic network speed prediction. Yu (2017) designed a spatiotemporal recurrent convolutional network for predicting network-wide traffic speeds. Meanwhile, big data has revolutionized the transportation industry over the past several years. These two hot topics have inspired us to reconsider the traditional issue of passenger flow prediction. In this paper, a HSR passenger flow forecasting model based on LSTM is proposed.

The passenger flow sequence of HSR is nonlinear time series. The interaction in the passenger time series should be considered to forecast the short-term passenger flow. Most of the current passenger flow prediction model cannot take advantage of the effective information in the passenger flow time series. LSTM is one kind of deep neural network and the model is fitted based on the big data of passenger flow. LSTM can capture the nonlinearity and randomness of traffic flow more effectively, as well as

overcome the issue of back-propagated error decay through memory blocks, and thus shows superior capability for time series prediction with long temporal dependency. In this paper, daily ticket data on the Beijing- Guangzhou HSR was collected from January, 2010 to December, 2015. The proposed LSTM passenger forecast model is applied to forecast the passenger flow of Beijing- Guangzhou HSR.

The remainder of this paper is organized as follows: Section 1 provides a general overview of the existing approaches of traffic flow forecasting and the application. The long short-term memory neural network architecture is present and the passenger prediction model based on LSTM is introduced in section 2. In addition, the performance of the LSTM is evaluated, compared to other models such as Support Vector Machine (SVM), K-Nearest Neighbor (KNN) and Random Forest (RF). In section 4 the effect of parameters in the LSTM on the prediction performance is analysed. Finally, conclusion and future envisions are discussed in section 5.

2 Passenger Flow Prediction model Based on LSTM

2.1 Structure of the Memory Unit of LSTM

Recurrent neural network (RNN) is a powerful deep neural network which can deal with sequence data using the internal memory. The architecture of RNN is illustrated in Figure 1. RNN contains input layer X , hidden layer S and output layer O . U , V , W are weight vectors. At the time t the hidden layer S_t and the output O_t can be calculated as Equation (1) and Equation (2).

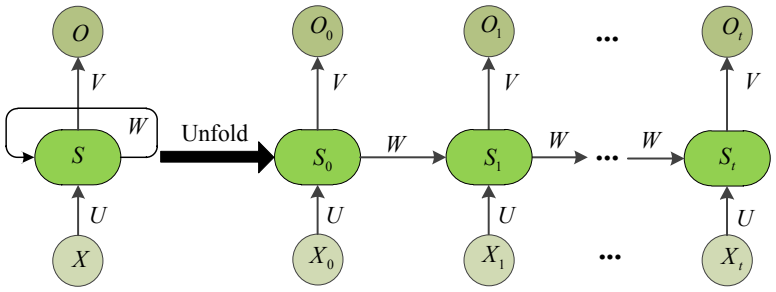


Figure 1: Standard RNN architecture

$$S_t = f\left(UX_t + WX_{t-1}\right) \tag{1}$$

$$O_t = g\left(VS_t\right) \tag{2}$$

Thus the hidden vector S_t at time t is determined by the input vector at time t and hidden vector at the previous time $t-1$ while the output O_t is determined by the historic input $X_t, X_{t-1}, X_{t-2}...X_1$.

In principle, RNN can map the whole historical input data to each output, relying on the key point that the recurrent connections allow the memory of previous input to affect the network's output. However, in standard RNN architecture, the given weight vector in the hidden layer plays an important role on the network output, which can lead to either decays or blows up exponentially as it cycles around the recurrent connections in the networks for too many times. This effect is often considered as the vanishing gradient problem. Thus, RNN is incapable of learning from long time lags, or saying long-term

dependencies (Bengio (2002)).

To address the problem, a LSTM is proposed to work well on modelling long-term time series. The difference between standard LSTM architecture and the RNN architecture is the hidden layer, which enhance the LSTM to avoid vanishing gradient problem. LSTM is a special kind of RNN. By treating the hidden layer as a memory unit, LSTM network can get the valid information in a long passenger flow time series. The typical architecture of LSTM memory unit is shown in Figure 2.

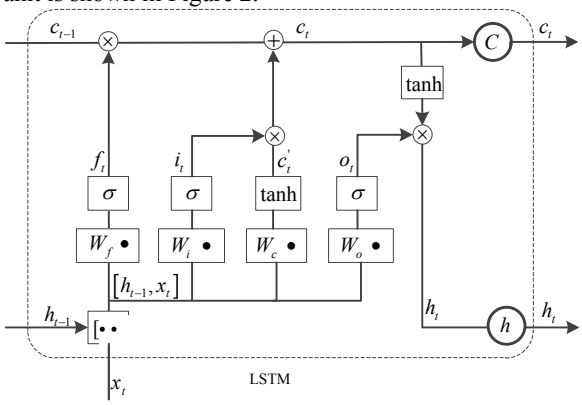


Figure 2: The structure of LSTM unit cell

There is a memory cell in the unit, denoted by C . Moreover, the LSTM memory unit contains three gates, namely input gate i_t , forget gate f_t and output gate o_t .The state of the memory cell at time t is indicated by c_t ,the input of every gate contains the preprocessed data x_t and the previous output of the LSTM unit, called h_{t-1} .Based on the information flow in the structure of memory unit, the update of the memory cell' state can be summarized as Equation (3) to Equation (8).

$$f_t = \sigma\left(W_f \bullet [h_{t-1}, x_t] + b_f\right) \tag{3}$$

$$i_t = \sigma\left(W_i \bullet [h_{t-1}, x_t] + b_i\right) \tag{4}$$

$$c_t^{'} = \tanh\left(W_c \bullet [h_{t-1}, x_t] + b_c\right) \tag{5}$$

$$c_t = f_t \circ c_{t\text{-}1} + i_t \circ c_t' \tag{6}$$

$$o_t = \sigma\left(W_o \bullet [h_{t-1}, x_t] + b_o\right) \tag{7}$$

$$h_t = o_t \circ \tanh\left(c_t\right) \tag{8}$$

i_t , f_t and o_t are the output of different gates, c_t is the new state of memory cell and h_t is the final output of the LSTM unit. W_f , W_i , W_c , W_o are coefficient matrixes, b_f , b_i , b_c , b_o represent the offset vectors, σ is the weight of the sigmoid function, and *tanh* is the hyperbolic tangent activation function. Via the function of the different gates, LSTM memory units can capture the complex correlation features within time series in both short and long term, which is a remarkable improvement compared with RNN.

2.2 LSTM Network for Passenger Flow Prediction

In the proposed LSTM model, the daily historical passenger flow data at one station can be viewed as a time sequence $P=\{p_1,p_2,...,p_i,...,p_n\}$. In general the passenger flow is non-stationary time series with increase and periodicity trend. The non-stationary of the time series affects the prediction accuracy of the LSTM model. Stationary data is easier to model and will very likely result in more skillful forecasts. A standard way to remove a trend is by differencing the data. That is the observation from the previous time step (t-1) is subtracted from the current observation (t). This removes the trend and we are left with difference series or the changes to the observations from one time step to the next.

Thus the passenger dataset should be framed for the long short-term memory (LSTM) model, considering the expectation input format of LSTM layers and the characteristics of the data.

The passenger flow $P=\{p_1,p_2,...,p_i,...,p_n\}$ can be transformed from time series to stationary by two steps, one is rolling window smoothing with M order and the other step is the differencing process. The rolling window smoothing process can remove the periodicity in the time series, after which the passenger data can be denoted as $P^W=\{p_1^w,p_2^w,...,p_{n-M+1}^w\}$.The differencing process can remove the increase trend in the time series, after which the passenger data can be denoted as $P^D=\{p_1^d,p_2^d,...,p_{n-M}^d\}$. Below are functions calculating the rolling window smoothing and differenced series.

$$p_t^w=(p_t+p_{t+1}+...+p_{t+M-1})/\mathrm{M}=\frac{1}{M}\sum_{i=t}^{t+M-1}p_i \tag{9}$$

$$p_t^d=p_{t+1}^w-p_t^w \tag{10}$$

Like other neural networks, LSTM expect data to be within the scale of the activation function used by the network. The default activation function for LSTMs is the hyperbolic tangent (*tanh*), which outputs values between -1 and 1. This is the preferred range for the time series data.

We can transform the dataset to the range [-1, 1] using the MinMaxScaler class. The function below inverts this operation. Again, we must invert the scale on forecasts to return the values back to the original scale so that the results can be interpreted and a comparable error score can be calculated.

$$x_{n,t+1}=\frac{p_n^d-\min(P_n^D)}{\max(P_n^D)-\min(P_n^D)} \tag{11}$$

The LSTM model in Keras assumes that your data is divided into input (X) and output (Y) components. Suppose we need to predict the passenger flow $P_{out}=\{p_{t+1},p_{t+2},...p_{t+n}\}$ of time duration $T=\{t+1,t+2,...t+n\}$ using the of m historical time steps passenger flow $P_{in}=\{p_{t-m+1},p_{t-m+2},...p_t\}$, we can concatenate these two series together to create data frame $\{X_{in},Y_{out}\}$ for supervised learning. Let us denote the input of LSTM model $X_{in}=\{X_1,X_2,...X_j...X_{n-M-L-F+1}\}$, $X_j=\{x_j,x_{j+1},...x_{j+L-1}\}$, $x_j\in X$, $|X_j|=L$ .The output of LSTM model $Y_{out}=\{Y_1,Y_2,...Y_j...Y_{n-M-L-F+1}\}$, $Y_j=\{x_{j+L},x_{j+L+1},...x_{j+L-1+F}\}$,

$x_{j+L} \in X$ and $|Y_j| = F$ .The $Dataframe = \{X_{in}, Y_{out}\}$ Should be divided into training datasets $Datatrain = (X_{Train}, Y_{Train})$ and test dataset $Datatest = (X_{Test}, Y_{Test})$. The training dataset are used to fit the model and the test dataset is used to evaluate the performance of the fitted model.

Given that the training dataset is defined as X inputs and Y outputs. Let us denote the input passenger time series as $X = \{x_1, x_2, ..., x_m\}$, hidden state of memory cells as $H = \{h_1, h_2, ..., h_m\}$ and output passenger prediction time series as $Y = \{y_1, y_2, ..., y_m\}$, LSTM works the computation as Equation (12) to Equation (13).

$$h_t = H(W_{hx} x_t + W_{hh} h_{t-1} + b_h) \tag{12}$$

$$p_t = W_{hy} y_{t-1} + b_y \tag{13}$$

The objective of the passenger flow prediction is to minimize the difference between the actual passenger flow and the predicted passenger flow. The square loss function given by the following formula is used as the objective function, in which y_t represents the actual passenger flow and p_t represents the predicted passenger flow.

$$e = \sum_{t=1}^{n} (y_t - p_t)^2 \tag{14}$$

In order to minimize training error and meanwhile avoid local minimal points, Adam optimizer, a modification of stochastic gradient descent (SGD) optimizer with adaptive learning rates, is applied for back propagation through time (BPTT).

The prediction accuracy of short-term traffic flow can be assessed by two commonly used metrics, i.e., Mean Absolute Percentage Error (MAPE) which evaluates the relative error and Root Mean Square Error (RMSE) which evaluates the absolute error. They are defined by Equation (15) and Equation (16).

$$MAPE = \frac{1}{n} \sum_{i=1}^{n} \frac{|p_{Prediction,i} - p_{Test,i}|}{p_{Test,i}} \times 100\% \tag{15}$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (p_{Test,i} - p_{Prediction,i})^2} \tag{16}$$

The flowchart of short-term passenger prediction based on LSTM is shown as Figure 3.

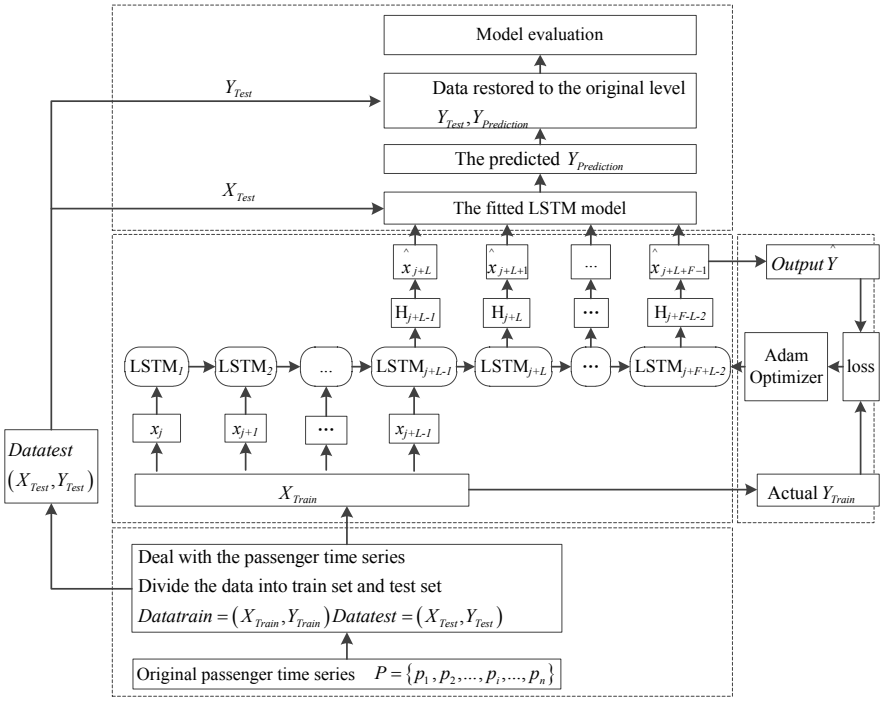


Figure 3: The flowchart of short-term passenger prediction model based on LSTM

3 Experiments and Results

3.1 Dataset Description

The passenger flow departing from Chenzhou station, Hengyang station and Shaoguan station, which locates on Wuhan-Guangzhou HSR, are taken as examples to demonstrate the efficiency of the LSTM based passenger prediction model.

The passenger volume data are collected every day from the booking tickets system, from 1st January, 2010 to 30th December, 2015, 2174 days in total. Part of the original dataset is shown in Figure 4. There is a big difference among the number of passengers departing form the three stations. Thus the performance of the LSTM Model on different grand passenger volume can be evaluated.

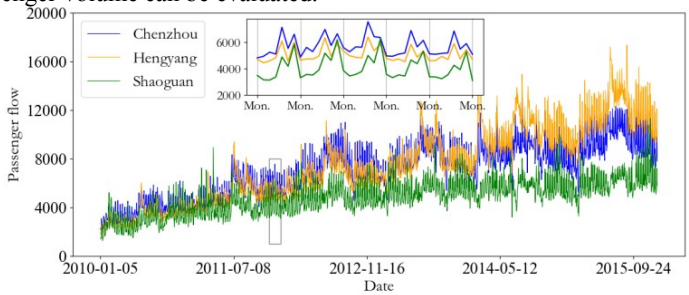


Figure 4: The passenger flow time series of the stations

The passenger volume increased during vacation, which may affect the prediction precision of LSTM model. Thus the passenger data of vacation are removed from the time series, and then 1673 days left.

The passenger time series of each station are shown in Figure 4. The passengers series present an increase trend as well as a significant cyclical with a period of 7 days. The passenger peak days appears on Friday and Sunday while the passenger trough appears on Sunday. The passenger data should be transformed from time series to LSTM data, just follow the data process in part 2. The prepared data for LSTM is shown in Figure 5.

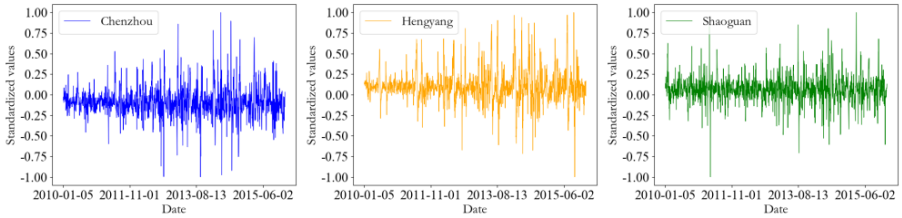


Figure 5: The standardized passenger flow time series

The passenger flow of Chenzhou station, Hengyang station and Shaoguan station are taken as examples to demonstrate the efficiency of the LSTM based passenger prediction model. Since the passenger volume presents a significant periodicity of 7 days, the objection of the model is to predict the passenger volume in the following days by means of the data of the previous seven days. To demonstrate the efficiency of the LSTM model as well as simplify the passenger prediction problem, LSTM model is just applied to forecast the passenger volume in the next day.

The passenger data is divided into two parts, the first 80% of the data is used to train the model, and the last 20% of the data is used to test the prediction accuracy of the model. To validate the efficiency of the proposed LSTM network, the performance is compared with some conventional forecast models, include ARIMA, SVM, RF, KNN.

Some key parameters should be determined for the short-term passenger flow prediction based on LSTM-RNN, including the size of input layer, the number of hidden layers, and the number of hidden units in each of hidden layer, the batch size and the size of output layer. The input historical data length is equal to the size of input layer, which is defined as 7 in the experiment. The number of hidden layers is assigned as 1,2,4,6,8 and the number of units in each hidden layer is assigned as 5,10,20,50,75,100. The size of output layer is 1, indicating the passenger flow in the next step will be forecasted. Grid search method is used to obtain the optimal parameters. The candidate values of the parameters are shown in Table 1. We performed a grid search over this parameter in order to find the size that leads to the best results. The grid search results of the optimal model parameters and the prediction precision are listed in Table 2.

Table 1: The parameters and hyper parameters of LSTM model

parameters	**Values**
Input historical data length	7
Output size	1
Epoch	1000
Optimizer	Adam
Learning rate	0.0001
Dropout	0.3
Loss function	Mean_Squared_Error
Activation function	Tanh
hyper parameters	**Values**
Batch size	1,2,4,6,8,10,12,14,16
Hidden Unit	5,10,20,50,75,100
Architecture	Input layer→LSTM layer→LSTM layer→Dropout layer→ Fully connected layer→Output layer

Table 2: The optimal parameters of LSTM model for different stations

Station	**Batch size**	**Hidden unit**	**MAPE**	**RMSE**
Chenzhou	1	10	7.21%	759.582
Hengyang	1	10	7.28%	800.227
Shaoguan	1	10	7.79%	562.000

3.2 Prediction Performance Analysis

In this section, we use the same experimental setup and fit the model for 1000 training epochs. A line plot of the series of RMSE scores on the train and test sets after each training epoch is also created, which is shown in Figure 6. The result clearly shows a downward trend in RMSE over the training epochs for the experimental runs of the three stations. The lines for the all the train case shows a sharp decrease before 200 epochs and then become more horizontal, but still generally show a downward trend, although at a lower rate of change. The lines for the test case of Chenzhou station, Hengyang station and Shaoguan station show a downward trend respectively before 500 epochs, 50 epochs and 400 epochs, and then the lines become more horizontal.

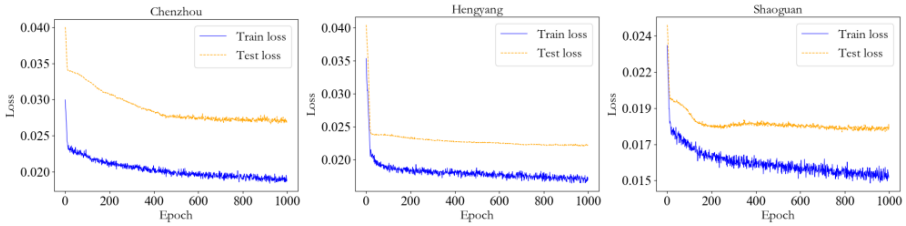


Figure 6: Diagnostic results with 1000 Epochs

Figure 7 shows the passenger flow comparison between observation values and the prediction values obtained by the LSTM. The prediction results are fairly good for Chenzhou, Hengyang and Shaoguan station, whose MAPE are 7.26%,7.33%,8.03% respectively.

In general the LSTM model is well capable of predicting the passenger volume trend.

The prediction accurate is high as the passenger flow shows a regular trend. However, when dramatic changes in the passenger flow are observed, the prediction accurate is low.

The distribution of MAPE over the predicted values is shown in Figure 7. We could find that most of the MAPE are located in （0,10%）. Specifically, 52.7%, 57.5%, 57.5% of the MAPE is less than 5%, and 81.9%, 79.2%, 82.4% of the MAPE is less than 10%, respectively for Chenzhou, Hengyang and Shaoguan station.

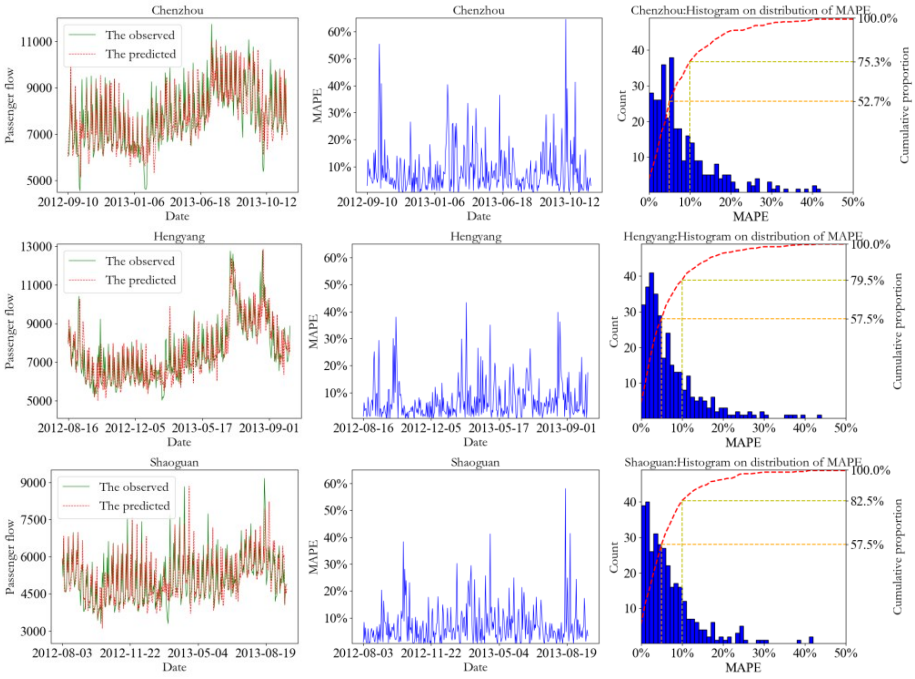


Figure 7: The prediction passenger flow and MAPE

To validate the efficiency of the proposed LSTM network, the performance is compared with some conventional forecast approaches; include ARIMA, SVM, RF, KNN. Each prediction method is tested for 10 times to avoid the randomness. The experimental results are shown in Table 3.As we can see from Table 3, compared to other methods, the MAPE of LSTM are the lowest. For the RMSE, the GB method performed best for the passenger volume prediction at Hengyang station while the RMSE of LSTM at other stations is the lowest.

Table 3: Prediction results of different models

Model	Chenzhou		Hengyang		Shaoguan	
	RMSE	**MAPE**	**RMSE**	**MAPE**	**RMSE**	**MAPE**
LSTM	**778.74**	**7.26%**	801.46	**7.33%**	**584.29**	**8.03%**
RF	861.45	8.14%	844.41	7.89%	628.60	8.96%
GB	831.37	7.80%	**801.15**	7.36%	602.41	8.44%
KNN	860. 74	8.11%	805.26	7.56%	625.07	8.51%
SVM	796.76	7.55%	816.51	7.34%	591.31	8.21%

For a further analysis of the prediction efficiency and the stability of different

prediction models, RMSE and MAPE distributions of each model is shown on a box and whisker plot in Figure 8.

The red line shows the median and the box shows the 25th and 75th percentiles, or the middle 50% of the MAPE. The values of the red line give an idea of the average expected performance of a configuration whereas the box gives an idea of the range of possible best and worst case examples that might be expected.

Looking at just the median RMSE scores, the results suggest that the choice of LSTM to predict the passenger volume is better than the other models since the median RMSE scores of LSTM for every station are the lowest and the average expected performance of LSTM is good. In terms of the stability of different prediction models, the comparison of the boxes suggest that the performance of the FR model is unstable since the gap between the 25th and 75th percentiles MAPE scores is large. The performance of LSTM model is relative stable while GB, KNN, SVM model is very stable.

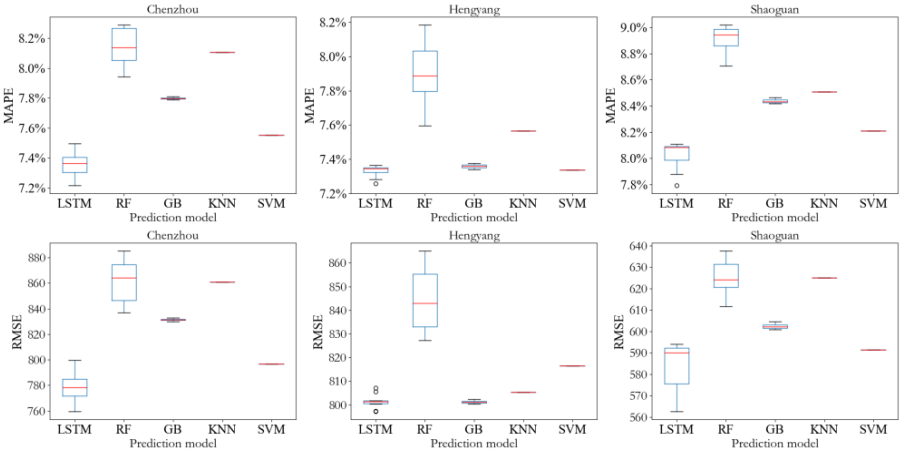


Figure 8: Boxplot of prediction RMSE and MAPE of different models

To sum up the above analysis, LSTM RNN model is capable of memorizing long historical data and achieving higher prediction accuracy even if the model is quite simple. Therefore, the proposed model is effective in short-term traffic flow prediction.

4 Analysis on the Influence of Model Parameters on Prediction Accuracy

4.1 The Number of Hidden Units

The number of hidden units in each of hidden layer affects the learning ability of the network. Generally, more neurons would be able to learn more structure from the problem at the cost of longer training time. More learning capacity also creates the problem of potentially over fitting the training data.

The effect of hidden units on the prediction results is investigated by assigning the number of units as 5, 10,20,50,75,100. We can objectively compare the impact of increasing the number of neurons while keeping all other network configurations fixed. We will use a batch size of 1 and 1000 training epochs

In order to alleviate the influence of random initialization for the model, we repeat each experiment 30 times and compare the average test RMSE performance with the

number of neurons, the result is shown in Figure 9.

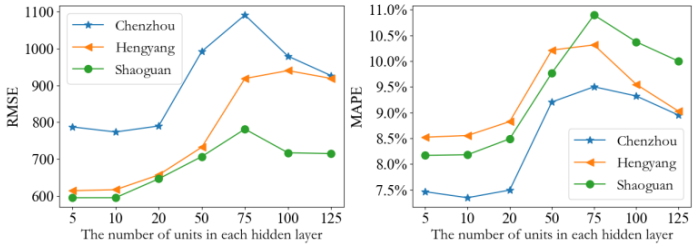


Figure 9: The distribution of prediction MAPE and RMSE of models with different hidden units

From Figure 9, we can see that the MAPE and RMSE remain stable when the number of the number of hidden units is less than 20, and the values of which are low. As the number of hidden units is more than 20 and less than 75, the MAPE and RMSE rise up with the increase of the number of hidden units. The MAEPs and MSEs of Hengyang station reach to the highest as the hidden units is 100, while the MAPE of the other stations decrease. As the number of hidden unit is 100, the LSTM model perhaps show an acceleration of over fitting.

Specially, diagnostic with1000 epochs and various neuron of Hengyang station are taken as an example to demonstrate the effect of the neurons on the LSTM. As the number of neurons is 5 and 10, both the line of train loss and test loss show horizontal. The results suggest a good, but not great, general performance. It shows a rapid decrease in test RMSE as the neurons is 10, which means the learning capacity of the network is improved as the number of the neurons increase from 5 to 10.

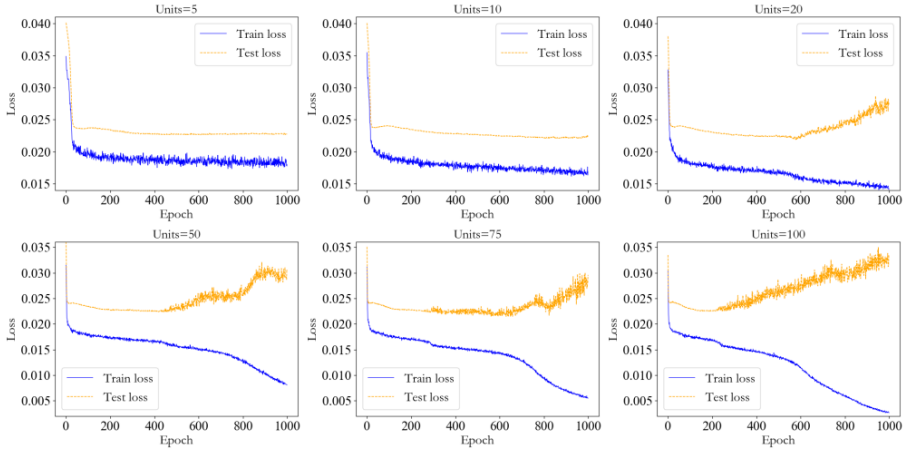


Figure 10: Diagnostic results of models with different hidden units

The diagnostic results of models with different hidden units are shown in Figure 10. As the number of the neurons is 20, 50, 75, diagnostic results shows a rapid decrease in test RMSE to about epoch 500-600. Meanwhile, the training dataset shows a continued decrease to the final epoch. These are significant signs of over fitting of the training dataset. When the number of the neurons is 100, the inflection point in the training dataset

seems to be happening sooner than the 20, 50, 75 neurons experiment, perhaps at epoch 300-400.

It can be proved that more neurons can enhance the learning ability of LSTM network. However, too many neurons may lead to an over fitting of the training dataset. These increases in the number of neurons may benefit from additional changes to slowing down the rate of learning, such as the use of regularization methods like dropout, decrease to the batch size, and decrease to the number of training epochs.

4.2 The Batch Size

Batch size is an important parameter in the LSTM configures, which limits the number of samples to be shown to the network before a weight update can be performed. Thus batch size controls how often to update the weights of the LSTM network. This same limitation is then imposed when making predictions with the fit model.

In this section, we will explore the effect of varying the batch size. In this study the batch size used are 1,2,3,4,5,6,7.We will hold the number of training epochs constant at 1000.As with training epochs, we can objectively compare the performance of the network given different batch sizes. Each configuration was run 10 times and summary statistics calculated on the final results.

A box and whisker plot of the prediction MAPE and MSE were created to help graphically compare the distributions, shown in Figure 11. The green line shows the average performance while the box shows the variability of the performance of the LSTM with different batch size.

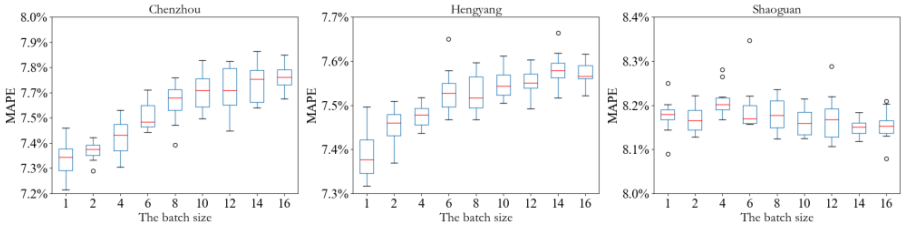


Figure 11: The distribution of prediction MAPE and RMSE of models with different batch size

In terms of the average performance, the median MAPE of Chenzhou station and Hengyang station showed an upward trend as the batch size increase from 1 to 16. The lowest low median MAPE are 7.36% and 7.39% as the batch size is 1, respectively for Chenzhou station and Hengyang station. For the Shaoguan station, the median MAPE fluctuates with the varying of the batch size, without obvious increase or decrease trend, indicating that the prediction accuracy of Shaoguan LSTM model for Shaoguan station is less affected by batch size.

The variability of the performance, the batch size has an influence on the stability of the LSTM model, since the variability of the box varies with the batch size. However the trend is not clear.

Tuning the batch size in a neural network is a tradeoff of average performance and variability of that performance. The ideal result should have a low mean error with low variability, meaning that it is generally good and reproducible. And the batch size should be decided according to the Data characteristics.

4.3 The Input Historical Data Length

The excellent performance of LSTM for short term traffic flow prediction mainly benefits from the memory ability of LSTM. For purpose of verifying the ability of LSTM to memorize long historical data, the performances of each model with different historical data length are compared. The input historical data length ranges from 7 to 35 with the interval of 7.Note that the input historical data length is always equal to the input size of each model. The five models' MAPE and RMSE are illustrated in Figure 12.

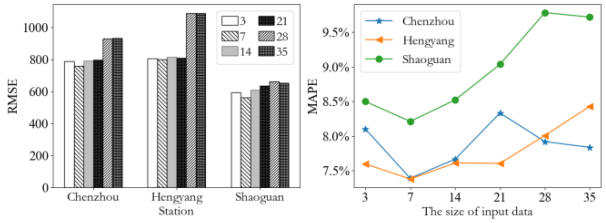


Figure 12: The distribution of prediction MAPE and RMSE of models with different input historical data length

There is a general trend of decreasing RMSE and MAPE as the number of historical data length increases from 3 to 7. As the historical data length increases from 3 to 7, the RMSE and MAPE for the passenger prediction at Chenzhou and Hengyang station rise up. For the passenger prediction at Shaoguan station, RMSE and MAPE increase as the historical data length increases from 7 to 21 and decrease as the historical data length increases from 21 to 35. The experiment results suggest a network configuration with historical data length of 7 having the best performance, the MAPE of which are 7.39%, 7.38%, 8.21%, Respectively for Chenzhou, Hengyang, Shaoguan station. It means that for one day prediction interval, the passenger flow in the past 7 days has a great impact on the current passenger flow, corresponding with the significant periodicity of 7 days presented by the passenger flow.

Specially, Diagnostic with1000 Epochs and various historical data length of Hengyang station are taken as an example to demonstrate the effect of the input of historical data length on the LSTM, and the result is shown in Figure 13.

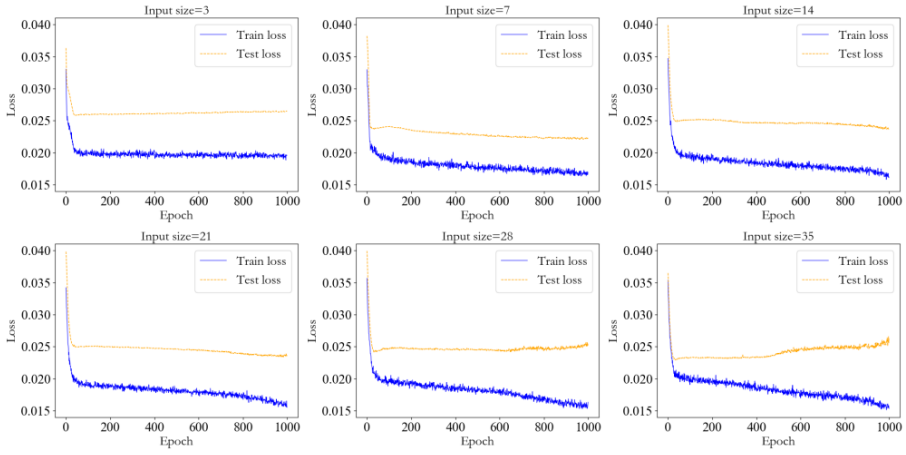


Figure 13: Diagnostic Results of models with different input historical data length

As the historical data length is 3, 7, 14, 21, both the lines of train loss and test loss show decrease trend and then keep horizontal, and the performance of these model seems to be reasonable. It shows a rapid decrease in test RMSE as the historical data length is 7 and the final RMSE score is 0.0222.

As the historical data length increases to 28 and 35, diagnostic results shows a rapid decrease in test RMSE to about epoch 500 and then rise up slightly until to the final epoch 1000. Meanwhile, the training dataset shows a continued decrease to the final epoch. There is a potential possibility of over fitting the training dataset.

To sum up, the performance of LSTM is effected by the input size of the data. LSTM can learning and memorize the complex interaction in the passenger time series and then predict the following passenger volume. For the passenger prediction in the experiment, the LSTM model is not well capable of getting the characteristics of the passenger time series as the input historical data length is too short. Meanwhile, as the input historical length is long, the limited learning ability cannot get the enough valid information contained in the data. In addition, and the longer the sequence data is, the more Interference noise information it contains, may lead a low prediction precision. Therefore, it is proper to model long-term dependencies and determine the optimal size of input data dynamically for the desirable results of short-term traffic flow prediction.

5 Conclusions

The paper analysis on the passenger flow characteristic of Wuhan-Guangzhou High Speed rail and proposes a passenger flow prediction method based on LSTM deep neural network. The results showed that:

(1) The LSTM passenger prediction model can cope with the correlation within long-term passenger time series and predict the trend of passenger flow accurately. The average prediction error MAPE of Chenzhou, Hengyang and Shaoguan stations are 7.36%, 7.33% and 8.03%, respectively. LSTM model is more effective and reliable than the other models, including RF, SVM, KNN, and GB models, while the stability of LSTM model is poor.

(2) The number of hidden units in each of hidden layer has a great influence on the prediction accuracy. While the number of hidden units is low, a slight increase of the hidden units of LSTM model can improve the convergence speed and prediction accuracy. As the number of hidden units in the LSTM model increase to a high level, the LSTM model may show an over-fitting state. In the experiment, the LSTM work show a better performance as the number of hidden units is set as 5 or 10.

(3) The input historical data length and the batch size have a great influence on the prediction accuracy of the LSTM model. When the historical data length is 7 and the batch size is 1, the passenger prediction accuracy is higher, which means the passenger flow in the past 7 days has a great influence on the following passenger flow.

Acknowledgement

This work was supported by the National Nature Science Foundation of China [grant number 71871188 and U1834209] and National Key R&D Program of China [grant number 2017YFB1200700]. We acknowledge the support of SAPIENZA Università di Roma and the China Scholarship Council. We are grateful for the contributions made by our project partners.

References

Asif, M.T., Dauwels, J., Goh, C.Y., et al., 2014. "Spatiotemporal patterns in large-scale traffic speed prediction". *IEEE Transactions on Intelligent Transportation Systems,* vol.15(2),pp.794–804.

Bengio, Y., Simard, P., Frasconi, P., 2002. "Learning long-term dependencies with gradient descent is difficult". *IEEE Transactions on Neural Networks,* vol.5(2),pp.157-166.

Çetiner, B.G., Sari, M., Borat, O., 2010. "A neural network based traffic flow prediction model". *Mathematical and Computational Applications*. vol.15 (2), pp.269-278.

Chandra, S.R., Al-Deek, H., 2009. "Prediction of freeway traffic speeds and volumes using vector autoregressive models". *Journal of Intelligent Transportation Systems*,vol.13 (2), pp.53-72.

Chen, H., Grant-Muller, S., 2001. "Use of sequential learning for short-term traffic flow forecasting". *Transportation Research Part C: Emerging Technologies,* vol.9(5), pp.319-336.

Chien, S.I.J., Liu, X., Ozbay, K., 2003. "Predicting travel times for the South Jersey real-time motorist information system". *Transportation Research Record,* vol.1855(1), pp.32-40.

Donahue, J., Hendricks, L A., Rohrbach M., et al. 2017. "Long-Term Recurrent Convolutional Networks for Visual Recognition and Description". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.39, pp.677-691.

Jiang, X., Zhang, L., and Chen, X.,2014. "Short-term forecasting of high-speed rail demand: a hybrid approach combining ensemble empirical mode decomposition and gray support vector machine with real-world applications in china". *Transportation Research Part C: Emerging Technologies,* vol. 44, pp.110-127.

Karlaftis, M., Vlahogianni, E., 2011. "Statistical methods versus neural networks in transportation research: differences, similarities and some insights". *Transportation Research Part C: Emerging Technologies,* vol. 19 (3),pp. 387–399.

Khashei, M , Bijari, M , Ardali, G A R ,2012. "Hybridization of autoregressive integrated moving average (ARIMA) with probabilistic neural networks (PNNs)". *Computers and Industrial Engineering,* vol.63(1),pp.37-45.

Kyungdoo, N., Thomas, S., 1995. "Forecasting international airline passenger traffic using neural networks". *Logist and Transportation Review,* vol.31(3),pp.239–251.

Li, L., Chen, X., Zhang, L., 2014. "Multimodel ensemble for freeway traffic state estimations". *IEEE Transactions on Intelligent Transportation Systems,*vol.15(3),pp.1323-1336.

Liu, H., Zuylen, H.V., Lint, H.V et al., 2006. "Predicting urban arterial travel time with state-space neural networks and Kalman filters". *Transportation Research Record,* vol.1968(1), pp.99-108.

Ma, X., Dai, Z., He, Z., et al., 2017. "Learning traffic as images: a deep convolutional neural network for large-scale transportation network speed prediction". *Sensors* vol.17(4),pp. 818.

Ma, X., Tao, Z., Wang, Y., et al. 2015. "Long short-term memory neural network for traffic speed prediction using remote microwave sensor data". *Research Part C: Emerging Technologies*, vol.54,pp. 187-197.

Moreira-Matias, L., Gama, J., Ferreira, M., et al.,2013. "Predicting taxi–passenger demand using streaming data". *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol.14, pp.1393–1402.

Polson, N G., Sokolov, V O., 2017.“Deep learning for short-term traffic flow prediction”. *Transportation Research Part C: Emerging Technologies,* vol.79,pp.1-17.

Smith, B.L., Williams, B.M., Oswald, R.K., 2002. “Comparison of parametric and nonparametric models for traffic flow forecasting”. *Transportation Research Part C: Emerging Technologies,* vol.10(4),pp.303-321.

Srivastava, N., Mansimov, E., Salakhutdinov, R., 2015, “Unsupervised learning of video representations using LSTMs”,In: *Proceedings of The International Conference on Machine Learning(ICML 2015),* Lille, France

Tan, M. , Wong, S. , Xu, J , et al, 2009. “An aggregation approach to short-term traffic flow prediction”. IEEE Transactions on Intelligent Transportation Systems, vol.10,pp.60-69..

Tang, Y.F., Lam, W.H.K., Ng, P.L.P., 2003., “Comparison of four modeling techniques for short-term AADT forecasting in Hong Kong”. *Journal of Transportation Engineering,* vol.129 (3),pp.271–277.

Tsai, T H., Lee, C K., Wei, C H., 2009, “Neural network based temporal feature models for short-term railway passenger demand forecasting”. *Expert Systems with Applications,* vol.36(2), pp.3728-3736.

Van Lint, J.W.C., 2008. “Online learning solutions for freeway travel time prediction”. *IEEE Transactions on Intelligent Transportation Systems,* vol.9 (1), pp.38–47.

Wang, Y., Papageorgiou, M., Messmer, A., 2007. “Real-time freeway traffic state estimation based on extended Kalman filter: a case study”. *Transportation Science,*vol.41(2), pp.167–181.

Wang,Y.,Papageorgiou, M., Messmer, A., 2006. “RENASSANCE: a unified macroscopic model-based approach to real-time freeway network traffic surveillance”. *Transportation Research Part C: Emerging Technologies,* vol. 14(3),pp. 190–212.

Wei,Y., Chen, M C., 2012.“Forecasting the short-term metro passenger flow with empirical mode decomposition and neural networks”. *Transportation Research Part C: Emerging Technologies,* vol. 21(1),pp.148-162.

Williams, B.M., 2001. “Multivariate vehicular traffic flow prediction: evaluation of ARIMAX modelling”. *Transportation Research Record*,vol.1776, pp.194–200.

Williams, B.M., Hoel, L.A., 2003. “Modeling and forecasting vehicular traffic flow as a seasonal ARIMA process: a theoretical basis and empirical results”. *Journal of transportation engineering,* vol.129 (6), pp.664-672.

Wu, C.H., Ho, J.M., Lee, D.T., 2004. “Travel-time prediction with support vector machine regression”. *IEEE transactions on intelligent transportation systems*,vol. 5(4), pp.276-281.

Yu, H., Wu, Z., Wang, S., et al. 2017. “Spatiotemporal recurrent convolutional networks for traffic prediction in transportation networks”. *Sensors,*vol.17(7), pp.1501.

Zhang, Y., Liu, Y., 2009. “Traffic forecasting using least squares support vector machines”. Transportmetrica . vol.5 (3), pp.193–213

Zhang, Y., Xie, Y., 2007. “Forecasting of short-term freeway volume with v-support vector machines”. *Transportation Research Record,* vol.2024(1), pp.92-99.

Zhu, Z., Sun, Y., Li, H., 2007. “Hybrid of EMD and SVMs for short-term load forecasting”. In: *Proceedings of the IEEE International Conference on Control and Automation(ICCA 2007)*, Guangzhou, China.

Tactical Capacity Assessment of a High-speed Railway Corridor with High Heterogeneity

Yanan LI [a], Ruihua XU [a,1], Chen JI [a], Han WANG [a], Di WU [a]
[a] The Key Laboratory of Road and Traffic Engineering, Ministry of Education
College of Transportation Engineering, Tongji University, China
[1] E-mail: rhxu@tongji.edu.cn, Phone: +86 13641650837

Abstract
Capacity assessment of high-speed railway corridor is critical in tactical planning process because it is beneficial to unearth the potential capacity and improve the capacity utilization without new investment in construction. China's high-speed railway corridor serves trains with high heterogeneity in different route, speed, and stopping plans. This paper first illustrates the necessity of assessing the corridor's capacity as a whole without decomposition. Based on the concept of base train equivalent (BTE), two methods named "capacity occupancy equivalent (COE)" method and "demand adaptation equivalent (DAE)" method are developed to standardize different types of trains into an equivalent unit. The case study of Jing-Hu high-speed railway corridor demonstrates that the methodology is concise in capacity assessment, and the impact of the long-distance direct service on corridor capacity utilization is also calculated.

Keywords
High-speed railway corridor, Capacity assessment, Base train equivalent, Heterogeneity, Demand adaptation

1 Introduction

The operation mileage of China's high-speed railway (CRH) is expected to be 30,000 km by 2020, forming a huge high-speed railway network. The high-speed railway corridor is the backbone of the network, providing local service by intra-line trains and long-distance direct service by cross-line trains (Figure 1). The origin and destination of the intra-line trains belong to the same corridor, while cross-line trains' origin and/or destination belong to the branch lines. The travel demands for different origins and destinations (OD) are extremely different over time and space dimensions along a long corridor. To meet with varieties of demand, trains run in different routes, different speed, and different stopping plans. Multiple types of trains running on the same corridor can cause different capacity impact and serious operational conflicts. Jing-Hu high-speed railway corridor, the busiest corridor in China's high-speed railway network, is facing the challenge of the increasing traveling demand. It is necessary to assess the corridor capacity and improve the capacity utilization.

Typically, the capacity of a rail corridor is defined as the number of trains that can safely pass within a period of time (Pouryousef, 2015). Considering the heterogeneity, Lai et al. (2012, 2015) use equivalent train unit to define capacity on lines. A few studies attempt to use "removal coefficient" to represent the impact of heterogeneous trains (Abramović B et al, 2004; Yang Z et al, 1995; Zhao, L.Z, 2001). However, most of the researchers divide the line or a corridor into sections as the first step of capacity assessment. The paradoxes of decomposition exist (Landex, A., 2008). When assessing the capacity of the corridor with

high heterogeneity, the shortcoming of the decomposition is more obvious. The number of the capacity by adding up the number of trains in different types directly makes the result incomparable. In the latest version of the UIC 406 method (2013), it recommends to look at entire routes without decomposition when assessing long-distance services. However, there is not an explicit method.

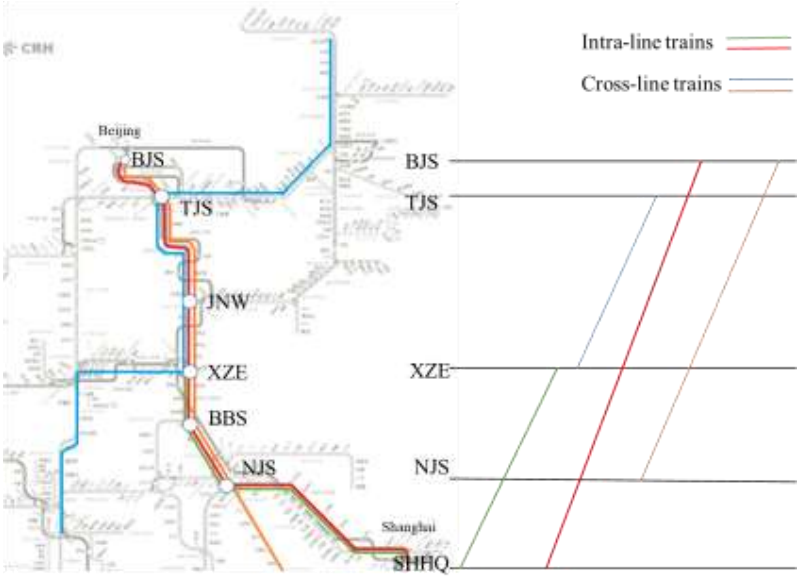


Figure 1 Intra-line trains and cross-line trains

This paper proposes a methodology for capacity assessment of high-speed railway corridor in the tactical level. We first illustrate that high-speed railway corridor with high heterogeneity should be regarded as a whole. Based on the concept of base train equivalent (BTE), "capacity occupancy equivalent (COE)" method and "demand adaptation equivalent (DAE)" method are developed to assess the corridor capacity by standardizing different types of trains into an equivalent unit. The case study of Jing-Hu high-speed railway corridor demonstrates that the methodology is concise in capacity assessment, and the cross-line trains' impact on corridor capacity utilization is calculated at the same time. In China, capacity assessment runs through strategic level (building of infrastructure), tactical level (timetabling), and operational level (short-term rescheduling and dispatching). Here we talk about the tactical capacity, which aims to unearth the potential capacity of the corridor under the current timetable. Therefore, this paper does not consider strategic level with little information of schedule (Jensen L W, 2017) or consider dynamic infrastructure occupancy under disturbances (Corman, 2010; C. Schmitz, 2017).

The remainder of the paper is organized as follows. In Section 2, we review relevant literature on capacity assessment methods. Section 3 explains the necessity of regarding the corridor as a whole in tactical capacity assessment. The methodology is then introduced in detail, and corresponding algorithm is developed and applied on Jing-Hu high-speed railway corridor as a case study in Section 4 and 5; and in the final section, conclusions and research extensions are also discussed.

2 Literature review

There exist various types of approaches for capacity assessment. Abril et al. (2008) classified the capacity methodologies as analytical methods, optimization methods, and simulation methods.

Analytical approach typically uses several steps of data processing through mathematical equations or algebraic expressions to determine theoretical capacity of the section/corridor at a planning level, as well as for the identification of bottlenecks in the infrastructure (Pouryousef, 2015; Riejos, 2016).

A widely used analytical method to assess capacity is defined by UIC (International Union of Railways) in their leaflet 406 (2013, 2004). The UIC406 leaflet describes how to assess the capacity consumed on a piece of infrastructure based on a given timetable using timetable compression. The definition of corridor in this leaflet is that corridors form the main structure of a railway network and are also considered to be a railway network's main source of revenue. In the UIC 406 method, the network is decomposed into sections for easier manageability. However, one of the shortcomings of this is that different network decompositions will lead to different results. Especially shorter line sections are a problem in the method (Landex, A., 2008). In the latest version of the UIC 406 method from 2013, it is recommended to look at entire routes without decomposition when assessing long-distance services. However, the latest version did not give a clear calculation method. In Germany, queueing based approach is common, focusing on knock-on delays (Wendler E, 2007; Weik N, 2016). In China, "removal coefficients of elimination" method (Abramović B et al, 2004; Yang Z et al, 1995; Zhao, L.Z, 2001; Xu, 2005) is used to represent the different impact of trains in different types. Base train equivalent (BTE) are proposed by Lai (2012) to identify the impact of heterogeneity on capacity. BTE models for headway-based analytical capacity analysis enables the standardization of rail capacity unit, facilitates assessment of the impact from heterogeneous trains, and allows comparison and evaluation of the capacity measurements from different lines and systems. The concept of equivalent is well known in many fields to deal with heterogeneity. Numerous of studies have been applied to use passenger car equivalent (PCE) for road capacity analysis (Elefteriadou L, 1997), pedestrian traffic (Galiza R J, 2012), etc.

Optimization methods are based on the design of saturated schedules and use mathematical programming models that achieve a high degree of saturation and simultaneously ensure certain level of quality of service (Abril et al, 2007). Majority of optimization methods are related to train timetabling problem (TTP).

Simulation methods attempt to replicate the actual operation of trains. Simulation tools are commonly used for detailed timetable analyses (Ralf Borndörfer, 2018). Several commercial software applications used as tools in the railway sector, including MultiRail, OpenTrack, Simone, RailSys, etc.

In conclusion, few researches concentrate on capacity assessment of high-speed railway corridor as a whole, and few papers pay attention to the demand adaptation when calculating the capacity utilization.

3 Corridor capacity assessment without decomposition

This section illustrates the necessity of regarding the corridor capacity as a whole. The weakness of corridor decomposition is explained from two aspects: the impact of cross-line trains and speed difference. To make a clear explanation, the necessary parameters are listed

as follows. The corresponding values are requested by signal system and safety technic norms from China Railway Corporation.

I Minimum interval between two trains in general, I=4min;
I_1 Minimum interval when the train arrives and the following train passes through the same station, I_1=3min;
I_2 Minimum interval when the train passes through and the following train departures from the same station,I_2=2min;
t_b Additional time when braking, t_b=3min (if the train runs at 350km/h) or t_b=2min (if the train runs at 300km/h)
t_s Additional time when starting, t_s=3min (either the train runs at 350km/h or 300km/h)
t_d Deviation time between the best position for capacity utilization and the actual position in the train diagram

3.1 Cross-line trains

With the expansion of China's high-speed railway network, the traveling demand is stimulated as CRH has greatly reduced travel time and it is more punctual, comfortable, convenient, and safe. Cross-line trains are operated, aiming to provide long-distance direct service.

The cross-line trains are almost fixed in the timetable for the corridor line. Once it is adjusted, the timetables of the branch lines must be adjusted at the same time. When cross-line trains run into the corridor or run away from the corridor, the deviation time (t_d) happens because of the cross-line trains' inflexibility. It is hard to make seamless connection for all the cross-line trains. If a corridor is decomposed when assessing the corridor capacity, the deviation time will be easily ignored as each decomposed cross-line train will be regarded as an independent intra-line train. According to the Jing-Hu corridor timetable in the tactical planning process (2018), the total deviation time of five important Jing-Hu nodes (NJS, BBS, XZE, JNW, TJS in Figure 1) is 109min in up-direction during one day's operation period. Take NJS station for example (cross-line trains run into corridor via NJS station), the total deviation time of cross-line trains is 34min from 6:00-23:00. If the corridor is decomposed into sections, this part of time will be regarded as unused capacity after compression. The capacity will be $\lfloor 34/I \rfloor$=8 trains more than the actual value according to the current method. Therefore, the idle time caused by the cross-trains' inflexibility should be considered.

3.2 Speed

Train speed is an important factor in capacity assessment. In Jing-Hu corridor, trains travel in 350km/h or 300km/h. The time difference in two kinds of trains for each section (Δt_i) is listed in Table 1. The high speed and the long distance increase the impact of the speed difference.

A fast train (350km/h) can at least overtake $\lfloor 27/I \rfloor$=6 slower trains (300km/h) according to the travel time differences (Table 1). The overlap leads to additional stops. Here, we define the slower train (300km/h) as standard train, denoted as $\varepsilon_{300km/h} = 1$, because the proportion of slower trains is more than 90% in China. Therefore, the capacity occupancy coefficient of a fast train $\varepsilon_{350km/h}$ consists of two parts: the basic capacity occupancy coefficient ε_b (Formula 1) caused by the additional stops, and the additional capacity

occupancy coefficient ε_d (Formula 2) caused by time deviation.

$$\varepsilon_b = \frac{(I_1 + t_b) + (I_2 + t_s)}{I}. \quad (1)$$

$$\varepsilon_d = \frac{t_d}{I} \quad (2)$$

The deviation time is between [0, I). In this paper, the granularity is 1min, and the initial feasible value of t_d is 0min, 1min, 2min or 3min. The algorithm of calculating t_d is as follows. The calculation results are listed in the right part of Table1, and the Figure 2 is the result when the initial t_d is 2min. Figure 2 and Table 1 are related, and the distance between the horizontal lines in Figure 2 reflects the distance between the two station.

Algorithm 1. Calculation of t_d

Data: t_d=[0,1,2,3], Δt_i (i=1, 2, ….n) (n: the number of sections)

for t_d=0 …3 **do**
  for i= n to 1 **do** $t_d^i = t_d^i + \Delta t_i$
    If $t_d^i - I > 0$
      $t_d^i = t_d^i - I$
    Else
      $t_d^i = t_d^i$
  end
  t_d=max$\{t_d^i\}$
end
return Average t_d

Therefore, $\varepsilon_{350km/h} = \frac{(3+2)+(2+3)}{4} + \frac{1.25}{4}$ =2.815, which means the capacity occupancy of a fast train equals to 2.815 slower trains. If the object is a section but not the corridor, the result must be smaller. The longer the distance is, the bigger the difference of the speed. For example, the capacity occupancy coefficient of the fast train is 2.625 for BBS-XZE by the same method. Therefore, considering the speed heterogeneity impact, the corridor should not be decomposed into sections when train runs along the whole corridor.

4 Methodology

In this section, “capacity occupancy equivalent (COE)” method is first developed to standardize different types of trains into an equivalent unit, aiming to assess the capacity of high-speed railway corridor as a whole. Then, “demand adaptation equivalent (DAE)” method is proposed. Traveling demand adaptation is taken into consideration, aiming to make the capacity utilization more efficient and profitable.

4.1 Capacity occupancy equivalent methods (COE)

There are multiple types of trains running along the corridor. As the capacity of high-speed railway corridor should be assessed as a whole, the key is to standardize different types of trains into an equivalent unit. The base train unit (BTU) in this problem is defined as the

whole journey intra-line train, traveling from the end of the corridor to the other end of the corridor. According to the capacity occupancy of different type of trains in time-space dimension, equivalent coefficient of non-standard intra-line trains and cross-line trains can be calculated. For a train running through the corridor, the more capacity occupancy, the less efficient capacity utilization. In other words, the equivalent coefficient calculated by COE method is less than 1, and the equivalent coefficient value is less if the non-standard trains take up more capacity but make less profit.

(1) Non-standard Intra-line trains

The traveling span (S) of non-standard intra-line trains is shorter than the length of corridor (L) (see the green train in Figure 1). Compared with the base train unit, non-standard intra-line trains occupy the corridor but only make profit in S distance if no other trains can occupy the rest of the corridor efficiently. The reasons that there are no "connecting" trains in the timetable are: 1) The space and time resources has been occupied by the neighbour trains, because of the speed difference, stopping plans, priorities, etc. (Figure 3a); 2) The demand of the unoccupied capacity is low, and there is no need to arrange more trains.

If no other trains can occupy the rest of the corridor resources, it means the capacity of the whole corridor is not fully used. The equivalent coefficient of non-standard intra-line train is denoted as θ_1. Therefore, the equivalent coefficient for the train i is $\theta_1^i = \mathrm{S}_i/L$ for non-standard intra-line trains with different traveling distances. Otherwise, the connection train and this non-standard intra-line train can be equivalent to a base train unit if the "connecting" time is less than the maximum of the dwelling time domain. The "connecting" time can be regarded as the station dwelling time of the base train unit although these two trains are not connected actually (Figure 3b).

(2) Cross-line trains

The equivalent of cross-line train is more complex than intra-line train because of its inflexibility explained in Section 3. The deviation time of cross-line trains is an inefficient capacity occupancy. It happens not only when cross-line trains running into the corridor (denoted as t_d^{in}), but also when cross-line trains running away from the corridor and no other trains can use the rest part of the corridor resources, denoted as $t_d^{connection}$.

As the deviation time may happen when the cross-line trains run into the corridor or run away from the corridor, cross-line trains can be divided into A type cross-line train (either origin or destination belongs to the other lines in Figure 1 orange train) and B type cross-line train (both of them belong to the other lines in Figure 1 blue train).

a）A-type

The equivalent coefficient of A-type cross-line train is denoted as θ_2. If the A type cross-line trains have connected trains, the equivalent coefficient $\theta_2' = 1 - (t_d^{in} + t_d^{connection}) * \frac{v}{60}/L$. If the A type cross-line trains don't have connected trains, $\theta_2'' = (S - t_d^{in} * \frac{v}{60})/L$ (Figure 3c). Therefore, if the proportion of connected A type trains is α and unconnected ones is (1-α), the final equivalent coefficient of A type is:

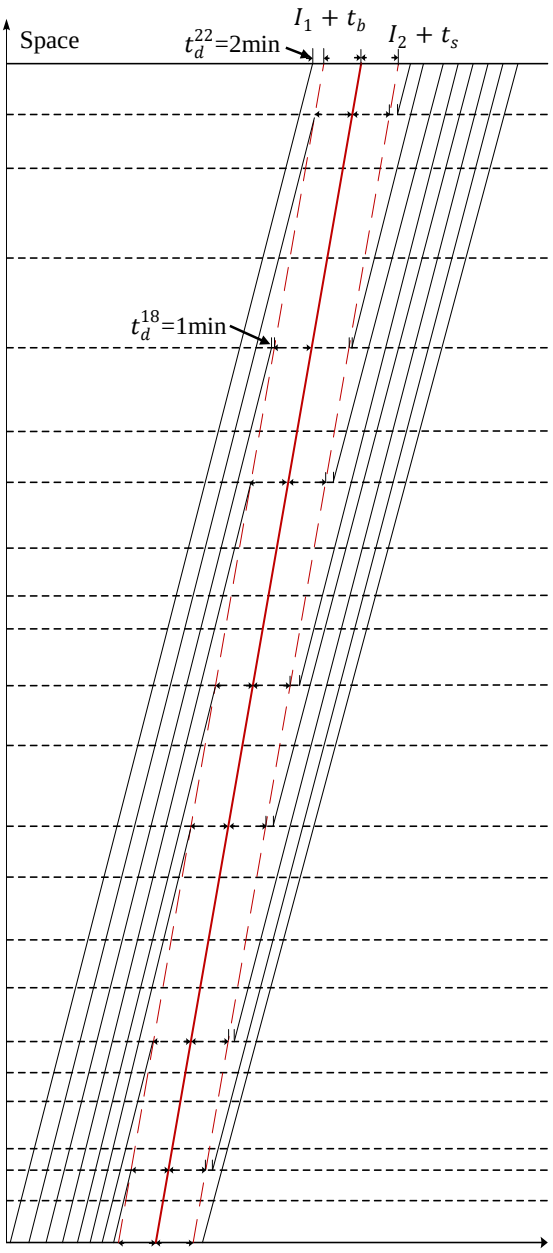


Figure 2 Example of Algorithm1

Name	i	Δt_i	t_d			
BJS	22	2	0	1	2	3
LF					0	1
TJS	21	1		0		
CZW	20	2	1			0
DZE	19	2		0	1	
JNW	18	2	1			0
TA	17	1			0	
QFE	16	1		0		
TZE	15	1	0			
ZZW	14	1				0
XZW	13	1			0	
SZE	12	2	0	1		
BBS	11	2			0	1
DY	10	1		0		
CZ	9	1	0			
NJS	8	1				0
ZJS	7	1			0	
DYN	6	1		0		
CZN	5	1	0			
WXE	4	1				0
SZN	3	1			0	
KSN	2	1		0		
SHHQ	1	3	2		0	1
maxt_d			2	1	1	1
average (max t_d)			1.25			

Table 1 Calculation of t_d

$$\theta_2 = \alpha \cdot \theta_2' + (1-\alpha)\sum_{k=1}^{k} \beta_k \cdot \theta_2'' \quad (3)$$

where β_k refers to the proportion of each kind of unconnected A type cross-line trains.

b）A-type

As there are two interfaces of B type cross-line trains, the equivalent coefficient should be divided into three parts. If it has two connected trains, $\theta_3' = 1 - 2 * (t_d^{in} + t_d^{connection}) * \frac{v}{60}/L$. If it has one connected trains, $\theta_3'' = S_{con}/L - (t_d^{in} + t_d^{connection}) * \frac{v}{60}/L$. S_{con} refers to the travel span of this kind of combination (Figure 3d). If B type train don't have connected trains, $\theta_2''' = (S - t_d^{in} * \frac{v}{60})/L$

Therefore, if the proportion of the above three kinds of B type trains is γ_1, γ_2 and γ_3, the final equivalent coefficient of B type is:

$$\theta_3 = \gamma_1 \cdot \theta_3' + \gamma_2 \cdot \sum_{m=1}^{m} \frac{\beta_m}{L}[S_{con}^m - (t_d^{in} + t_d^{connection}) * \frac{v}{60}] + \gamma_3 \cdot \sum_{n=1}^{n} \beta_n \cdot (S_n - t_d^{in} * \frac{v}{60})/L \quad (4)$$

where β_m, β_n refer to the proportion of corresponding trains.

Here, we define the proportion of intra-line trains and cross-line trains is φ and (1-φ), the proportion of A type cross-line trains and B type cross-line trains is ω and $(1-\omega)$. The equivalent coefficient based on "capacity occupancy equivalent" is:

$$\theta = \varphi \cdot \theta_1 + (1-\varphi) \cdot [\omega \cdot \theta_2 + (1-\omega) \cdot \theta_3] \quad (5)$$

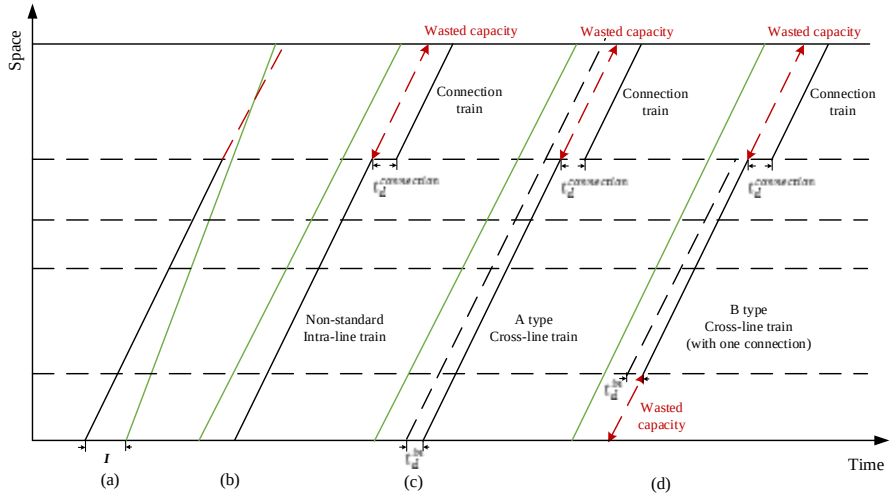


Figure 3 Illustration of parameters

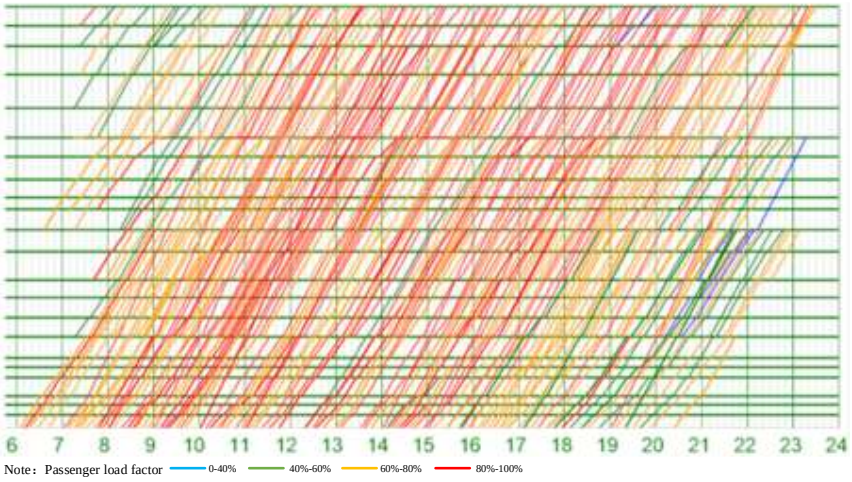


Figure 4 Capacity occupancy of speed difference

4.2 Demand adaptation equivalent method

The demand adaptation is directly related to the income. More trains don't equal to more passengers. The key of "demand adaptation equivalent (DAE)" method is adding the passenger load factor (σ) to the equivalent coefficient.

Figure 4 shows the average passenger load factor of each train running along the Jing-Hu corridor (in up direction) in 2017. The basic equivalent coefficient of a train is θ calculated in Section 4.1, then the demand adaptation equivalent coefficient can be $\theta' = \sigma \cdot \theta$.

5 Case study

Jing-Hu corridor is a typical high-speed railway corridor, connecting with metropolitan area and branch lines in the high-speed railway network. According to the train diagram (2018) in tactical planning process, there are 231 trains operating on Jing-Hu corridor in the up-direction (SHHQ-BJS) in one day.

The values of relative parameters can be statistically computed from the timetable as follows: $\varphi = 74\%, \omega = 70\%, \alpha = 22\%, \gamma_1 = 7.7\%,\ \gamma_2 = 38..5\%,\ \gamma_3 = 53.8\%$. $t_d^{in} \in (0,\ I)$, $t_d^{connection} \in (0, t_b + I_1 + I_2 + t_s)$. According to the central limit theory of great numeral, $t_d^{in} = 2$min and $t_d^{connection}$=5min as a simplification. The passenger load factor is clustered into four k partitions to simplify the demand analysis. The cluster centers and corresponding rates are 0.77 (40.6%), 0.90(25.0%), 0.65(24.6%), and 0.47(9.8%). Therefore, $\sigma = 0.77 * 40.6\% + 0.90 * 25.0\% + 0.65 * 24.6\% + 0.47 * 9.8\% = 0.744$. In this paper, we propose "DAE" method as a research direction, and give a simplified method for calculating the demand adaptation. The value of σ can be more specific according to each type of trains. The rest of the input data are listed in Table 2 and Table 3.

Table 2 Input data for A type cross-line trains

Section	Number of trains with connection	Number of trains without connection	β_k(%)	S_k (km)
SHH-NJS	9	6	6.5%	285.6
SHH-XZE	8	17	24.7%	616.1
SHH-JNW	1	6	6.5%	901.7
SHH-TJW	0	9	7.5%	1185.6
NJS-BJS	3	11	10.8%	1022.1
BBS-BJS	1	13	14.0%	846.8
ZXE-BJN	0	4	4.3%	691.6
JNW-BJS	10	24	25.8%	406

Table 3 Input data for B type cross-line trains

Connected Section	Train number	β_m	S_{con}^m	Section no connection	Train number	β_n	S_n
SHHQ-XZE	4	19.0%	616.1	NJS-XZE	7	25.0%	330.5
SHHQ-JNW	5	23.8%	901.7	NJN-JNW	2	7.1%	616.1
SHHQ-TJW	1	4.8%	1185.6	NJN-TJW	1	3.6%	900
NJS-BJS	7	33.3%	1022.1	BBS-XZE	6	21.4%	155.2
XZE-BJS	4	19.0%	691.6	BBS-JNW	1	3.6%	440.8
				BBN-TJW	1	3.6%	724.7
				XZE-JNW	7	25.0%	285.6
				XZE-TJW	1	3.6%	569.5
				JNW-TJW	2	7.1%	283.9

Based on "capacity occupancy equivalent (COE)" method, $\theta_2 = 0.521 + 0.433\alpha = 0.62$, $\theta_3 = 0.908\gamma_1 + 0.612\gamma_2 + 0.261\gamma_3 = 0.45$. In other words, an A type cross-line train is equivalent to 0.62 base train unit, and a B type cross-line train is equivalent to 0.45 base train unit. It is clear that is impact of B type is serious than A type cross-line trains because of $\theta_2 > \theta_3$. Among the 231 trains, there are 41 trains running from SHH to BJS, and non-standard intra-line trains are equivalent to 8.5. Therefore, the final equivalent capacity of Jing-Hu corridor is:

$$N = 231 \cdot [0.62\omega + 0.45(1-\omega)] \cdot \varphi + 41 + 8.5 = 147 \quad (6)$$

In other words, the capacity of the whole corridor is 147 base train units by "COE" method. In addition, the proportion of cross-line train(φ), the proportion of A and B (ω), and the connecting proportion of cross-line train (α, γ_i) are the key to the equivalents. To optimize the capacity utility of the main corridor, it is necessary control the proportion of cross-line trains and improve the level of coordination.

6 Conclusion

Most previous studies on railway capacity assessment is based on sections, and paid little attention to long-distance direct service by cross-line trains. This paper contributes a methodology for high-speed railway corridor's capacity assessment in tactical level. Based on the concept of base train equivalent (BTE), "capacity occupancy equivalent (COE)"

method and "demand adaptation equivalent (DAE)" method are developed to standardize different types of trains into an equivalent unit. The equivalent method makes the different timetable comparable, especially the corridor with high heterogeneity. This paper also proposes the demand oriented capacity assessment method, which will be more instructive in capacity utilization. Considering the serious impact of cross-line trains, the proportion should be controlled, and more efforts are necessary to achieve a compromise between accessibility of long-distance direct services and efficiency of the whole network.

Acknowledgements
The study was financially supported by China Railway Project (2017X009-M), and National Natural Science Foundation of China (71701152). The authors wish to acknowledge Shanghai Railway Co., Ltd, for collaboration during the research, and wish to thank anonymous referees for their useful comments.

References

Abril M, Barber F, Ingolotti L, et al. "An assessment of railway capacity", *Transportation Research Part E Logistics & Transportation Review*, Vol. 44(5), 2008, pp. 774-806.

Abramović B, Čičak M, Mlinarić T J. "Methods for determining throughput capacity of railway lines using coefficients of elimination", *Promet - Traffic - Traffico*, Vol. 16(2), 2004, pp. 63-69.

Corman, Francesco, et al, 2010. "A tabu search algorithm for rerouting trains during rail operations." *Transportation Research Part B: Methodological*, Vol. 44.1, pp: 175-192

Christoph Schmitz, Norman Weik, Stephan Zieger, Nils Nießen, Anke Schmeink, 2017. "Markov Models for the Performance Analysis of Railway Networks", In: *Proceedings of The 7th International Conference on Railway Operations Modelling and Analysis (RailLille2017)*, Lille, France.

Elefteriadou L, Torbic D, Webster N. "Development of Passenger Car Equivalent for Freeways, Two-Lane Highways, and Arterials", *Transportation Research Record Journal of the Transportation Research Board*, Vol. 1572(1), 1997, pp. 51-58.

Galiza R J, Ferreira L. "Developing Standard Pedestrian-Equivalent Factors Passenger Car–Equivalent Approach for Dealing with Pedestrian Diversity", *Transportation Research Record Journal of the Transportation Research Board*, Vol. 2299(2299), 2012, pp. 166–173.

Jensen L W, Landex A, Nielsen O A, et al., 2017. "Strategic assessment of capacity consumption in railway networks: Framework and model", *Transportation Research Part C*, vol. 74, pp. 126-149.

Lai Y C, Liu Y H, Lin T Y. "Development of Base Train Equivalent to Standardize Trains for Capacity Analysis", *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2289(2289), 2012, pp.119-125.

Lai Y C, Liu Y H, Lin Y J., 2015. "Standardization of capacity unit for headway-based rail capacity analysis", *Transportation Research Part C*, Vol. 57, pp. 68-84.

Landex, A., 2008. Methods to Estimate Railway Capacity and Passenger Delays (Ph.D. thesis). Lyngby, Denmark.

Pouryousef, Hamed, P. Lautala, T. White. "Railroad capacity tools and methodologies in the U.S. and Europe", *Journal of Modern Transportation*, Vol.23.1, 2015, pp.30-42.

Ralf Borndörfer, Torsten Klug, Leonardo Lamorgese, Carlo Mannino et al.. “Handbook of Optimization in the Railway Industry”, *International Series in Operations Research & Management Science*, Vol. 268, 2018, Springer, pp. 32.

Riejos F A O, Barrena E, Ortiz J D C, et al. “Analyzing the theoretical capacity of railway networks with a radial-backbone topology”, *Transportation Research Part A*, Vol. 84, 2016, pp. 83-92.

Ruihua Xu. Operational Organization of Railway Transportation System. China Railway Publishing House, 2005.

UIC, 2004. Capacity (UIC Code 406). first ed. pp. 1-21.

UIC, 2013. Capacity (UIC Code 406). second ed. pp. 1-51.

Weik N, Niebel N, Nießen N. “Capacity analysis of railway lines in Germany – A rigorous discussion of the queueing based approach”, *Journal of Rail Transport Planning & Management*, Vol. 6(2), 2016, pp. 99-115.

Wendler E. “The scheduled waiting time on railway lines”, *Transportation Research Part B*, Vol. 41(2), 2007, pp. 148-158.

Yang Z, Yang Y, SunQuanxin, et al. “A Study on Parameters for Calculation Block Section Carrying Capacity and Trains' Coefficient of Removal on Beijing-Shanghai High Speed Railway”, *Journal of Northern Jiaotong University*, Vol. S1, 1995, pp.1-8.

Zhao, L.Z.. “Calculation and analysis of carrying capacity of high-speed railway section”, *China Railway Science*, Vol. 22(6), 2001, pp. 54-58.

A Collection of Aspects Why Optimization Projects for Railway Companies Could Risk Not to Succeed – A Multi-Perspective Approach

Christian Liebchen [a], Hanno Schülldorf [b]
[a] TH Wildau, Engineering and Natural Sciences
Hochschulring 1, 15745 Wildau, Germany
[1] E-mail: liebchen@th-wildau.de, Phone: +49 (0) 3375 508 755
[b] DB Analytics – Optimization
Poststraße 20, 60329 Frankfurt am Main, Germany

Abstract
You might be aware of the following gap: There are by far more publications on promising projects on how mathematical optimization could improve the performance of railway companies, than true success stories in the sense that operations research methods really entered the practice of railways.

In this paper, we shed a bit of light on those projects, which finally did *not* enter the practice of railways. We do so by conducting a survey in which we ask both, railway practitioners who served as ordering party, and optimization experts who served as R&D solution provider.

We summarize and comment the most frequent replies to our question about the key factors why in the past mathematical optimization methods did not enter the practice of railways: expert capacity for validation, management attention, quality of input data, and "moving target" objectives. Hereby, we offer a knowledge base to future project managers. Acting accordingly with respect to definition of project goals, project design, and project management, hopefully lets them come up with even more true success stories of operations research methods in the practice of railways.

Keywords
Railway Optimization, Operations Research, Project Management, Limiting Factors, Do's and Don't's

1 Introduction

Planning for and operations of railway systems are prominent fields of application for mathematical optimization models and algorithms. In particular, during the last decades there had been reported many projects in which in particular (mixed) integer linear programming technologies had been the technology of choice to solve the real-world problems of railway companies. The particular tasks to be covered include for instance:

- network design for cargo traffic
- line planning in passenger transport
- design of the basic hourly patterns for periodic timetables
- track allocation as it is usually performed by infrastructure managers

- vehicle scheduling (locomotives, passenger train units; rotations)
- crew scheduling (train drivers and/or conductors)
- shunting planning at shunting yards
- delay management (from an infrastructure manager's perspective and/or with the focus on passenger connections)
- schedule for ticket control staff
- and many more.

Some of these elementary tasks already have been even equipped with features which are reasonable to add, but which make the mathematical task itself even more complex. Think of periodic timetabling, for which there have been included robustness issues and demand feedback loops, in particular passenger re-routing. Due to the sensitivity of the subject that we are presenting in this paper, we refrain from providing references to particular papers. Rather, we generally refer to the references that are included in survey papers such as Borndörfer et al. (2018), Cacchiani and Toth (2012), Caimi et al. (2017), and Harrod and Gorman (2011).

Nevertheless, there do not seem to be dozens of papers that do not just report on promising projects, but rather on success stories in that operations research methods really entered the practice of railways. The most striking one is of course Kroon et al. (2009), for which the authors received the INFORMS Edelman Award – "The Oscar® of O.R." – for their various mathematical contributions for the Dutch Railways. Besides, other papers in which optimization results have been used and applied for railway operations, include for instance Kohl (2003) and Liebchen (2008). In contrast, we are only aware of the paper by Gorman (2016), in which he explicitly describes the failure of a railway optimization project.

To summarize, we feel there is a kind of gap between the number of projects in which mathematical optimization experts and railway practitioners work together, and the number of success stories in the sense that the operations research methods are applied on a regular basis in the practice of railway companies. This impression is based on our personal experience in a couple of projects on both sides: the practice of railways as ordering party, as well as research institutions as solution providers. Notice that we are not limiting ourselves to daily operations, but we would also consider it as a success story, if for regular strategic questions (such as in the context of public tenders) the respective methods are applied regularly.

This is why in this paper we shed a bit of light on those projects, which did not become a "success story" in the above sense. We are interested in such projects, whose project goal in the beginning has been the application of the developed mathematical optimization methods on a regular basis, but which did not attain this goal: Are there any common key properties, which prevented several of these projects to become true success stories?

We are aware of some personal summaries and collections of general hints on selected specific success factors for railway optimization projects, provided by some experienced railway optimization experts, e.g. Borndörfer et al. (2017) and Schülldorf (2018). Yet, we think it might be of interest to set such collections on a broader basis, both for the number of experts who are sharing their experience, and for the fact that both sides – including railway practitioners as ordering party – shall contribute with their experience. This is why we initiate a survey in which we investigate this question by asking both railway

practitioners and mathematical optimization experts who were involved in such projects through a questionnaire.

The objective of this paper is to enable future project managers to setup their project goals and project design & management upon the negative experiences that other projects faced in the past. Hereby, we hope to improve the possibility that in the future more mathematical optimization projects for railway companies will become true success stories.

The paper is structured as follows. The questions that we are asking about the projects can be separated into two more or less separate classes. First, there is the general – and mainly administrative – framework of the projects (e.g., its duration, its partners, and its funding), which we describe in Section 2.1. Next, in Section 2.2 we add some problem-specific properties of the projects, some of which we suppose to be critical for a project to become a true success story. In Section 3, we shortly sketch the realization of our survey, before in Section 4, we present the results of our survey. On the one hand, we are aware that a number of 24 filled questionnaires is indeed "limited", and in particular far from being representative. On the other hand, to the best of our knowledge, this still constitutes the largest knowledge base in this specific field.

We report the results separately for the replies that we obtain from railway practitioners (N=10), and for the mathematical optimization experts (N=14), because there was a slightly different *awareness* regarding the most important reasons for project failures. Finally, we propose some conclusions for the future design and implementation of optimization projects for railway companies in Section 5.

2 General Administrative Framework of the Projects

In this section we essentially list the questions that we ask the former project members. We start with some questions to classify the projects according to some rather general properties in Subsection 2.1. Hereafter, in Subsection 2.2, we list our questions regarding rather content-related and method-related features that a project could show, and of which we can imagine that some of them might have significant influence why certain methods finally are not used in practice on a regular basis.

This distinction between the sets of questions is motivated by our goal to relate certain specific reasons for failure to some general framework properties of the projects (e.g. project duration), see Subsection 4.4 for some selected correlations.

2.1 General Administrative Framework of the Projects

In the sequel, we list the general properties of a project for which we ask in our questionnaire.

(a) Is it a railway practitioner of a mathematical optimization expert who is answering?

(b) Goal: Has the project goal been the reduction of (operational) cost and/or some increase of quality?

(c) Cost components: If the project goal was mainly cost efficiency, did the calculation of the estimated benefit of the project only include the expenses for the research, or also the full integration into the software landscape of the company including interfaces, education etc.?

(d) Funding: To what extent have the expenses for research been funded apart from public money from some research agency?

(e) Suppliers: Who has been responsible for the R&D part: universities, research institutions, software companies?

(f) Changes: Has the project goal been modified significantly during the project ("moving target")?

(g) Horizon: Which attribute fits best the target of the project: strategic decisions, planning for the operations process, the operations process itself, or any other?

(h) Target: Have the following been affected by the by the optimization results: vehicles, operational staff (train drivers, shunting assistants, conductors)?

(i) Timeline: What has been the timeline duration of the R&D part of the project (up to 12 months, 13-24 months, at least 25 months)?

(j) Volume: What has been the project volume of the R&D part in the sense of manpower (up to 12 months times men, 13-36 months times men, at least 37 months times men)?

(k) Urgency: Have there been alternative ways (without mathematical optimization) to come up with *some* solution(s) for the questions that should be answered by the new optimization methods?

(l) Input: What has been the structure of the input that was necessary to feed the mathematical optimization models: all data – except for optimization specific parameters – have been available in one existing IT system, all data have been available in IT systems but had to be combined from more than just one system, some of the data that had been necessary to feed the mathematical optimization models had not been available in any existing IT system?

(m) Output: What has been the existing IT-infrastructure to receive and further process the result of the mathematical optimization: Does the optimization result have the same data structure as it is already stored in some IT system(s), e.g., to manage solutions that earlier had been designed manually, or is there any manual post-processing required to fit the optimization data into the existing IT-infrastructure, or is there even a completely new IT system or organizational structure required in order to further work with the optimization results?

(n) Interpretability: How complex is it to "understand" the solution returned by the mathematical optimization model? It is just accessible at the level of key performance indicators (KPIs), or is it possible to comfortably dive into the very details of the solution, maybe even supported by some appropriate visualization?

In addition to these rather organizational properties of optimization R&D projects for railway companies, in the next section we present rather content-related and method-related features that a project could show, and of which we can imagine that some of them might have been decisive for the lack of success of some particular projects.

2.2 Problem-Specific Properties of the Projects

Now, we switch from the organizational perspective on the projects' framework to some of their content-related and method-related properties, which seem to us to have the potential having been a limiting factor more than only just one time. In the summary of the results of our survey (Section 4), we will put emphasis on these features, in order to identify those constellations which in the past had been the most likely show-stoppers for optimization

projects for railway companies to become true success stories – and on which future projects should pay most attention right from the very beginning.

(1) Data: The available input data finally did not meet the quality that was necessary to be able to come up with high-quality optimization results (e.g. better than solutions that were designed manually).

(2) Partial Fixing: The optimization missed the ability to accept some particular fixation for certain "variables" that were key in the point of view of the railway practitioners.

(3) Features: During the project timeline, the optimization model had been confronted with more and more detailed requirements, which finally let the performance and/or quality of the optimization methods collapse.

(4) Validation: The railway company didn't allocate a sufficient amount of expert staff to validate in detail the results of the optimization methods during the entire project timeline.

(5) Post-processing: The optimization environment lacked an editor that enabled the railway practitioners to (slightly) adjust the solution that was returned by the optimization algorithm to meet their actual practical needs and expectations?

(6) Quality: The optimization results failed to outperform the previously manually designed solutions and/or the optimization results did not achieve the quality which has been assumed in the cost-benefit-analysis that had been the basis to initiate the project.

(7) Regularity: The optimization results didn't show a certain "regularity pattern", which in the end had been expected by the railway practitioners (although not communicated as a key feature at the project kick-off).

(8) Transparency: The structure of the optimized solutions stayed somehow intransparent – "sealed" – to the railway practitioners which let them refrain from continuing to work with them in the sequel.

(9) Integration: The solution indeed optimized the specified task, but from a process perspective, subsequent tasks let expect a poor performance, when fed with the optimized solution.

(10) Strict Feasibility: The optimized solution satisfied all constraints – but other "solutions" have been preferred (e.g. designed manually by railway practitioners), although they violated some *less important constraints*.

(11) Reliability: The optimization software did not provide useful solutions on a regular basis (e.g. due to software bugs, or due to unreliable quality given that randomized elements have been deployed).

(12) Obsolescence: During the project duration, there have been new algorithmic findings which made the optimization methods in the project obsolete.

(13) Cost: The cost to make the optimization methods available in a productive context blast the cost which has been assumed in the cost-benefit-analysis that had been the basis to initiate the project.

(14) Attention: During the project duration, the "management attention" decreased, e.g.

because some protagonist within the railway company left the project.

(15) Others: These shall be specified by the respondents.

3 Realization of the Survey

For our survey, we used the online survey tool LamaPoll (2018). The survey had been designed anonymously, and it was only accessible with designated access codes. In total, we sent more than 98 access codes to both, mathematical optimization experts and managers or practitioners within railway companies. In addition, the authors filled four questionnaires about projects in which they were active. The geographical focus has been Europe (in particular Germany, the Netherlands, Switzerland, Denmark, Italy, France, Great Britain, Sweden), but we also asked experts from Northern America and China. The survey had been open from January 7th until January 21st 2019. In our inviting email, we were asking: "Hence, if ever in the past you had any project in which mathematical optimization had been intended to enter the practice of a railway company, but finally did not (fully) succeed, then we will be most thankful if you share with us your experience by answering the following questionnaire."

We received 24 questionnaires, in which at least some of the problem specific features had been answered, including 22 questionnaires that had been finished, i.e. in which 100% of the mandatory questions had been answered. These include four questionnaires of the authors. Ten questionnaires had been filled by railway employees, and the other 14 by mathematical optimization experts. Moreover, since we had been interested in the personal experience of the protagonists, when we had been asked by two experts who were active in the very same project, we invited them to fill one questionnaire each.

We were also asking – optionally – for the projects' names. Our intention was to possibly compare the answers of a railway manager on the one side, and an optimization expert on the other side, for the very same project. Indeed, in eight questionnaires the projects had been referred to with their names. But all projects had been different, so we are not able to perform such a comparison. Yet, this proves that the survey had not just been filled with ten questionnaires for the very same project.

Nevertheless, we are fully aware that N=24 is far away from letting us interpret the answers as being representative! Yet, we still consider the answers that we were able to collect as one step to provide possible explanations for the gap between the large total number of railway optimization projects, and the somehow limited number of both, true success stories from a fully practical point of view, and reports on project failures.

4 Results of the Survey

It had been our initial intention of the questions that we collected in Section 2 to be able to subdivide the answers on features that had been critical for the project's success. Unfortunately, in view of just ten replies from railway managers, we do not consider it being appropriate to subdivide this small number of answers even further.

Let us shortly explain a somehow technical step that we did for our evaluation: In Section 4.4, we are going to consider correlations between framework properties of a project and the features that could have been critical for the overall practical success of an optimization project for a railway company. To this end, we translated the text answers that the participants were able to select into points:

For instance, for the question "Who has been mainly responsible for the R&D part?",

we defined a “scale” from three (University) via two (Research Institute) to one (Software Company). Similarly, for the feature “Input: What has been the structure of the input that was necessary to feed the mathematical optimization models?”, we defined the following “monotone” scale:

(3) All data – except for optimization specific parameters – have been available in one existing IT system

(2) All data have been available in IT systems but had to be combined from more than just one system

(1) Some of the data that had been necessary to feed the mathematical optimization models have not been available in any existing IT system.

Now, we are ready to report on the answers given by the 24 participants. The scale of all the answers of Section 2.2 that we are reporting on in the sequel ranges from zero (“not relevant”) to five (“decisive”).

4.1 Most Decisive Features in the Eyes of Railway Managers

We start by providing the project features that railway managers and practitioners rated to be most critical for the practical success of an optimization project.

The Top 3 such features are:

- Attention: During the project duration, the “management attention” decreased, e.g. because some protagonist within the railway company left the project
- Validation: The railway company didn’t allocate a sufficient amount of expert staff (time capacity) to validate in detail the results of the optimization methods during the entire project timeline
- Data: The available input data finally did not meet the quality that was necessary to be able to come up with high-quality optimization results (e.g. better than solutions that were designed manually)

In Figure 1, the problem-specific features of a project are ordered decreasingly according to the relevance that railway managers and practitioners associated with them on average why the developed methods did not enter practice on a regular basis. In addition, we display the range from the minimum value (light-gray, bottom) to the maximum value (light-gray, top), as well as the 25%-75% percentile (dark-gray).

We shortly comment on the Top 3 features. Regarding “Management attention”, at least in business-oriented companies, let us have a closer look on projects that suggest a contribution to the company’s benefit (e.g. by reduction of cost). Here, we believe that the management shall mainly be driven by economical goals, which typically can be expressed in terms of money. So, we believe that only *very* rarely, a decrease of management attention can be the *only* decisive feature if a project is terminated without entering practice on a regular basis. Rather, we fear that in most of the cases there might have been deviations from the initial profit estimate (higher cost for development/implementation, less savings for the application phase), too. For primarily service-oriented projects, if additional quality cannot be “translated” precisely into additional earnings, we are fully convinced, that a loss of management attention can be the initial cause for a project to be cancelled. Very much compatible to this consideration, let us shortly include the optimization experts’ answers: In total, there have been 12 of 22 questionnaires, in which a “loss of management attention” had been rated (much) important, i.e. “4” or “5” – and *none* of these replies appeared in any of the 5 (of 22) projects, whose exclusive goal had been a reduction of cost.

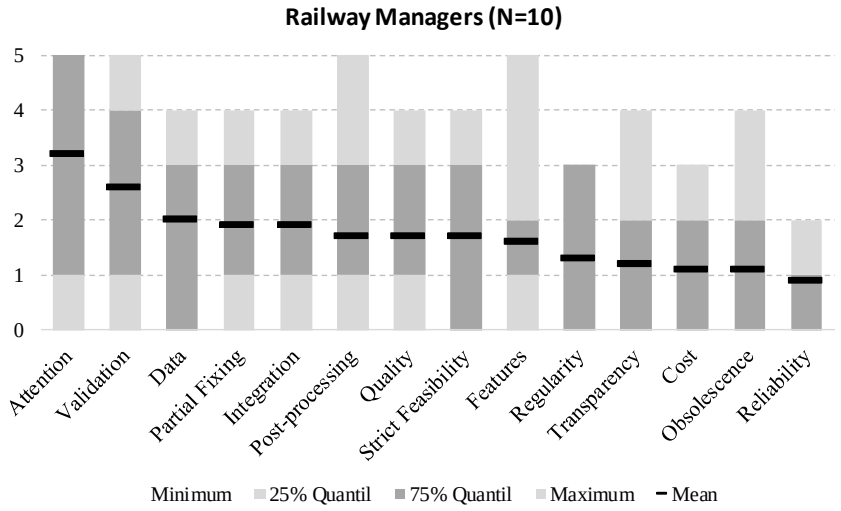


Figure 1: Relevance that railway managers and practitioners give to the problem-specific features of a project why it did not enter practice on a regular basis. For instance, "Data" had been given an average relevance of 2.0, a maximum of 4, and the 25%-75% Percentile ranges from 0 to 3

To be honest, we have been surprised in a positive way, that also railway managers seem to be aware that a shortage of expert capacity for validation – and thus, in the sequel, for the improvement of software prototypes – can indeed be a decisive feature for the unsuccessful end of a project. Nevertheless, optimization experts associate with it an even larger relevance, see also Sections 4.2 and 4.3.

In Section 4.4, we will see that the fact that the availability and/or quality of input data has been rated with 3 or 4 by 50% of the railway managers in particular correlates with a general feature of the project, namely the IT system environment on the input data side. However, recall that when discussing correlation, we are still aware that N=10 and N=24 are not suited to guarantee any true statistical significance.

4.2 Most Decisive Features in the Eyes of Optimization Experts

Now, in Figure 2 let us turn to the perspective of the mathematical optimization experts. Much like the railway managers and practitioners, they rated the shortage of expert capacity for validation being relevant – but with a by far more striking average of 3.64 out of 5.

Among the Top 5, here we also get what we called "strict feasibility", for short, i.e., the fact that in the end practitioners might have made use of the possibility to "relax" some of the constraints that have been imposed to the optimization algorithms, still considering their manually designed "solution" to be "practically feasible".

In addition, mathematical optimization experts consider the cost for making the optimization methods available in a productive context relevant, if they exceed the initial cost-benefit-analysis of the project (2.50).

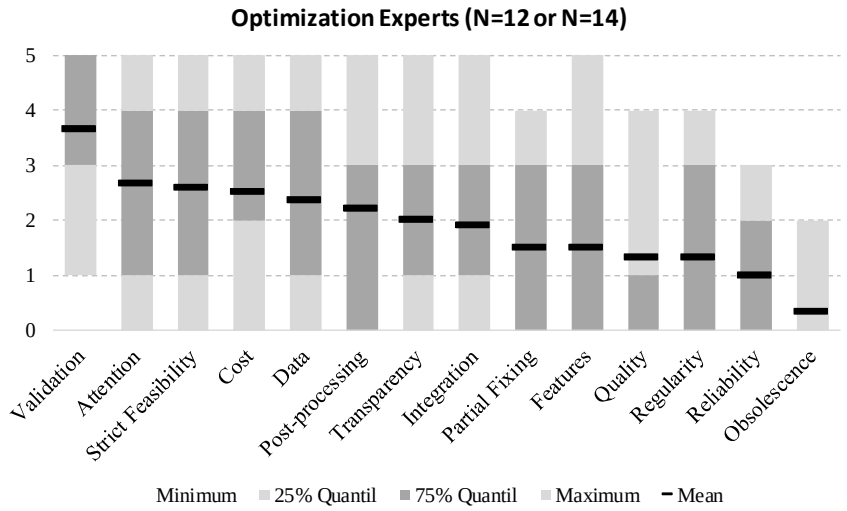


Figure 2: Relevance that mathematical optimization experts give to the problem-specific features of a project why it did not enter practice on a regular basis.

4.3 Features with Largest Deviations Between Managers & Optimization Experts

Even more interesting insights might arise from comparing the relevance that the railway managers and practitioners assigned with the one that the mathematical optimization experts did assign. In Figure 3 we subtract the mean of the latter from the mean of the former.

For any of the differences that are displayed in the chart, keep in mind that we have to assume that there is *no* project, for which we got the answers from both sides, i.e., railway managers and optimization experts. Hence, the primary reason for any differences between the two perspectives could still simply lie in a different nature of the projects. Nevertheless, assuming that the major source for the difference could indeed lie in the role of the protagonists, we propose the difference values as a kind of indication.

At first sight, one could observe that from the perspective of the railway managers, obsolescence of the algorithm appears to be much more relevant for a project not to attain its full goals, compared to the understanding of the mathematical optimization experts with respect to *their* methods. But recall from the previous figures that the values for the feature "obsolescence" are 1.1 for the railway managers and practitioners, but only 0.3 for the mathematical optimization experts, which yields the value 1.1–0.3 = 0.8. In particular, both partners did only observe a (very) small relevance in "obsolescence" of the methods.

At the other end of the scale, it had been the impression of the mathematical optimization experts that the full integration of their methods into the software landscape of the company turned out to be too costly in the end, and thus become a "show-stopper". This is reflected by a value of 2.5. Interesting enough, this is not confirmed by the railway managers and practitioners, who rate this feature only 1.1, which thus provides a difference of 1.1–2.5 = –1.4.

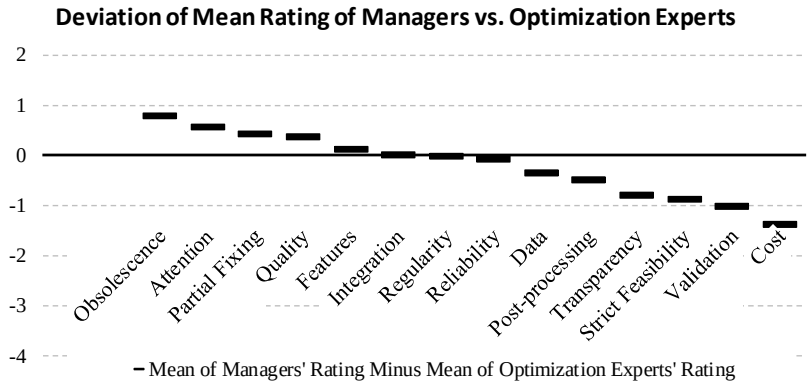


Figure 3: Difference between the mean rating assigned by the railway managers and practitioners and the mathematical optimization experts

Although the feature "Validation" (expert capacity to (in-) validate intermediate results) belongs to the Top 2 features of both, railway managers and optimization experts, there is one of the largest *gaps* between the intensity that they assigned to this feature: 2.6 by the railway managers, but even 3.6 by the optimization experts, hereby marking the top score of the entire survey. This provides a difference of 2.6–3.6 = –1.0.

Moreover, we find another result interesting. Consider the "moving target" question (f), which we put in the "general framework" section of our questionnaire. Only 20% of the railway managers said there has been a "moving target" within their projects – while as many as 57% of the mathematical optimization experts report this as their impression! We suppose that this could be due to different understandings regarding the degree of specification at the very beginning of a project: Maybe, railway managers sometimes cannot (or do not want to?) specify any requirement in most detail when launching a project. Then, later, when they "add" some piece of specification, the mathematical optimization experts could experience such a late specification already as a significant modification of the project goal, or "moving target".

4.4 Selected Correlations Between General Framework and Specific Features

Finally, although we are fully aware that 10+14 filled questionnaires unfortunately cannot be representative for all projects and project members, we still perform some correlation test and invite the reader to interpret it as a slight indication.

To this end, we computed the correlation between each pair of feature of the general administrative framework of the projects (see Section 2), and of the problem-specific properties of the projects (see Section). Among the roughly 250 possible combinations, one can detect three where the absolute value of the correlation is larger than 0.5, and thus could tend to be "significant" (which it is not, due to our relatively small sample size).

- 0.65
  The more academic the partner who has been mainly responsible for the R&D part (3 = university, 2 = research institution, 1 = software company)…
  … the more severe the lack of railway expert capacity for validation for the situation when a mathematical optimization project does not meet its full

goals.

- –0.55
  The better the structure of the input data (3 = all data available in *one* existing IT-system, 2 = all data available in some IT-system, but these must be combined, 1 = some of the data necessary for the mathematical optimization models have not been available in any existing IT system)…
  … the less likely the project's results did not get used in practice on a regular basis because the available input data finally did not meet the quality that was necessary to be able to come up with high-quality optimization results.
  Of course, this relationship sounds absolutely reasonable. So we primarily interpret the observed correlation as a kind of cross-check question to evaluate the consistency of the answers, rather than some new insight.
- 0.54
  The fact that the project goal has been modified significantly during the project ("moving target") correlates positively with…
  … the priority that had been assigned to the fact that in the end solutions that had been designed manually by railway practitioners made their way into practice, although they violated some (less important) constraints which the optimization software still had to respect ("strict feasibility").
  A similar positive correlation (0.48) can be observed between the "moving target" property in the general framework, and the project-specific feature "regularity" (The optimization results didn't show a certain "regularity pattern", which in the end had been expected by the railway practitioners, although not communicated as a key feature at the project kick-off).
  Also here, we consider these two correlations very much reasonable: A moving target and either of "regularity requirements" (not communicated at the very beginning) and "strict feasibility" (relaxations at the very end) can be regarded as two sides of the same coin. This makes us believe in the quality of the answers that we received, despite their small number.

To summarize, we were able to statistically observe some correlations between features that we were asking in the context of "general framework" of a project, and problem-specific reasons why a project finally did not enter practice on a regular basis. While the second and the third one that we are reporting on are rather confirming somehow trivial assumptions, the first one might constitute a "lesson learnt": In particular, when the R&D part is contributed by a university, it is even more critical for the actual success of the entire project that the railway company allocates a sufficient amount of capacity of practical experts in order to evaluate intermediate results.

4.5 Further Comments by the Experts

The last – optional – question of our survey has been: "Have there been other features for the project finally not to meet its full goals?", i.e. those, which our questionnaire did *not* include already (see Section 2.2). In the sequel, we report some answers that we received for this question:

- Complexity of Control
  In a sense symmetric to a lack of transparency of the solution, it can also be negative for an optimization tool, if it leaves too many control parameters to the end user, where the effects (and interactions) of the parameters could not be anticipated adequately.

- Employee Participation
  In a project that touched the working times of staff, the best solutions that the optimization tool was able to deliver have not been accepted by the unions, and thus decreased the benefit of the project, and/or delayed its implementation significantly.
- Management Implementation
  The change required towards an automated optimization method was significant: therefore, a relevant contribution from the management would have been required to make it used in real practice.
- Managerial Consistency
  The gap between the management expectation to substantial benefits on the one hand vs. the very detailed "parameter battle" with the experts on the other side could not be closed. The correct parameters have an extremely high impact on the optimization result. Therefore, a lot of time of the railway experts is needed (see also "Validation"). This could not be communicated to the management.
- Organizational Changes
  Suboptimization within organizations and organizational units meant the global optimum provided required large organizational changes to be implemented in practice.
- Performance
  The runtime of the optimization was much higher than expected. The optimization approach used wasn't suitable for the size of the problems as it is relevant in practice. If the scientist is able to deliver a high quality solution after a computation time of 48h, then it is only of limited use for a practitioner, if he requires the results in a „live" context.
- Rolling Horizon
  If a shift plan had to be designed for some general week, it should of course "glue well" between Sunday 23:59 and Monday 0:00, without leaving an expensive transition back to the initial state outside the objective function.

5 Conclusions

Even though the number of replies that we received stayed rather small, we feel able to provide some suggestions for the future design and management of operational research R&D projects for railway companies. Recall that here we are not referring to projects, in which just some study for the potential of some new algorithmic ideas is to be conducted. Rather, we are considering projects that have the goal, that at the end the optimization methods will be used in practice on a regular basis.

- The by far most reported reason why in the past the results of optimization projects for railway projects did not enter practice on a regular basis, is a lack of *expert capacity within the railway companies for the validation of intermediate results*. An appropriate amount of their capacity must be planned from the very beginning of the project, and then guaranteed throughout the lifetime of the project.
- This point has been rated even more important, if the R&D part in the project has been developed by a university partner – presumably, software companies

fixed the required capacity allocation already in their contracts? In any case, we encourage in particular university partners to do so for future projects.

- The *availability, consistency, and quality of input data* can of course be decisive for the success or failure of any project. Hence, we recommend in particular to the railway companies to let their R&D partner evaluate the quality of the available input data in detail prior to launching the actual project for the development of algorithms. If there were some significant deficiencies detected, then it could make sense to postpone the optimization project until the input required for it is available.
- Regarding *management attention*, let us only consider quality-oriented projects, where the contribution to the benefit of the company cannot be expressed explicitly (in terms of money). We agree that management attention risks to be volatile in particular in this case. Here, we can only recommend to the companies only to initiate such projects, of which they can be sure that their (strategic?) quality goals will *not* change during the timeline of the project.
- Finally, let us recall the "*moving target*" property of a project, which we observed to be much more present in the eyes of mathematical optimization experts. To prevent a project to fail due to this feature, we recommend to the railway managers to put very much emphasis on a detailed description of the requirements for the optimization tool, and prevent any deliberate "lazy specification".

6 Acknowledgements

We are most thankful to the 24 respondents of our survey, to another half a dozen experts who answered our open "other features" question whose answers are reported in Section 4.5, and to Verena Appeldorn (DB Cargo AG), whose feedback to an earlier initiative laid the ground for this survey.

References

Borndörfer, R. et al., 2017. "Recent success stories on integrated optimization of railway systems", *Transportation Research Part C: Emerging Technologies*, vol. 74, pp. 196 – 211. https://doi.org/10.1016/j.trc.2016.11.015

Borndörfer, R. et al. (eds.), 2018. *Handbook of Optimization in the Railway Industry*, ISBN 978-3-319-72152-1, Springer International Publishing.

Cacchiani, V. and Toth, P., 2012. "Nominal and Robust Train Timetabling Problems", *European Journal of Operational Research*, vol. 219, pp. 727 – 737. https://doi.org/10.1016/j.ejor.2011.11.003

Caimi, G. et. al., 2017. "Models for railway timetable optimization: Applicability and applications in practice", *Journal of Rail Transport Planning & Management*, vol. 6, pp. 285 – 312. https://doi.org/10.1016/j.jrtpm.2016.11.002

Gorman, M., 2016. "From Magnum Opus to Mea Culpa: A Cautionary Tale of Lessons Learned from a Failed Decision Support System", INFORMS Journal on Applied Analytics, vol. 46m pp. 185-195. https://doi.org/10.1287/inte.2015.0818

Harrod, S. and Gorman, M.F., 2011. “Operations Research for Freight Train Routing and Scheduling”, In: Cochran, J.J. et al. (eds.), *Wiley Encyclopedia of Operations Research and Management Science*, ISBN 978-0-470-40063-0, John Wiley & Sons, Inc.

Kohl, N., 2003, “Solving the World’s Largest Crew Scheduling Problem”, *orbit Xtra*, pp. 8 – 12.

Kroon, L. et al., 2009. “The New Dutch Timetable: The OR Revolution”, *Interfaces*, vol. 39, pp. 6 – 17. https://doi.org/10.1287/inte.1080.0409

LamaPoll, 2018. An online survey tool by Lamano GmbH & Co. KG, Berlin, Germany. https://www.lamapoll.de/Support/Impressum

Liebchen, C., 2008. “The First Optimized Railway Timetable in Practice”, *Transportation Science*, vol. 42, pp. 420 – 435. https://doi.org/10.1287/trsc.1080.0240

Schülldorf, H., 2018. “Der weite Weg der Optimierung zum produktiven Einsatz“, presentation at the *Sitzung GOR - Arbeitsgruppe Logistik und Verkehr*, May 18th, 2018, Karlsruhe, Germany, in German.

Computing Base Train Equivalents for Delay-Based Capacity Analysis with Multiple Types of Trains

Tzu-Ya Lin [a], Ying-Chun Lin [a], Yung-Cheng (Rex) Lai [a, 1]
[a] Department of Civil Engineering National Taiwan University
Room 313, Civil Engineering Building,
No. 1, Roosevelt Road, Sec. 4, Taipei, 10617 Taiwan
[1] E-mail: yclai@ntu.edu.tw, Phone: +886-2-3366-4243

Abstract
Different types of trains may have substantially dissimilar characteristics, resulting in various capacity impacts. The concept of base train equivalent (BTE) was proposed to standardize different train types into a universal unit, namely, base train unit (BTU). However, the previously developed delay-based model suffers from consistency issue, and its application is limited to only two train types. Thus, this study proposes a new concept of delay-based BTE computation and corresponding BTE models. The dynamic BTE model considers volume and heterogeneity and aims to reflect fully the actual capacity impact of non-base trains. The fixed BTE model identifies the most appropriate BTE value at a particular traffic heterogeneity. Results from the case studies demonstrate that the proposed method can address scenarios with all types of traffic mixes and multiple train types. The unit of delay-based rail capacity can be converted into a standard unit using the proposed models. The effect of an additional train can be easily assessed, and the capacity measurements from different lines or systems can be compared and evaluated.

Keywords
Rail Transport, Capacity Analysis, Base Train Equivalent

1 Introduction

Multiple types of trains usually operate on a railroad line to accommodate different demands. Different train types may have substantially dissimilar characteristics, resulting in various capacity impacts. Lai et al. (2012) proposed the use of base train equivalent (BTE) to convert different train types into a universal unit, namely, base train unit (BTU) Delay-(Lai et al. (2012)) and headway-based approaches (Lai et al. (2015)) were developed to compute BTE depending on the types of capacity model.

Delay, which uses parametric and simulation models, is a common output of capacity analysis in North America (Confessore et al. (2009); Dingler et al. (2014); Krueger (1999); Lai et al. (2012); Lai and Barkan (2009); Pouryousef and Lautala. (2013); Prokopy and Rubin. (1975); Sogin et al. (2013) and Shih et al. (2015)). Although the delay-based BTE model was established by Lai et al. (2012), their model adopted the delay-based approach from highway research and defined BTE as the delay ratio of a marginal non-base train over a base train. A deficiency of this method is that the BTU converted from a mixed traffic through the BTE may be different from the number of base trains at the same delay level. In addition, the delay-based BTE model cannot handle scenarios with more than two train types (Lai et al. (2012)).

In the present study, we proposed a new concept and developed a set of corresponding delay-based BTE models. Furthermore, we extended the model framework to accommodate multiple types of trains. The unit of delay-based rail capacity could be converted into a standard unit through the proposed models. The capacity measurements from different lines or systems could be compared and evaluated.

2 Methodology

Figure 1 demonstrates the new concept proposed in this study for determining BTE. The two points from the mixed and base flows at the same delay level are used to compute the BTE for non-base trains. For example, the delay level of mixed traffic for 18 days in the mixed flow is equivalent to that of the homogeneous traffic with 52 base trains in the base blow. Therefore, if the non-base trains in the mixed flow are converted into base trains through BTE, then the total number of base trains after the conversion (30×1+10×BTE) should be 52, thereby resulting in a BTE value equal to 2.2. In this way, we can easily compare different traffic flows in the same standard and convert the mixed flow to the base flow meaningfully and consistently.

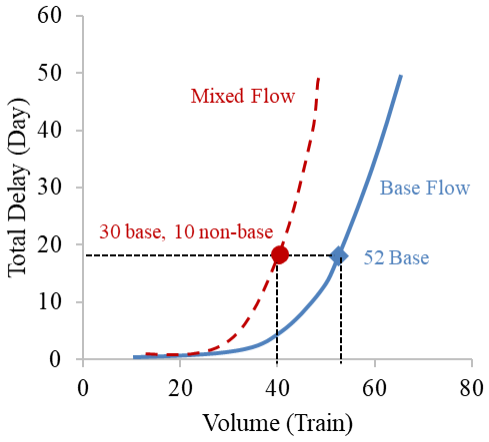


Figure 1: Concept for BTE computation

Several types of BTE model are developed on the basis of the new concept proposed in this study. In terms of a particular route, BTE will only vary with traffic volume and heterogeneity because most of the other factors are fixed. Therefore, this study initially develops dynamic BTE models with consideration of volume and heterogeneity. Furthermore, we develop a fixed BTE model with consideration of only heterogeneity because its influence to BTE value is considerably higher than that of traffic volume.

Another breakthrough of this study is enabling the possibility to compute BTEs for multiple types of train. If only two types of trains exist, BTE can be directly computed. However, the same model cannot be applied directly to scenarios with multiple types of train due to additional unknown BTEs. Therefore, this study also adopts the concept of

projecting vector to identify a suitable BTE for each type of non-base trains.

2.1 Dynamic BTE Model

Given the new concept of the BTE computation, the computational process should first determine the number of base trains in the homogeneous flow that corresponds to the delay level of the mixed traffic in the heterogeneous flow. Equation (1) can be used to determine BTE to non-base trains by allocating impacts to non-base trains, where n_b is number of base type of trains in the mixed flow; n_i is number of ith type of non-base train in the mixed flow; n_B is number of base type of trains in the base flow; i is index for train type; I is total number of types of non-base train in the mixed flow; E_i is BTE of the ith type of non-base train; E_b is BTE of the base train in the mixed flow (= 1); E_B is BTE of the base train in the base flow (= 1).

$$n_b E_b + \sum_{i}^{I} n_i E_i = n_B E_B. \tag{1}$$

If only one type of non-base train is found in mixed flow, then the BTE value for non-base trains (E_i) can be easily determined by Equation (1). However, if more than one type of non-base train is observed, then multiple unknown BTEs with only one equation exist. We build the coordinates in three-dimensional space to determine each relative position (Figure 2).

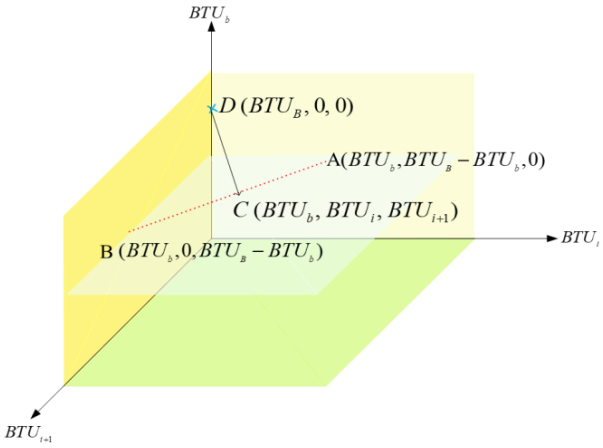


Figure 2: Schematic of BTUs for the three types of train

According to Equation (1), the BTU of each type of train can be summarized as Equation (2). The right-hand side is the BTU of the base flow, and the left-hand side is the BTU of the mixed flow.

$$BTU_b + \sum_{i=1}^{I} BTU_i = BTU_B. \tag{2}$$

This example can be illustrated as a three-dimensional space in Figure 2. In this figure, point *D* demonstrates the number of BTUs in the base flow, that is, BTU_B, and the red dashed line ($\overrightarrow{AB}$) represents the feasible region for BTU_i and BTU_{i+1} (and the corresponding E_i and E_{i+1}) in the mixed flow. Each point in the feasible region ($\overrightarrow{AB}$) reflects the same delay and BTU with the base flow. To determine the appropriate values of E_i and E_{i+1}, we project point *D* to $\overrightarrow{AB}$ by setting the inner product of the direction vector [$\overrightarrow{AB}$ = (0, BTU_B - BTU_b, -(BTU_B - BTU_b))] and normal vector [$\overrightarrow{CD}$ = (BTU_B - BTU_b), -BTU_i, BTU_{i+1}] to zero, as described in Equation (3), where $\vec{u}$ is direction vector; $\vec{v}$ is normal vector. Equations (4) and (5) demonstrate the process of determining the BTEs (i.e., E_i and E_{i+1}) for the two types of non-base train. Equation (4) corresponds to the detailed process of the inner product. Finally, the BTEs (i.e., E_i and E_{i+1}) can be obtained by using Equation (5). Although we take three types of train as examples here, the proposed process can be easily applied to scenarios with four or more types of train.

$$\vec{u} \cdot \vec{v} = 0. \tag{3}$$

$$\vec{u} \cdot \vec{v} = 0 = \overrightarrow{AB} \cdot \overrightarrow{CD} = \left[0, BTU_B - BTU_b, -(BTU_B - BTU_b)\right] \cdot \begin{bmatrix} BTU_B - BTU_b \\ -BTU_i \\ -BTU_{i+1} \end{bmatrix}. \tag{4}$$

$$E_i = \frac{(BTU_B - BTU_b)}{n_i}, \quad E_{i+1} = \frac{(BTU_B - BTU_b)}{n_{i+1}}. \tag{5}$$

2.2 Fixed BTE Model

The fixed BTE model adopts the same concept used in the proposed delay-based BTE computational process in this study. However, the fixed BTE model aims to identify the most appropriate BTE value to represent a specific heterogeneity regardless of the traffic volume. In the fixed BTE model, the mixed flow is no longer only a point but a line with the same heterogeneity (red line in Figure 3).

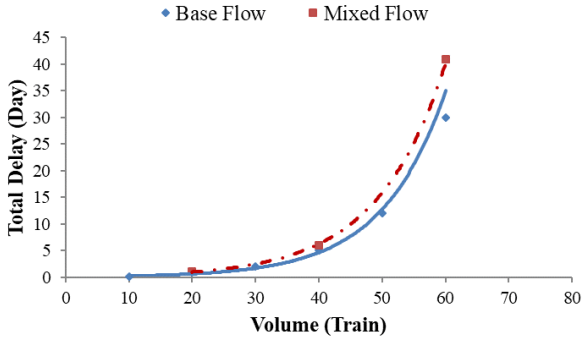


Figure 3: Delay–volume curve of the fixed model

As shown in Figure 3, the most appropriate BTE value should convert the mixed flow (red dashed line) to the base flow (blue solid line) at any delay level. Thus, the proposed process applies an iterative process to determine the most appropriate BTE value by minimizing the difference between the BTU in the base flow and that in the mixed flow (after conversion) with the given delay levels (K) (Equation (6), where K is number of selected delay levels; BTU_{mk} is BTU in the mixed flow at the Kth delay level; BTU_{bk} is BTU in the base flow at the Kth delay level; n_{ik} is number of type i non-base trains in the mixed flow of Kth delay level; n_{bk} is number of base trains in the mixed flow of Kth delay level.).

If three delay levels are selected in the model and we expanded Equation (6), and remove the squared and root denoting the difference between the base and mixed BTUs of each selected delay base as d_k, then we can move the number of base trains in the mixed flow (n_b) and the base flow (n_B)and dk can be moved to the right-hand side because they are all known values (Equation (7)). The fixed BTE model aims to identify only one BTE for a particular heterogeneity. E_i in each of the three equations in Equation (7) is the same. Therefore, it can be regarded as one equation. The right-hand side denotes constant f, and the general equation is Equation (8). If only one type of non-base train is found in the mixed flow, then the most appropriate BTE value (E_i) can be determined by Equation (8) with a given set of delay levels. However, if more than one type of non-base train is found, then more than one possible solution exists.

$$\begin{aligned} &\min \sum_{k=1}^{K} \sqrt{(BTU_{mk} - BTU_{Bk})^2} \\ &= \min \sum_{k=1}^{K} \sqrt{(\sum_{i=1}^{I} n_{ik} \times E_i + n_{bk} \times E_b - BTU_{Bk})^2}. \end{aligned} \tag{6}$$

$$\begin{aligned} \sum_{i=1}^{I} n_{i,1} \times E_i &= d_1 + n_{B,1} - n_{b,1}, \\ \sum_{i=1}^{I} n_{i,2} \times E_i &= d_2 + n_{B,2} - n_{b,2}, \\ \sum_{i=1}^{I} n_{i,3} \times E_i &= d_3 + n_{B,3} - n_{b,3}. \end{aligned} \tag{7}$$

$$\sum_{i}^{I} n_i \times E_i = f. \tag{8}$$

In Figure 4, we also take three types of train as example, which are easily presented in the three-dimensional space. From the figure, points D_1–D_3 demonstrate the number of BTUs in the base flow with different volumes but same heterogeneity, and the red dashed line ($\overrightarrow{A_1B_1}$~$\overrightarrow{A_3B_3}$) represents the feasible region for BTU_i and BTU_{i+1} in the mixed flow. Figure 4 shows three sub-spaces based on the three selected delay levels (K). The mixed flow distribution is proportional, and the base line is coordinated with the BTU_b axis. The aforementioned concept shows that only one delay base conduct projection can be selected.

The solution process for multiple types of train is almost the same as that in the dynamic BTE model for a similar case. Each point in the feasible region ($\overrightarrow{AB}$) reflects the same delay and BTU with base flow. To determine appropriate values of E_i and E_{i+1}, we project point D to $\overrightarrow{AB}$ by setting the inner product of the direction vector [$\overrightarrow{AB}$= (0, BTU_B - BTU_b, -(BTU_B- BTU_b))] and normal vector [$\overrightarrow{CD}$= (BTU_B - BTU_b), -BTU_i, BTU_{i+1}] to zero. Equations (4), (9),

and (10) demonstrate the process to determine the BTEs (i.e., E_i and E_{i+1}) for the two types of non-base train. Equation (9) can be derived from Equations (4) and (8). Finally, the BTEs (i.e., E_i and E_{i+1}) can be obtained using Equation (10). Similarly, although we take three types of train as example here, the proposed process can be applied to scenarios with four or more types of train.

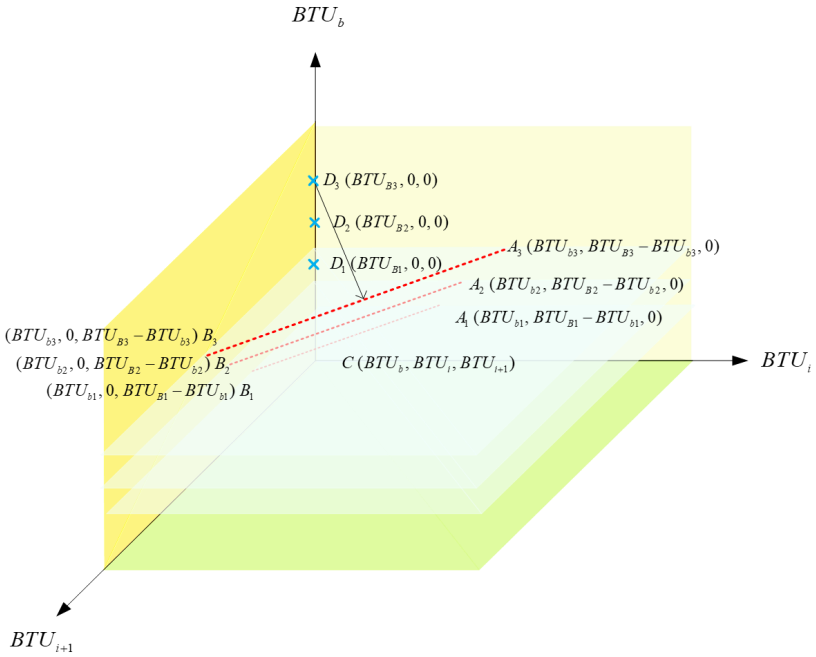


Figure 4: Schematic of BTUs of the three types of trains for the fixed BTU

$$BTU_i = BTU_{i+1} = \frac{f}{2}. \tag{9}$$

$$E_i = \frac{\frac{f}{2}}{n_i}, \quad E_{i+1} = \frac{\frac{f}{2}}{n_{i+1}}. \tag{10}$$

3 Case study

To demonstrate the use of the proposed model, dynamic and fixed BTE models are applied to scenarios with three train types. For the three train types, we add intermodal trains and regard it as a base train to understand the changes in the BTE values for coal and passenger trains. Table 1 shows the characteristics of all train types.

Table 1: Train Characteristics

Train	Passenger train	Intermodal	Coal train
Locomotive	P42-DC locomotives	5 SD70 locomotives	3 SD70 locomotives
No. of Cars	13 cars	93 cars	115 cars
Weight (tons)	500 tons	5,900 tons	16,445 tons
Train Length	500 feet	5,649 feet	6,325 feet
HP/TT	15.4	3.64	0.78
Max Speed	79 mph	70 mph	50 mph

RTC simulation software is used to obtain the delay data. This case study is based on a set of inputs that represent the typical characteristics of a Midwestern North American single-track main line. The route characteristics are as follows: (1) section length: 262.25 miles; (2) siding spacing: 2.75 miles; (3) signal spacing: 2.75 miles; (4) three-aspect signaling system; (5) sidings are evenly distributed in the section; (6) the number of bidirectional train departures is consistent; and (7) passenger train stops at three stations on the section are evenly distributed, and dwell time is 2 minutes (Dingler et al. (2013)). According to each different combinations of train type, we perform 30 different random seeds in RTC to acquire average delay. An alternative method is the use of other types of delay-based capacity model, such as the parametric capacity model. These delay data can then be used to compute BTE values by using the proposed computational process.

3.1 Analytical Results for Multiple Train Types

Dynamic BTE of Three Train Types

In the three train types, intermodal is added as a base train. The non-base trains are coal and passenger trains. We use 10% of train heterogeneity for the interval unit. A total of 36 heterogeneous groups are found in the three train types, and each heterogeneous group has three volumes, that is, 20, 40, and 60 trains. Therefore, 108 types of train combinations are found.

Figure 5 shows the BTEs of the three train types. For the case of 20 mixed trains (10% intermodal, 10% passenger, and 80% coal trains), the BTEs of these three train types are 1, 7.73, and 0.8. However, the BTEs of 20 mixed trains with 10% intermodal, 80% passenger train, and 10% coal train can also be considered 1, 0.5, and 4.02. In other words, when the percentage of a train type in the traffic mix is lower, its BTE is usually higher because these trains are more special than other trains that have a higher tendency to disturb the traffic

flow and incur higher delay.

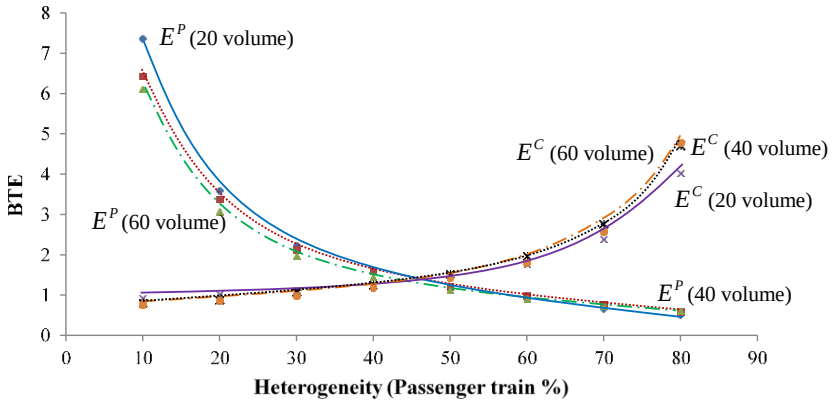


Figure 5: BTE of three train types for the dynamic model
(given the proportion of intermodal trains is 10%)

Fixed BTE of Three Train Types

In this case, the proportion of every type of train ranges from 10% to 80%, thereby resulting in 36 combinations. A fixed BTE should be the most appropriate one among the 36 combinations evaluated in the process.

Figure 6 shows the fixed BTE value of two non-base trains assuming that the proportion of intermodal trains is 10%. If the proportion of one type of train is lower, then its BTE is higher, and vice versa. This trend is the same as the previous case, in which a train type with lower percentage affects the flow more considerably.

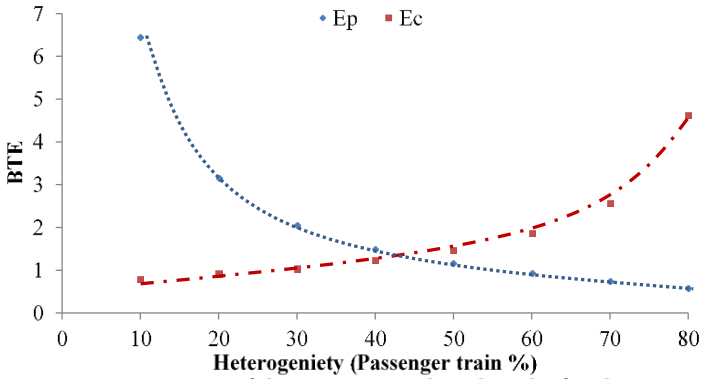


Figure 6: BTE of three train types based on the fixed BTE model
(given the proportion of intermodal trains is 10%)

4 Discussion: BTE Application

Capacity is usually defined as the maximum system throughput. We can further define the maximum throughput to the maximum base trains using BTE. Table 2a presents a capacity evaluation by using capacity and BTU for different traffic compositions among various dates. In terms of traffic volume, 40 trains exist for each day of the periods. However, if the traffic volume is converted into BTU, then they are all relatively different. Similarly, Table 2b shows three different sections. Capacity and BTE in different sections are dissimilar due to the difference in route characteristic. The comparison in BTU is considerably more meaningful than that in the number of trains.

Table 2: Capacity Evaluation Based on BTU

(a) Same Section

Date	n_P	n_I	n_C	E_P	E_I	E_C	N	BTU	C	V/C
3/1	4	32	4	1.72	1	1.72	40	45.76	55	0.832
3/2	20	8	12	1.04	1	1.72	40	49.44	55	0.899
3/3	20	16	4	0.7	1	3.49	40	43.96	55	0.799
3/4	12	12	16	1.67	1	1.24	40	51.88	55	0.943
3/5	8	8	24	2.92	1	0.97	40	54.64	55	0.993

(b) Different Sections

Section	n_P	n_I	n_C	E_P	E_I	E_C	N	BTU	C	V/C
1	12	12	16	1.67	1	1.24	40	51.88	55	0.943
2	32	4	4	0.48	1	3.87	40	34.84	66	0.528
3	4	16	20	4.41	1	0.88	40	51.24	51	1.005

Section 2：length = 161.75 miles, siding = 5.5 miles, signal = 2.75 miles

Section 3：length = 109.75 miles, siding = 16.5 miles, signal = 1.375 miles

5 Conclusions

This study proposes a new concept of delay-based BTE computation and the corresponding BTE models. The dynamic BTE model considers volume and heterogeneity and aims to reflect fully the actual capacity impact of non-base trains. The fixed BTE model identifies the most appropriate BTE value at a particular traffic heterogeneity. The results from the case studies demonstrate that the proposed method can address scenarios with all types of traffic mixes and multiple types of trains. The unit of delay-based rail capacity can be converted into a standard unit using the proposed models. The capacity measurements from different lines or systems can be compared and evaluated

References

Confessore, G., Cicini, P. and Luca, P.D. 2009. A simulation-based approach for estimating the commercial capacity of railways. Proceedings of the 2009 Winter Simulation Conference, New York, NY, USA.

Dingler, M.H., Lai, Y.C. and Barkan, C.P.L. 2014. Effect of train-type heterogeneity on single-track heavy haul railway line capacity. Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit, 228 (8): 845-856.

Krueger, H. 1999. Parametric modeling in rail capacity planning. Simulation Conference Proceedings, Proceedings of Winter Simulation Conference, Phoenix, AZ. 2: 1194-1200.

Lai, Y.C. and Barkan, C.P.L. 2009. Enhanced parametric railway capacity evaluation tool. Transportation Research Record: Journal of the Transportation Research Board, 2117: 33-40.

Lai, Y.C., Liu, Y.H. and Lin, T.Y. 2012. Development of Base Train Equivalents to Standardize Trains for Capacity Analysis, Transportation Research Record: Journal of the Transportation Research Board, 2289: 119-125.

Lai, Y.C., Liu, Y.H. and Lin, Y.J. 2015. Standardization of Capacity Unit for Headway-based Rail Capacity Analysis, Transportation Research Part C: Emerging Technologies, 57: 68-84.

Pouryousef, H. and Lautala, P. 2013. White. Review of Capacity Measurement Methodologies; Similaries and Differences in the U.S. and European Railroads. Presented at 92th Annual Meeting of the Transportation Research Board, Washington, D. C.

Prokopy, J.C. and Rubin, R.B. 1975. Parametric analysis of railway line capacity. Federal Railroad Administration, Washington.

Sogin, S., Lai, Y.C., Dick, C.T. and Barkan, C.P.L. 2013. Comparison of Capacity of Single- and Double-Track Rail Lines, Transportation Research Record: Journal of the Transportation Research Board, 2374: 111-118.

Shih, M. C., Dick, C.T. and Barkan, C.P.L. 2015. Impact of Passenger Train Capacity and Level of Service on Shared Rail Corridors with Multiple Types of Freight Trains, Transportation Research Record: Journal of the Transportation Research Board, 2475: 63-71.

Finding feasible timetable solutions for the Stockholm area

Olov Lindfeldt
Head of Traffic Planning, MTR Pendeltågen, Stockholm
Box 10038, 121 26, Stockholm Globen, Sweden
E-mail: olov.lindfeldt@mtr.se, Phone: +46 (0) 729 80 25 09

Abstract
MTR (Mass Transit Railway) is contracted by Stockholm Public Transport (SLL) to operate the Stockholm commuter trains. The number of passengers is increasing and traffic is expected to increase by 50 % in ten years. This will therefore require further investigations to enable investments in additional infrastructure and rolling stock.

A generic model has been developed in order to screen future timetable situations and find resource efficient timetable alternatives and investments needed to enable the expected traffic increase.

Short turning traffic lines is one way to reach high efficiency for a commuter system. However, the sequence of short turning and full route lines will affect congestion heavily. Consequently different permutations of a termination pattern results in different passenger distributions on the traffic lines. The core idea of the timetabling model is to combine congestion efficient permutations for the four branches into network timetables.

A number of important features of the timetable are influenced by the choice of termination patterns, permutations of these patterns, the time rotation of the entire timetable and the requisite of symmetry. The latter is required in order to enable long distance traffic on shared line sections. Examples of important features are: the termination times, the number of train set needed, the need for additional termination tracks and the recovery and punctuality that can be reached.

A brief description of the commuter rail network, the demand and the prerequisites for the timetable are presented and discussed. Similarly the main ideas of the generic model are outlined. The method is elucidated by an illustration of a future traffic increase by 25 %.

Keywords
Planning, scheduling, robust timetables, congestion management

1 Introduction

From December 2016 MTR, Mass Transit Railway, is contracted by Stockholm Public Transport (SLL) to operate the commuter trains in Stockholm. The political ambition is to increase peak hour traffic by 50 % until 2030 (Tillväxt- och regionplaneförvaltningen (2017)). The railway network, owned and administrated by the National Transport Administration (Trafikverket), is however already heavily utilized, resulting into lower punctuality and higher passenger congestion on the services than desirable.

Feasibility studies addressing infrastructure measures are initiated. MTR, as key operator holds extensive knowledge and insights within the operational sector, is actively engaged and involved in these studies. One of MTR´s contributions is a timetable

generating model that screens the possibilities to find resource efficient timetable solutions for the future.

This screening approach is useful since parts of the network are shared with long distance and regional traffic, which requires a coordination where the timetable for the commuter trains cannot be optimized independently.

1.1 Network and demand

The network is presented in Figure 1 and consists of four branches. All line sections are double or quadruple lines, except for the southern part of Nynäs line (Hemfosa – Nynäshamn) that is still single line with crossing loops. The commuter traffic is well separated from other rail traffic with quadruple lines on most sections shared by long distance and regional traffic. Two important exceptions are the end section of Mälar line (Kallhäll – Bålsta) and East Coast line (Upplands Väsby – Märsta/Uppsala) where a thorough timetable coordination is required to manage the traffic mix.

The mid-section consists of the new commuter train tunnel, City line, launched in 2017, that separates commuter traffic through central Stockholm. This line however, has a limited capacity (Lindfeldt (2017)).

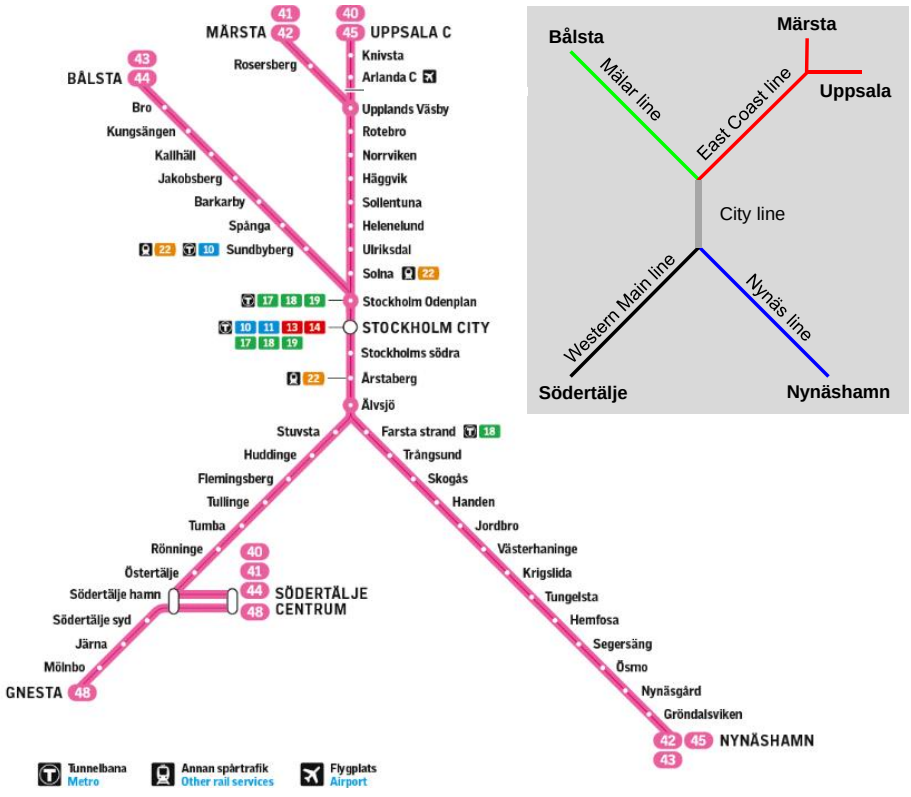


Figure 1 Network. Colour scheme for the branches shown in the small figure. The Gnesta line (48) is not included in the study.

All services on the branches operate City line and the capacity limitation in this line is a major condition for the timetable generation. The color scheme used for the branches in Figure 1 will be used throughout the article in order to increase readability.

The distribution of demand is shown in Figure 2. The average load per timetable cycle (30 minutes) in morning peak period, 07:20 – 08:50, is shown for all four branches. A train set has 750 seats and the diagram gives a first idea of the traffic needed to meet the demand. One important indication is that the demand corresponds to a system where not all traffic lines are full route lines. Hence, a major planning task will be to find feasible solutions for short turning lines.

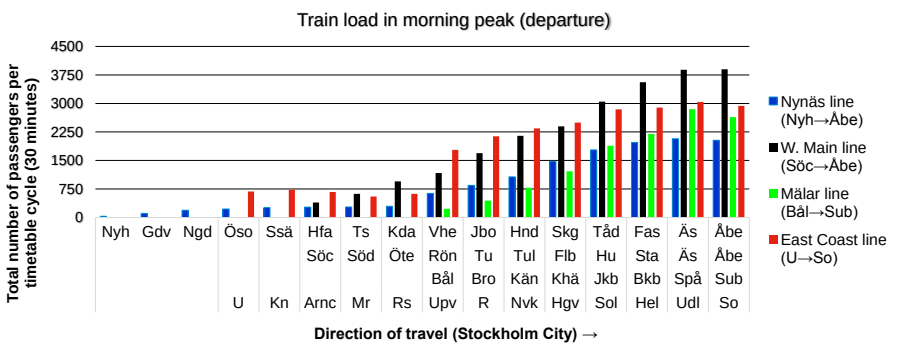


Figure 2 Demand during the morning peak.

1.2 Definitions

In the following section definitions for five concepts that will be frequently used throughout the article are discussed.

Branch: the infrastructure stretch from the separation junction to the farthest located termination station in the actual direction. The East Coast line is divided into two sub branches in Upplands Väsby, but modelled and referred to as one.

City line: is the common line section under Stockholm city. It imposes dependency between traffic lines to the four branches. This calls for a timetable coordination that is a natural starting point in timetable generation.

Timetable period: refers to a period, in minutes, with which the periodic timetable is repeated. The current timetable period is 30 minutes and during peak hour traffic the timetable cycle is repeated six times.

Line: a traffic relation between two termination stations that is operated by a service once per timetable period. Frequency of service between the two stations may be increased by adding more lines and coordinating them in time to get the desired frequency.

Termination pattern: a vector that is defined for each branch and shows the number of terminating lines per station, for example Södertälje C: 4 and Tumba: 2 implying that 4 out of 6 lines terminate in Södertälje C and 2 out of 6 lines in Tumba. The termination pattern gives rise to permutations of line sequences that are of great importance for congestion management.

1.3 Important timetable features

Four operational features of importance have been identified: punctuality, congestion management, timetable symmetry and resource efficiency (high utilisation of vehicles, infrastructure and train staff). It is worth noting that political priorities may also be implemented regarding the distribution of traffic resources in the network, stopping patterns, network extension etc.

Punctuality is highly commended by commuters. A high punctuality can be ensured through a thorough scheduling using balanced and optimized time supplements, buffer times and termination times. Consequently the termination times are in focus in this article, since they also impact the need for additional train sets and infrastructure.

Congestion management might be the most important factor in scheduling of commuter traffic since it has a direct impact on demand, overall customer satisfaction, resource efficiency and even punctuality. Congestion management is closely interconnected to short turning patterns and these two factors form the core when it comes to scheduling of commuter traffic in urban areas.

Symmetry is widely used throughout Europe as a way of achieving coordination between rail services (Liebchen (2004)). In a symmetric timetable the inbound direction is a reflection of the outbound as shown in Figure 3. Each traffic line is represented by a colour and line style (dashed for services on Nynäs line and solid for services on Western Main line). It can be seen that each line crosses itself on the so called symmetry times, i.e. minute 0, 15, 30, etc.

Since almost all other regular rail traffic in Sweden, such as long distance and regional traffic, is scheduled symmetrically it becomes a technical requirement also for the Stockholm commuter traffic in order to manage coordination of traffic on shared line sections. However, the symmetry has at least two major drawbacks for the commuter traffic:

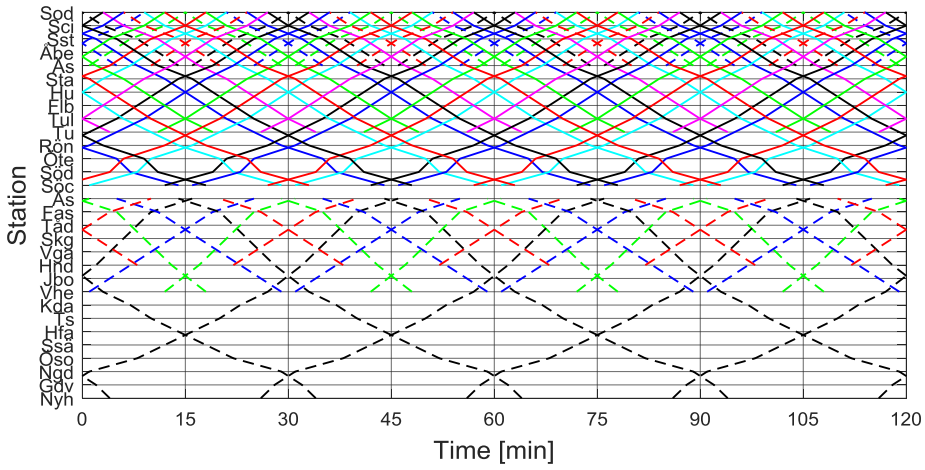


Figure 3 Timetable example showing symmetry. Southern half of the network shown.

- The termination pattern for in- and outbound directions cannot be chosen independently. This have rather severe, limiting effects on congestion management since the demand is different for the in- and outbound direction.
- The termination time, which is important for punctuality as well as resource efficiency, can be chosen less freely since a specific departure time requires a specific arrival time if the timetable is to be symmetric.

This implies that the coordination with other traffic impose a less efficient commuter operation as regards number of train sets and termination tracks (infrastructure) as well as congestion management and punctuality. This is previously shown by (Liebchen (2004)). However, this efficiency decrease might be less costly than a complete separation through additional quadruple line sections.

2 Finding feasible timetable solutions

A feasible timetable might be defined as a timetable that, for a given demand distribution in the network, has a low spread in passengers' loads on the traffic lines, allows enough termination times to reach a reasonable level of punctuality and efficient use of train sets and termination tracks.

As the evaluated system is limited and closed, a generic approach might be applied to find these timetable solutions. The method can be described by the following steps:

1. Establish a slot system with traffic lines in City line, the common link, to define available capacity.
2. Distribute available lines on the four branches, using demand data in Figure 2, and construct alternative branch patterns. These are referred to as line sequences. See Table 1 for two examples.
3. Create a nominal timetable between end stations of the branches according to the defined sequence. This step includes definition of stopping patterns, location of time supplements, buffer times etc.
4. Define termination patterns that are expected to be of interest. Existing as well as investigated termination stations might be listed for evaluation. Table 2 shows two alternative termination patterns for the two sequences.
5. For each branch:
    a. Permute the termination pattern in order to cover all permutations.
    b. For each permutation: adjust the nominal timetable by cutting off short turning lines according to the permuted termination patterns and distribute demand on the lines. Ortúzar and Willumsen (2011) share useful ideas of demand modelling.
    c. Compare load distribution on the traffic lines and select permutations with even passengers' load. One example, based on pattern 1 in sequence 1, is shown in Figure 4.
6. For each branch and selected permutation: calculate termination times and track usage in each termination station. Perform this for all rotation steps within a timetable cycle. The principle of rotation is shown in Figure 5. For a timetable period of 30 minutes the rotation gives 30 timetable variants for each permutation. Data for termination times and track usage are compiled for each permutation and rotation step.
7. Create complete network timetables through combinations of permutations on the four branches including rotation variants. The number of train sets needed,

additional termination tracks needed, expected termination punctuality and data for passenger load distribution is compiled for each timetable variant.

The number of timetables found depends heavily on the complexity in termination patterns and whether only symmetric solutions are accepted. For symmetric solutions a chosen permutation for a branch in one traffic direction defines the permutation also in the opposing direction for the same branch, since they are each other's reflections.

The following tables and figures illustrate the method used. Table 1 shows the slot system on City line with one train path every three minutes. This is followed by two alternative line sequences that divide traffic on the northern branches slightly differently. Table 2 shows two examples of termination patterns for sequence 1 and two for sequence 2. The pattern on Western Main line is simple, with either 4 or 3 out of six lines terminating in Södertälje (Söc), resulting in 15 and 20 permutations respectively for this branch.

Figure 4 shows the unique permutations for the two southern branches. The passengers' load on each line is affected by the line extension and the time distance to the preceding service, shown in the text box in upper left corner. Please note the difference between the most balanced permutation (leftmost) and the least balanced one (rightmost). A train set has 750 seats and a traffic line might be regarded as overloaded from a comfort perspective when the load reaches 1 000 passengers.

Figure 5 shows the principle of symmetry, meaning that a clockwise time shift in departure time imposes an equal counterclockwise shift in arrival time.

Table 1: Line sequences, two examples. ML: Mälar line, NL: Nynäs line, ECL: East Coast line, WML: Western Main line. * Indicates express lines.

Line	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Dep time S bound	00	03	06	09	12	15	18	21	24	27
Seq 1										
North	ML	ECL	ML	ML	ECL	ML	ECL	ECL	ML	ECL
South	NL*	WML	WML	NL	WML	NL*	WML	WML	NL	WML
Seq 2										
North	ML	ECL	ECL	ML	ECL	ML	ECL	ECL	ML	ECL
South	NL*	WML	WML	NL	WML	NL*	WML	WML	NL	WML

Table 2: Termination patterns. Two examples per sequence, 10 lines per timetable period.

		Southern lines					***Northern lines***					
		Western Main line		**Nynäs line**			**Mälar line**			**East Coast line**		
Seq	**Pattern**	Söc	Tu	Nyh	Vhe	Hnd	Khä	Kän	Bål	Upv	Mr	U
1	1	4	2	1	2	1	2	1	2	1	2	2
1	2	3	3	1	3	-	2	2	1	-	3	2
2	1	4	2	1	2	1	2	0	2	2	2	2
2	2	3	3	1	3	-	1	2	1	1	3	2

The procedure discussed above might be applied to screen for feasible timetable solutions. One example is presented in the following section. The procedure has several similarities with the TVEM model that is described in Lindfeldt (2010). Major differences are that demand, congestion, terminations and vehicle rotations are included in the current model.

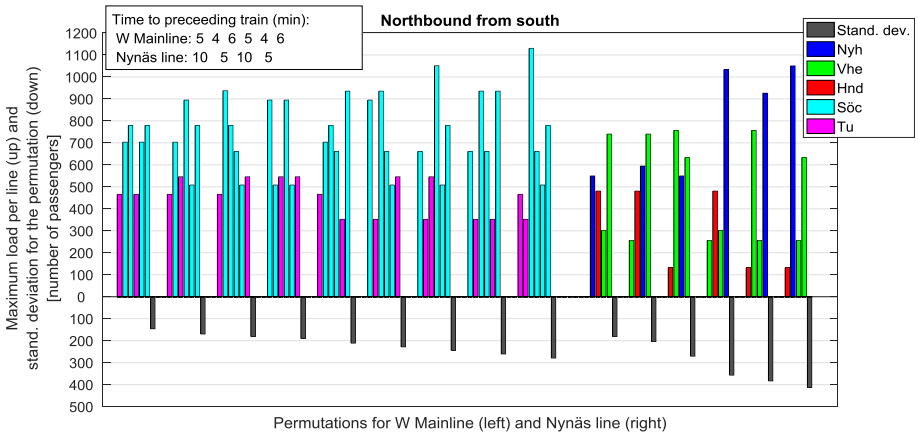


Figure 4 Line permutations and corresponding passenger distributions on lines. Sorted according to ascending standard deviation in passengers' load. Only unique permutations are shown for space reasons.

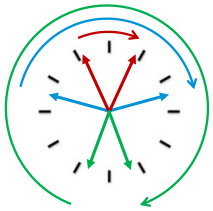


Figure 5 Principle of rotation with symmetry.

3 Application supported by an example

The first pattern in sequence 1 in Table 1 is assumed to meet demand efficiently and serves as a good example. This pattern corresponds to a 25% traffic increase, compared to the operated traffic in 2019. As most permutations have a high variance in passengers' load, (Figure 4) it is feasible to choose only permutations with a load standard deviation lower than 200 passengers. After that selection, only 36 000 timetables remain. All of these constitutes a reasonable passengers' load on all traffic lines.

Despite the fact that the total operated time and distance is exactly the same in all these timetables, termination times differ. This results in different number of train sets needed. Moreover, the existing number of termination tracks might not be sufficient, implying a lack of infrastructure in some locations. Distributions for these resources are shown in Figure 6. The diagrams indicate that a minimum number of 62 train sets is needed for the traffic and at least two additional termination tracks have to be constructed.

Termination punctuality values can be estimated through combination of scheduled termination time, minimum (technical) termination time and historic delay distributions for

arrivals at the different stations.

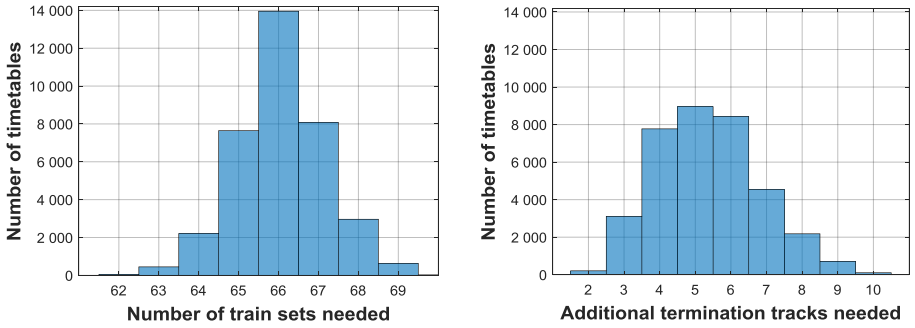


Figure 6 Distribution for number of train sets and additional termination tracks needed.

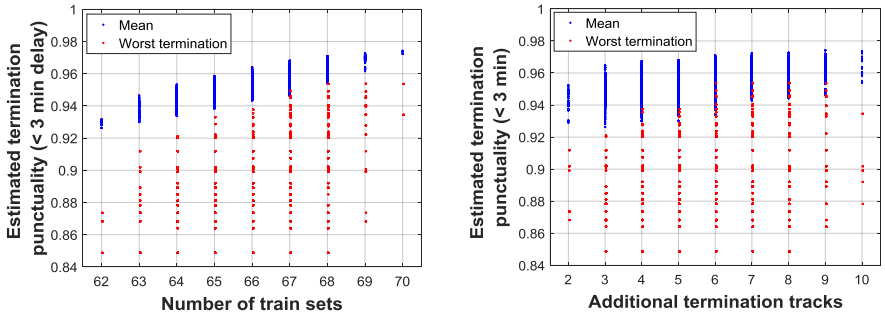


Figure 7 Estimated termination punctuality for different number of train sets and additional termination tracks. Blue: mean for all 20 terminations and red: termination with lowest punctuality.

Figure 7 shows punctuality statistics. All 36 000 timetables are represented by a blue and a red marker, indicating mean termination punctuality and lowest termination punctuality respectively.

As the diagram indicates, additional train sets and/or termination tracks provide a better recovery and higher punctuality through longer termination times. This is a rather unprecise way to estimate punctuality, as the real arrival delays are influenced by the timetable solution and other factors such as recovery in the other end of the line. However, it may be a suitable way to sort out timetable solutions with a distribution of termination time that will ensure an acceptable level of termination recovery.

Further analysis calls for additional filtration, since it is unreasonable to invest in extensive number of train sets and/or termination tracks. Therefore, in the final evaluation only timetables that requires 62-64 train sets are analyzed. These numbers of train sets correspond to 2-5 additional termination tracks.

It is not enough to know the number of lacking termination tracks. It is also important that their locations are specified as well, if the infrastructure is going to be completed in

order to meet the analyzed traffic increase.

If the remaining feasible timetables are compiled according to need of additional termination capacity a diagram like the one in Figure 8 can be drawn. As demonstrated in the diagram one additional termination track ought to be constructed in Bålsta and Uppsala respectively. Such an extension would be enough to manage about 15% of the feasible timetables. A further analysis of the corresponding timetables has to be performed in order to assess whether this portion is enough to manage future changes in operation. Complementary flexibility in the choice of timetable requires more termination tracks.

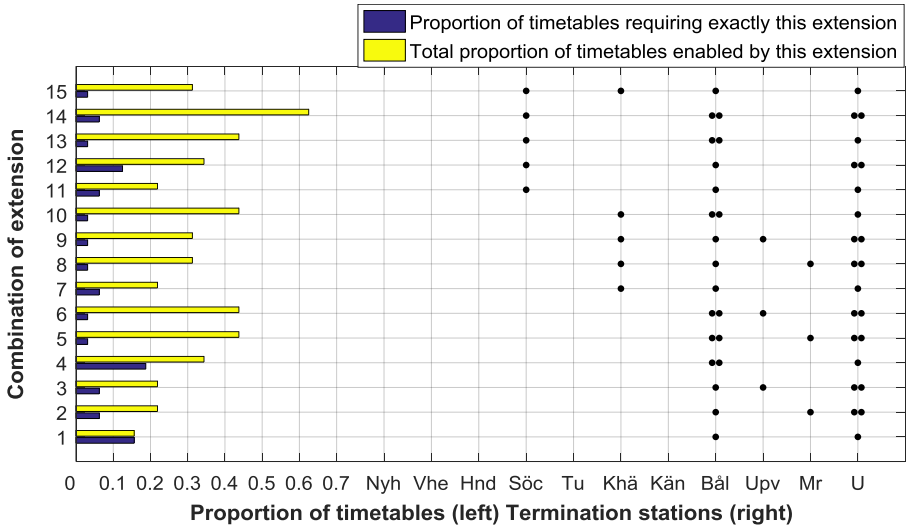


Figure 8 Alternative combinations of additional termination tracks needed to enable the evaluated traffic and the proportion of feasible timetables that they enable. A double dot indicates need for two additional tracks.

4 Conclusions

A generic method for timetable screening of a four branch commuter rail system has been discussed. The method is based on congestion management and combinatorics of short turning traffic lines (termination patterns). The overreaching objective is to explore timetable solutions that are efficient, namely, congestion balanced, that require a limited number of train sets and additional termination tracks, but still have termination times that are long enough and well distributed to reach a reasonable recovery level and punctuality.

The number of available timetable solutions is limited by the common line section, City line, where a slot system has to be applied in order to manage capacity. As in most other public transport systems, short turning is an efficient way to adjust traffic supply to demand. The short turning lines constitute termination patterns and permutations of these.

The permutations influence passenger distribution and congestion and has to be selected with care. A network timetable can be constructed through a combination of permutations for the four branches. Each combination can be rotated through the timetable period, in this

case 30 minutes. The choice of permutations and degree of rotation determine the termination times, which in turn define important features such as: recovery/punctuality, number of train sets needed and number of termination tracks needed.

Special attention ought to be paid to the mode of rotation. Coordination with other rail traffic on shared line sections requires the timetable to be symmetric. The symmetry implies that the permutation can only be chosen freely in one direction per branch, since the permutation for the opposing direction has to be a reflection of the first one. This fact strongly limits the number of available timetable solutions, the possibility to reach an efficient congestion management in both traffic directions and to limit the number of train sets and/or termination tracks needed. As such the coordination with other traffic impose a less efficient commuter operation. Nevertheless, it might be less costly than a complete separation through supplementary quadruple line sections.

Further evaluation studies are planned to be carried out to analyse the benefits of asymmetric timetables. It would also be interesting to delve into available capacity for long distance and regional services. The current model can easily be updated to cover timetable solutions for asymmetric timetables as well as other traffic.

Acknowledgements

The author expresses his thanks the Stockholm Public Transport (SLL) for its encouragements during the development of the model. The author also acknowledges the executive board, administration and the planning teams of MTR Pendeltågen. Thanks also to the National Transport Administration (Trafikverket) that manages the planning of future rail capacity as well as the national timetable coordination.

References

Lindfeldt, O., 2017. "Designing the future commuter traffic through central Stockholm". In: *Proceedings of The 7th International Seminar on Railway Operations Modelling and Analysis (RailLille 2017)*, Lille, France.

Lindfeldt, O., 2010. "Railway operation analysis - Evaluation of quality, infrastructure and timetable on single and double-track lines with analytical models and simulation". PhD thesis KTH, Stockholm.

Liebchen, C., 2004. "Symmetry for Periodic Railway Timetables", Electronic Notes in Theoretical Computer Science, vol. 92, pp. 34–51.

Ortúzar, J., and Willumsen, L. G., 2011. "Modelling Transport" (4th ed.), Wiley, Chichester.

Tillväxt- och regionplaneförvaltningen, SLL, 2017. "Strukturanalyser för Stockholms län och östra Mellansverige år 2050 - konsekvenser av strukturbild år 2030–2050 - Underlag till utställningshandling", Stockholm. (In Swedish)

An Iterative Approach for Profit-Oriented Railway Line Planning

Di Liu [a,b,c,1], Pieter Vansteenwegen [c,2], Gongyuan Lu [a,b,3], Qiyuan Peng [a,b,4]
[a] School of Transportation and Logistics, Southwest Jiaotong University
610031, Chengdu, Sichuan, China
[b] National United Engineering Laboratory of Integrated and Intelligent Transportation, Southwest Jiaotong University
610031, Chengdu, Sichuan, China
[c] KU Leuven Mobility Research Center-CIB, KU Leuven
Celestijnenlaan 300-box 2422, 3001, Leuven, Belgium
[1] E-mail: diliu5545@gmail.com
[2] E-mail: pieter.vansteenwegen@kuleuven.be
[3] E-mail: lugongyuan@home.swjtu.edu.cn
[4] E-mail: qiyuan-peng@home.swjtu.edu.cn

Abstract
With the rapid development of the Chinese high-speed railway network, more and more railway lines are becoming oversaturated, leading to inefficient operations and reducing the service quality. To improve the network's performance, this paper proposes a profit-oriented line planning model for optimizing both the operational costs and passenger travel times. Due to the complexity of the problem, an iterative approach is designed to solve the problem efficiently. Two case studies are implemented to verify the performance of the approach. The results of the case studies show that the proposed approach can improve the profit while balancing the operational cost and the passenger travel time, with reasonable computation times. For the resulting line plan, the optimal passenger routes, including transfers when necessary, are also determined. The proposed iterative approach increases the profit of an initial solution with on average more than 20% for a medium and a large-scale network.

Keywords
High speed railway network, Line planning problem, Iterative approach

1 Introduction

The Chinese high-speed railway (HSR) network has developed rapidly during the past ten years. Currently, the 4 "vertical" and 4 "horizontal" tracks (4V4H) of the HSR network are the backbone to connect the major cities in China. The four vertical high-speed railway lines are Beijing-Harbin (1800 km), Beijing-Shanghai (1318 km), Beijing-Hongkong (2383) and Hangzhou-Shenzhen (1449 km). The four horizontal high-speed railway lines are Qingdao-Taiyuan (940 km), Xuzhou-Lanzhou (1434 km), Chengdu-Shanghai (2066 km) and Kunming-Shanghai (2056 km). The majority of these HSR lines are only used for providing passenger transportation services. At the end of 2017, the length of the total high-speed railway was more than 25 thousand kilometres, which accounts for more than 60% of HSR

lines in the world. The amount of passenger volume transported by HSR trains is 56.8% of the total railway passenger demand in China (China National Bureau of Statistics 2018). With the rapid expansion of the current HSR network, the total length of the HSR lines will reach 30 thousand kilometres and cover more than 80% of the cities in China by 2020. Figure 1 shows the 4V4H network and its associated HSR lines.

Comparing to the HSR lines in Europe and Japan, the large-scale HSR lines in China are different according to its actual operation practice. There are a large number of long-distance HSR trains operating per day to serve as many passengers travel demand as possible. However, the average passenger travel distance is usually much shorter than the HSR line distance. For instance, the average travel distance of the two main HSR lines is about 558 km (Beijing-Guangzhou HSR) and 621 km (Beijing-Shanghai HSR), while the distance of the whole HSR lines are 2281km and 1318 km respectively (Fu et al. (2015)). This may lead to the inefficient use of railway resources such as train capacity and HSR line capacity.

Since the Chinese HSR lines were constructed gradually, single lines were firstly designed between selected stations, and then merging and diverging lines were added to the line plan. Therefore, the previous line plan was designed without considering the network as a whole. After the basic 4V4H HSR network is formed, the operation plans should be made on the consideration of passenger demand features based on the whole HSR network. Because of its large-scale size, high transportation demand and network capacity limitations, it is required to develop efficient operation plans to improve the whole network's operational performances. Line planning is one of the crucial planning stages when designing a railway service.

To improve service quality and reduce the operational cost, this paper proposes an iterative approach for optimizing both operational cost and passenger travel time. This paper

Figure 1. 4 vertical and 4 horizontal HSR network and its associated HSR lines

aims to design a network line plan for the 4V4H HSR network and to determine the frequency of the lines, optimizing operational cost and passenger travel time, while considering transfers when necessary. In order to obtain this, a profit-oriented objective function is introduced. The detailed contributions of this paper are summarized as follows.

(i) A profit-oriented objective is proposed that uses a time value parameter to consider the travel time in the ticket price.

(ii) An iterative algorithm is designed to solve the line planning problem.

(iii) Different local search improvements are considered to generate neighborhood solutions, such as extending a line, reducing a line, inserting a line and removing a line.

(iv) Fast and heuristic evaluation methods are designed to choose the most promising neighborhood solution in order to obtain a better line plan efficiently.

(v) The passenger route choice is optimized by assigning each passenger to its shortest path through the network.

The remainder of this paper is organized as follows. Section 2 gives a brief literature review of the line planning problem. In Section 3, the profit-oriented line planning problem and the proposed mixed integer linear programming model are presented. Section 4 shows an iterative solution approach in detail. Section 5 shows the description and evaluation of the numerical experiments. In Section 6, the major conclusions and further studies are presented.

2 Literature Review

Recently, several studies have addressed the network design or line planning problem (LPP) integrated with traffic assignment or passenger assignment and focusing on the operational cost and passenger preferences (Karbstein (2014); Borndörfer and Karbstein (2012); Friedrich et al. (2017); Nachtigall and Jerosch (2008); Fu et al. (2015); Rosalia (2017); Borndorfer et al. (2007)). In general, the available infrastructure and passenger demand between each origin-destination (OD) pair are considered as given input data of the LPP. The LPP aims to determine the appropriate set of lines, each serving a sequence of stops, together with its associated frequencies, so that the total passenger demand is satisfied directly or with a limited number of transfers.

Past studies typically differ in how they consider the interests of passengers and operator costs. E.g., Friedrich et al. (2017) investigate a cost-oriented line planning model with passenger assignment evaluation. In this case, the line planning solutions focus on the operational cost rather than service quality, such as travel time, transfers and passenger waiting times, which may lead to a reduction of the demand and lower revenues. With this concern, Nachtigall and Jerosch (2008) integrated the cost-oriented objective and customer-oriented objective into a single model by transforming one of them into a constraint. Instead of converting one objective into the constraint, Borndörfer et al. (2007) used a weighted sum of cost-oriented and customer-oriented objectives while the lines are generated dynamically with flexible passenger paths. In addition, Fu et al. (2015) developed a bi-level programming approach to optimize the line planning and passenger assignment sequentially.

Rosalia (2017) developed a model to optimize the operational cost and travel time iteratively on a city road network, which solves the minimum operational cost line plan and then minimizes the travel time based on that line plan

In large-scale networks, transfers are unavoidable because of the infrastructure capacity limitation and the operational costs of operating all direct connections. Moreover, if long-distance trains stop at each intermediate station, the resulting lower speed will have a negative effect on the HSR network capacity and also reduce the attractiveness of the HSR trains. In order to consider transfers, Borndörfer and Karbstein (2012) presented a direct connection approach to integrate line planning and passenger routing optimization by encouraging the direct connection and penalizing the transfers. Furthermore, Karbstein (2014) introduced a new model for integrated line planning and the passenger routing problem by involving a variant of the 2-terminal Steiner connectivity problem as the pricing problem and applying it to handle the transfers in LPP. However, for some passengers it may be beneficial to have transfers rather than spending much more time on a train with a direct connection. Therefore, convenient and time-saving transfers should also be considered. Readers interested in LPP are referred to a review conducted by Schöbel (2012). None of these works are profit-oriented and integrate operator costs and passenger travel time in a single objective by making the revenues dependent on the passenger travel time.

Although more and more researches are focusing on the LPP in recent years, the trade-off between long distance direct connections and transfer services have not been studied greatly. Currently, the passenger flows on the Chinese HSR network have significant characteristics. According to the current timetable, there are 595 stations on the HSR network. While the number of direct connections accounts for around 9% of the total OD pairs according to current line plan (Liu and Li (2018)). Therefore, the majority of the passengers need to take transfers. An explicit decision should be made on which passengers should be able to travel directly and how to facilitate transfers for the remaining passengers.

In this paper, instead of choosing between a cost-oriented and a customer-oriented objective, we propose a profit-oriented line planning model which maximizes the ticket price income minus the operational cost. The ticket price (and thus the operator revenues) are reduced when passengers need a transfer or a detour and have no direct train from origin to destination. Moreover, the operational costs consider fixed and length dependent costs for operating the different lines.

3 Profit-oriented Line Planning

The profit-oriented line planning problem presented in this paper focuses on making a trade-off between a cost-oriented objective related to the number of trains operated to meet all the passenger demand and the customer-oriented objective by minimizing the travel inconvenience. This travel inconvenience is defined here as additional travel time compared to the travel time of having a direct connection along the shortest path in the infrastructure network. The following assumptions are made throughout this paper.

- Assumptions:

  (i) Stopping pattern: Since only major stations are considered as nodes in the network, the stopping pattern of the line plan is an all-stop pattern for the major stations. The passenger demand of small stations can be assigned to the major stations in a pre-calculation phase. After designing the lines, a stopping pattern optimization can be used to determine exactly which (small and large) stations

will be served by each line.

(ii) Demand: All demand in the network is served with at most two transfers. In this network of limited size (only considering the major stations) two transfers should be more than enough.

(iii) Passenger route choice: Passengers will always choose the shortest travel time path no matter what the price of the path is. Passengers of the high-speed railway normally pay more attention to the travel time rather than the ticket price.

(iv) Train type: Two train types are considered, a single train-set with 500 seats and a double train set with 1000 seats. For now, there is only one speed of train considered on the network, i.e., the 300-350 km/h high-speed train. Currently, the operation speed on this network is set to 250km/h for some safety reasons. With the development of the high-speed railway in China, there is a tendency to operate higher speed trains in the future. In addition, the train speed is just a parameter in our model, which could easily be changed to 250km/h. Moreover, we do not impose a maximum number of trains for a certain track yet. Including track capacity and trains with different speed on the network is considered as future work.

(v) Line attributes: There is no limitation on the line length considered and lines can start and end in any station.

(vi) Passenger demand is considered symmetrical and therefore each line is assumed to operate in both directions.

- This is considered as input:

(i) Passenger OD matrix: The number of passengers traveling between any two stations is given in the symmetrical OD matrix. The passenger OD matrix represents the daily passenger demand.

(ii) HSR network topology: The available stations (nodes) and tracks (links) are fixed and the distance of each link between two stations is known.

Variables and Notations

In this study, we use the following variables and notations. The physical network topology is considered as the undirected graph $N = (V, E)$. The node set $V = \{1,2,3, \dots, n\}$ represents the stations and the edge set $E = \{e, e \epsilon V \times V\}$ represents the connections of two stations in the network. Before solving the LPP, we introduce the train service network (TSN). In order to take transfer times into account and to calculate the approximate travel time, the TSN is constructed to depict the itineraries of passengers (Fu et al. (2015)). This is also called the Change & Go network in Schöbel and Scholl (2006).

D a set that represents the passenger demand of all the OD pairs.

Inc the operational income.
Cos the operational cost.
L_{cur} the current line plan.
C^{Fix} the fixed cost for operating a line with frequency one.
C^{Var} the variable cost per line per kilometre.
d_{v_i,v_j} the number of passengers want to travel from station v_i to station v_j.
$T^P_{v_i,v_j}$ the length of the shortest travel time of each OD pair $(v_i, v_j)$with respect to the physical network independent of the line plan.
T_v the time value (the ticket price per unit of time) to convert the passenger travel time into the ticket price by multiplying with the riding time and dwelling time.
$StaInc$ the ideal income, if each passenger would have a direct train on his/her shortest path: i.e., $\sum_{v_i,v_j \in V} T^P_{v_i,v_j} * T_v * d_{v_i,v_j}$.
$T^S_{v_i,v_j}$ the length of the shortest travel time path of each OD pair (v_i, v_j) on the TSN.
T^{Pen}_v the penalty time value: a fixed value for each transfer on a path and per unit of time for the detours.
k_l the length of line l (in kilometres).
f_l the frequency of line l.

Objective

Instead of using the weighted sum of a cost-oriented and a customer-oriented objective, this model uses a profit-oriented objective which considers the operational cost and the passenger total travel time represented by the operational income. The operational cost is composed of a fixed cost per line per train and a variable cost depending on the length of the line. By introducing the time value, the passenger travel time can be converted into operational income. Thus, the operational income can be formulated as the passenger total travel time multiplied with the time value and minus the transfer and detour penalties. The goal is to maximize the profit.

$$\max Z = Inc - Cos. \tag{1}$$

$$Inc = StaInc - \sum_{v_i,v_j \in V} (T^S_{v_i,v_j} - T^P_{v_i,v_j}) * T^{Pen}_v * d_{v_i,v_j} \tag{2}$$

$$Cos = \sum_{l \in L_{cur}} (C^{Fix} + C^{Var} * k_l) * f_l. \tag{3}$$

Equation (2) gives the specific composition of the operational income. The left side of the minus is the ideal income calculated by the ideal shortest travel time of the direct connection of each OD pair. The right side of the minus is the penalty fee for transfers and detours by using the results of the comparison of actual travel time and the ideal travel time multiply with the penalty time value and the associated passenger demand. The ticket prices reduction is determined in such a way that it (partly) compensates the discomfort or lost time of having to travel longer (than the ideal shortest path). This also implies that passengers will never prefer to travel even longer because it would be cheaper. The operational cost is presented as equation (3), which is related to the number of lines and associated frequencies.

Constraints
The constraints used in the iterative algorithm are mainly about satisfying passenger demand and capacity limitations of the trains: All passenger OD pairs should be served with at most two transfers. Moreover, for each arc of the network, the summed capacity of all the lines (each with a certain frequency and vehicle type) on that arc should be sufficient to meet the passenger demand on that arc.

The solution is represented by a set of lines, each associated with a certain frequency and a vehicle type. A line consists of a sequence of nodes.

4 An Iterative Approach

The TSN is constructed based on the given line plan. A small example is introduced to illustrate the construction of TSN. In Figure 2, the topology of the physical railway tracks and the line plan are given. According to the line plan, a TSN is built as shown in Figure 3. In the TSN, the passenger routes of different OD pairs can be seen as the combinations of several types of arcs with the associated nodes. For example, the dotted line depicts the passenger from station B to station D take the sequence of boarding arc, riding arc, dwelling arc, riding arc, transfer arc, riding arc and alighting arc to reach their destination.

This paper presents an iterative framework for solving the network LPP in two stages. With the idea of minimizing the operational cost and saving the passenger travel time, an initial line plan is generated with a constructive heuristic in the first stage. Then the initial

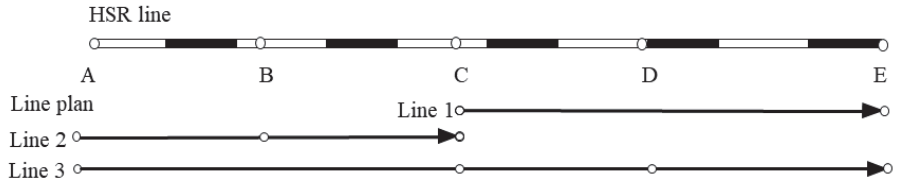


Figure 2: An example of railway topology and given line plan

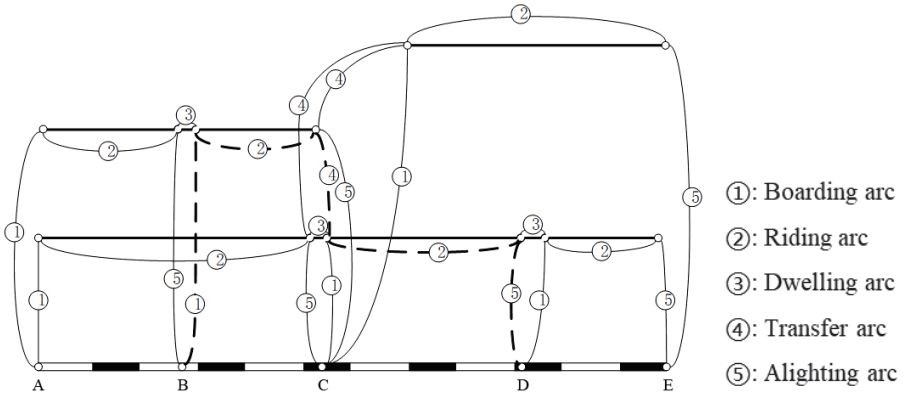


Figure 3: A train service network (TSN) constructed based on the information in Figure 2

line plan is given as the input for the second stage, where an iterative approach is performed to optimize the initial line plan. Each iteration of the second stage starts with the results of the passenger assignment process and determining the most appropriate frequency and vehicle size for each line. Then, some indicators are calculated to evaluate the current line plan and to determine the modifications required to determine a new and better line plan. The second stage is repeated until the algorithm reaches the stopping criterion. The different steps of the approach are now explained one by one in more detail.

Initial Line Plan Generation

The basic idea of the initial line plan generation is to select those lines that serve directly as much passenger demand as possible. First, a line pool $L^P_{v_i,v_j}$ is constructed containing the shortest path between every OD pair in the physical HSR network. Then, one by one, from that pool those lines are selected that serve the most passengers directly, not only from the starting towards the ending station of the line, but also from and towards all stations in between on that line. As in many research papers on railway line planning (Goerigk, Schmidt (2017), Yang et al (2016)), we assume that passengers travel according to their shortest path.

During this evaluation, only OD pairs that are not served yet directly by previous lines are considered. The selection of lines ends as soon as all nodes are covered. After that, the passenger assignment process is performed to check whether the transfer constraint (at most two transfers for each passenger) is satisfied. If not, the passenger OD pairs that break the transfer constraints are selected to be served directly and the corresponding lines from the line pool are added to the initial line plan. The outline of the process described above is shown in Algorithm 1.

Algorithm 1: Heuristic algorithm for calculating the initial line plan

Calculate the shortest path set S of each OD $(v_i, v_j) \epsilon D$ w.r.t track lengths
for path $l^P_{v_i,v_j} \epsilon L^P_{v_i,v_j}$ **do**
  for $(v_i, v_j) \epsilon D$ **do**
    - assign direct passengers on $l^P_{v_i,v_j}$
  end for
  - calculate the number of direct passengers assigned on the path $l^P_{v_i,v_j}$
end for
repeat
  - select the path $l^P_{v_i,v_j}$ with the most direct passengers into the initial line plan
until all the nodes in the network are covered.
Checking the transfer constraint. The set of initial line plan L^0 is obtained

Passenger Assignment and Frequency Setting

We assume that the passengers look for the shortest path in travel time among all the possible paths between their origin and destination in the TSN. Firstly, the TSN of the initial line plan is constructed. The shortest paths are found using a modified Floyd algorithm (Floyd (1962)), which takes into account the transfer constraint by counting the number of transfers of the possible shortest paths and choose those paths with less than two transfers. Due to the TSN used, possible transfers are also considered when determining the shortest path. The line plan generally contains (partially) overlapping lines, for some OD pairs, there

may exist several paths with the same shortest travel time. As the consequence, the passenger route choice during the passenger assignment process is assigned randomly based on these paths.

In order to determine the frequency of each line, the number of passengers assigned to each part (link between two consecutive stations) of a line is considered. The part with the highest number of passengers determines the frequency and vehicle size assigned to the line. After calculating the frequencies of lines, the cost of operating all the lines can be obtained.

Line Plan Evaluation and Modification

In order to improve the line plan, four modification methods are considered, namely, reducing a line (*Reduction*), extending a line (*Extension*), removing a line (*Removal*) and inserting a line (*Insertion*). Each type of modification leads to a neighbourhood of possible line plans. The Reduction neighbourhood of a current solution contains all line plans where one terminal node of one line is removed. The Extension neighbourhood contains all line plans where one node, adjacent to the terminal node in the physical network, is added to one of the lines. When it comes to Insertion, all lines corresponding to OD-pairs without a direct connection in the current line plan, are considered. For Removal, the neighbourhood contains all line plans where a line of the current line plan is removed. In each neighbourhood, only feasible line plans are considered.

The detailed evaluation and modification process of the line plan is illustrated in Figure 4 and is now explained. Given the results of passenger assignment, two evaluation calculations are applied on the current line plan considering Reduction and Extension a line. Here we consider all neighbourhood solutions implicitly by heuristically evaluating how promising the neighbourhood solutions are.

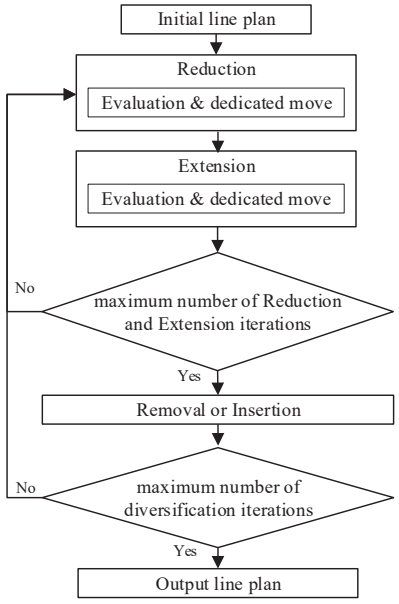


Figure 4: The iterative approach framework for line planning optimization

When considering Reduction, the load factors of terminal edges of each line are calculated as the evaluation indicator of the current line plan. The load factor is the actual passenger volume on the edge of a line divided by the frequency of the corresponding line. Then the terminal edge with the lowest load factor is selected to be reduced. After Reduction, Extension is considered by comparing the number of indirect passengers that can be transported directly through the extended edges. The available extended edge that can provide the most additional direct connections for passengers is selected. After each modification, the passengers need to be assigned to the TSN in order to evaluate the total profit (ticket sales income minus line operating costs). Only solutions that actually improve the profit are accepted and implemented. Reduction and Extension are executed until a predetermined number of iterations. This number is discussed later in Section 5. It was necessary to predetermine such a number because continuing until a local optimum is reached turns out too time consuming. It is well-known that the passenger assignment process, unavoidable when evaluating the profit of a new line plan, is computationally very expensive.

In order to diversify the algorithm, two disturbances are implemented as well: Removal and Insertion. One of both is selected randomly and the line to remove or insert is also selected randomly. The solution of the disturbance is accepted whether the solution is better or worse than the current solution. The number of diversification iterations is also fixed beforehand as a stopping criterion for the algorithm.

The previous process is presented in Algorithm 2. The correlated notations are as follows.

Z_{cur} the profit of the current line plan.
Z_{nei} the profit of the neighbourhood solution based on the current line plan.
L_{nei} the selected neighbourhood line plan.
E_{red} the neighbourhood solution set of reducing a line.
E_{ext} the neighbourhood solution set of extending a line.
E_{rem} the neighbourhood solution set of removing a line.
E_{ins} the neighbourhood solution set of inserting a line.
A_{undir} the set of OD pairs with passengers requiring at least one transfer.
TSN_{cur} the train service network based on the current line plan.
TSN_{nei} the train service network based on the neighbourhood solution of the current line plan.

Algorithm 2: Iterative evaluation and modification of line plan

Repeat
  Repeat
    Reduction method
      Select the e_{red} with the lowest load factor in E_{red}
      Modify the L_{cur} and TSN_{cur} to obtain L_{nei} and TSN_{nei}
      if network connectivity check = true **do**
        Calculate passenger assignment and the profit of L_{nei}
        if $Z_{nei} > Z_{cur}$ **do**
          - $L_{cur} \leftarrow L_{nei}, TSN_{cur} \leftarrow TSN_{nei}, Z_{cur} \leftarrow Z_{nei}$
          - update E_{red} and E_{ext}
        else do
          - $L_{nei} \leftarrow L_{cur}, TSN_{nei} \leftarrow TSN_{cur}, Z_{nei} \leftarrow Z_{cur}$
          - delete this e_{red} from set E_{red}
      else do
        - roll back to L_{cur} and TSN_{cur}
    Extension method
      Select e_{ext} with the highest number of increased passengers in E_{ext}
      Modify the L_{cur} and TSN_{cur} to obtain L_{nei} and TSN_{nei}
      Calculate passenger assignment and the profit of L_{nei}
        if $Z_{nei} > Z_{cur}$ **do**
          - $L_{cur} \leftarrow L_{nei}, TSN_{cur} \leftarrow TSN_{nei}, Z_{cur} \leftarrow Z_{nei}$
          - update E_{ext} and E_{red}
        else do
          - $L_{nei} \leftarrow L_{cur}, TSN_{nei} \leftarrow TSN_{cur}, Z_{nei} \leftarrow Z_{cur}$
          - delete this e_{ext} from set E_{ext}
  until maximum number of Reduction and Extension iterations
  Disturb
    Index = random (0,1)
    if index = 0 **do Removal**
      - select a line l_b randomly from L_{cur} and remove it
      - modify the TSN_{cur}as TSN_{nei}
        if network connectivity check = true **do**
          - calculate passenger assignment and the profit based on L_{nei}
          - $L_{cur} \leftarrow L_{nei}, TSN_{cur} \leftarrow TSN_{nei}, Z_{cur} \leftarrow Z_{nei}$
          - turn to step 2
        else do
          - $L_{nei} \leftarrow L_{cur}, TSN_{nei} \leftarrow TSN_{cur}, Z_{nei} \leftarrow Z_{cur}$
          - turn to step 27 and taboo l
    else do Insertion
      - calculate A_{undir}
      - randomly select an OD pair from A_{undir}. Add its shortest path as line l
      - modify the TSN_{cur} as TSN_{nei}
      - calculate passenger assignment and the profit based on L_{nei}
      - accept the L_{nei} as L_{cur} and TSN_{cur} as TSN_{nei}
      - turn to step 2
until maximum number of diversification iterations
Output the line plan and objective profit

5 Experimental Results

The experimental results of implementing the above iterative approach are presented in this section. The algorithm is implemented in C# and runs on an Intel i7 2.81GHz with 24GB RAM in the environment of Microsoft Win10. The input contains are the HSR network infrastructure, a fixed passenger OD matrix, link distance and link travel time (based on the single average speed considered). In order to show the performance and effectiveness of the proposed dedicated modification methods, we compare it with random modification methods without heuristic evaluation of the current line plan in order to determine the most promising Reduction and Extension (the random method is listed in Algorithm 3 in Appendix). Three passenger demand scenarios are considered corresponding to two networks, i.e., medium-scale network and 4V4H network, which contain 26 nodes (676 OD pairs) and 35 nodes (1225 OD pairs) respectively. The medium-scale network considers a part of the 4V4H network.

The specific passenger demand over the network is hard to obtain due to the confidentiality of the China Railway company. Therefore, the passenger demand used in this research is generated randomly. Additionally, other parameters used in the numerical experiments are shown in Table 1.

Table 1: Parameters setting

Name	Value	Unit
Train speed	300	km/h
Transfer time penalty	30	min
Stopping time	3	min
Ticket rate	0.5	CNY/km
Time value	2.5	CNY/ person, min
Penalty time value	0.55	CNY/min
Fixed cost of different train types	15000/10500	CNY/train
Variable cost of different train types	150/105	CNY/km
Train capacity of different train types	1000/500	seats/train

We assume that there are two types of train capacities, namely the doubled train and the single train, corresponding to different fixed cost and variable cost. The cost values of the single train are 0.7 times of the doubled train. The time value of Table 1 is computed as ticket rate multiply train speed, i.e., 0.5 * 300 / 60 = 2.5 CNY per person per minute. We assume that the average income of citizen is 33 CNY/hour. Thus, the penalty time value is 33/60=0.55 CNY/min. A small example is given in Figure 5 to show how these parameters work in the profit calculation. The example line plan is shown in Figure 6.

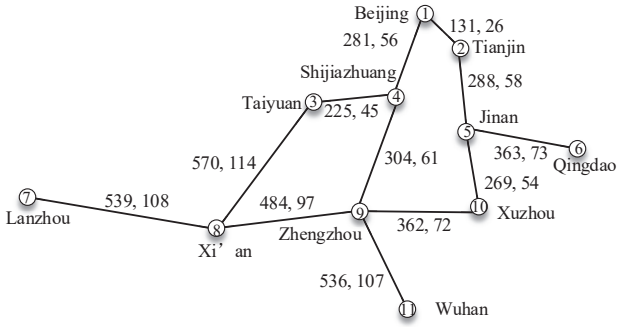


Figure 5. Small example for presenting the profit calculation

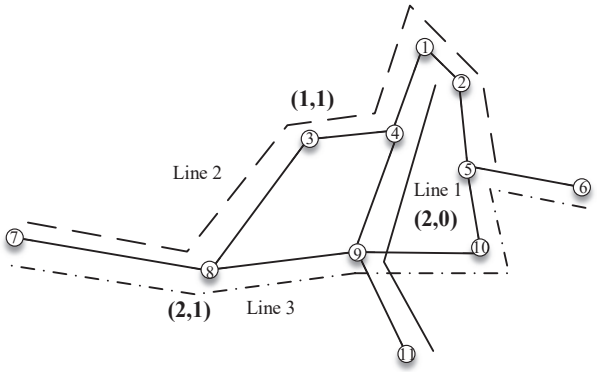


Figure 6. Example line plan of small network

In Figure 5, the first number besides every link indicates the distance between two nodes and the second number gives the corresponding travel time (since all trains have the same speed). In Figure 6, the first number between the brackets is the number of doubled trains (large size) on that line and the second number is the number of single trains. The operational cost of line 1 is 15000+150*(281+304+536)) * 2=366,330. After computing the operational cost of line 2 and line 3, the total operational cost equals to 1,767,855.

The operational income is calculated as the ideal income minus the penalty caused by transfers and detours. We assume here that the passenger demand of each OD pair is 100. The ideal income is the price that all the passenger OD pairs would pay when they are served by direct connections on their shortest paths. For instance, the ideal income of passengers from node 1 to node 5 equals the shortest travel time multiplied with the time value and the passenger demand between node 1 and node 5, i.e., (26+3+58) * 2.5 * 100 = 21,750. According to the ideal income calculation method of node 1 to 5, the total ideal income of all the passenger OD pairs is obtained as 2,435,250. When computing the penalty fee of transfers or detours, the penalty time is calculated as the actual travel time minus the shortest travel time with respect to the physical network. For example, the penalty fee (actually a reduction in the ticket price) for passengers between node 3 and node 9, requiring a transfer in node 4, equals the penalty time multiplied with the penalty time value and the

passenger demand of that OD pair: i.e., ((45+30+61) - (45+3+61)) * 0.55 * 100 = 1,485. The total penalty fee of those who have transfers and detours is 102,245. The final operational income of all the passenger is 2,333,005.

Medium-scale Network Study

A medium-scale example is driven from the Chinese 4V4H HSR network (Figure 1) with 26 nodes and 676 OD pairs. The topology of the network can be seen in Figure 7.

For this experiment, we applied the approach presented in Section 4 and tested 3 different passenger demand scenarios. After testing several combinations of different number of Reduction and Extension iterations and diversification iterations on a small network, we concluded that the number of diversification iterations should be much higher than the number of Reduction and Extension iterations in order to obtain a high-quality solution efficiently. Therefore, for now, the number of Reduction and Extension iterations is set to 10 and the number of diversification iterations is set to 50.

The algorithm is executed during ten runs for each passenger demand scenario. For these ten runs, the maximum profit, average profit, the average percentage of improvement in profit compared to the initial line plan and the average computation time for a single run are selected as the parameters to illustrate the performance of the iterative approach. The results are listed in Table 2.

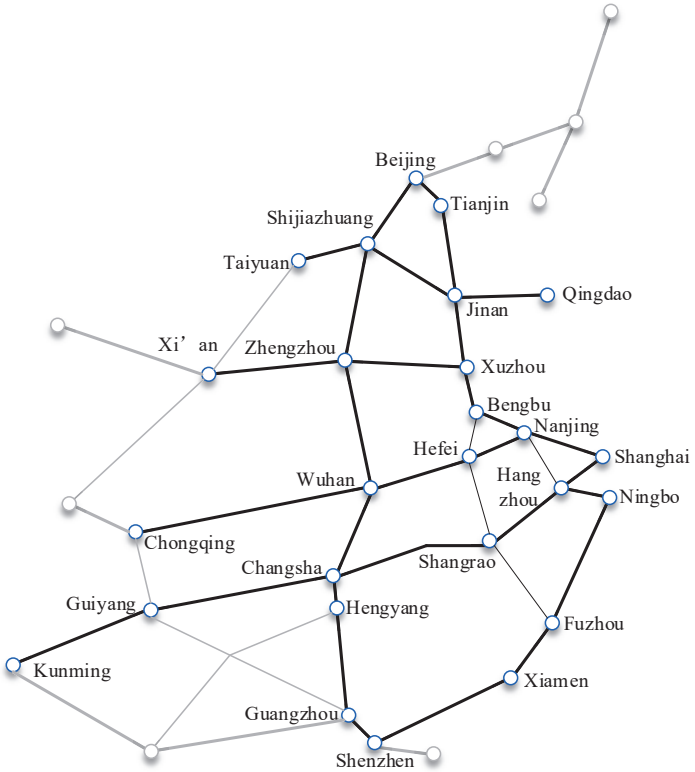


Figure 7: The medium-scale example HSR network (grey links and stations are not considered)

Table 2: Numerical experiment results on medium-scale network

Passenger demand scenarios	Maximum profit (*10^7)	Average profit (*10^7)	Improvement	Average running time
Initial 1	1.91	-	-	14s
Random 1	2.31	2.26	18%	2229s
Dedicated 1	2.34	2.26	18%	2711s
Initial 2	1.69	-	-	6s
Random 2	2.34	2.24	33%	2873s
Dedicated 2	2.33	2.24	33%	3017s
Initial 3	1.80	-	-	3s
Random 3	2.09	2.06	15%	1071s
Dedicated 3	2.10	2.08	16%	1786s

It can be seen from the improvement column that the proposed iterative algorithm of both random modification and dedicated modification performs well in this medium-scale network by increasing the profit from 15% to 33% compared to the initial line plan. The average of the improvements of random modification and dedicated modification are 21.8% and 22.4%. The improvement of the dedicated modification is the same or slightly better than the random modification in all three passenger demand scenarios.

From the aspect of computation time, the dedicated modification needs more time than the random modification because of the extra time for the heuristic line plan evaluation. However, in general, the computation time of both modifications are acceptable. To sum up, the solution approach is effective and efficient, and for the medium-scale network the dedicated and the random modification perform similarly.

Large-scale HSR Network Study

Also the large-scale HSR network (Figure 1) is used to show the efficiency and performance of the proposed algorithm. The running configurations setting is the same as the medium-scale network, however only 5 runs are performed for each scenario. The numerical experiment results are shown in Table 3.

Table 3: Numerical experiment results on large-scale network

Passenger demand scenarios	Maximum profit (*10^7)	Average profit (*10^7)	Improvement	Average running time
Initial 1	4.72	-	-	59s
Random 1	6.03	5.79	23%	11258s
Dedicated 1	6.11	6.04	28%	20694s
Initial 2	5.18	-	-	84s
Random 2	6.21	6.17	19%	6890s
Dedicated 2	6.27	6.24	20%	14805s
Initial 3	4.65	-	-	87s
Random 3	5.58	5.55	19%	7120s
Dedicated 3	5.64	5.57	20%	9586s

Due to the increase in the size of the network and the network connectivity, the network provides more choice for passengers to travel, which makes it much more complicated to solve the LPP and requires more computation time. In addition, the size of the solution neighbourhoods, mostly related to the number of lines considered, has increased significantly.

Also for the large network, both modifications improve the results of the corresponding initial line plan. The random modification and dedicated modification increase the profit with on average 20% and 23%. The results of the dedicated modifications are now clearly better than the random modifications for all three passenger demand scenarios. Obviously, the approach with dedicated modifications requires more computational time.

When looking at all six scenarios, some tendencies can be observed. The operational cost and income of all these scenarios are shown in Table 4. The operational income is an indication of the direct passengers and the operational cost indicates the frequency and number of lines. The M and L in Table 4 represent the medium-scale network and the large-scale network.

From Table 4, we can see that the operational costs of the modified line plans are always significantly lower than that of the initial line plan. However, the majority of the scenarios have a higher number of lines. The reason may that the proposed algorithm tends to reduce the frequencies of the long-distance lines and add short-distance lines on the busiest part instead.

Table 4: The maximum results of different scenarios

Passenger demand scenarios	**Average profit (*10^7)**	**Improvement**	**Average running time**
M Initial 1	3.92	2.01	8
M Random 1	3.90	1.59	10
M Dedicated 1	3.91	1.58	11
M Initial 2	3.95	2.26	8
M Random 2	3.94	1.61	11
M Dedicated 2	3.94	1.60	9
M Initial 3	3.74	1.94	7
M Random 3	3.73	1.65	8
M Dedicated 3	3.75	1.65	14
L Initial 1	10.60	5.88	9
L Random 1	10.65	4.62	13
L Dedicated 1	10.62	4.51	16
L Initial 2	10.73	5.55	11
L Random 2	10.66	4.45	10
L Dedicated 2	10.63	4.35	14
L Initial 3	9.90	5.24	11
L Random 3	9.87	4.29	10
L Dedicated 3	9.86	4.22	11

Taking large-scale network with passenger demand scenario 1 as an example, we compare the performance of the proposed approach. The results of its initial line plan and modified line plan are given in Table 5. The average length of the lines is weighted by frequencies.

The number of lines increase while the operational cost decrease when using the

dedicated modification. It is because that the average length of the lines is lower in the dedicated optimized line plan. It can be seen that he dedicated modification tends to reduce the line length.

Table 5: The results of large-scale network with passenger demand scenario 1

Line plan	Initial line plan	Random optimized line plan	Dedicated optimized line plan
Maximum profit (*10^7)	4.72	6.03	6.11
Operational cost (*10^7)	5.88	4.62	4.51
Standard income (*10^7)	10.82	10.82	10.82
Actual income (*10^7)	10.60	10.65	10.62
No. lines	9	13	16
Ave length of lines (km)	42695	23083	18227
No. single trains	7	5	7
No. doubled trains	128	114	123

Comparison with Different Objectives

We think there are two ways to compare our model with models from the state of the art with a cost-oriented objective or a customer-oriented objective. In the first method, the profit-oriented objective is modified by removing the cost part or the customer part. The cost-oriented objective can be obtained by ignoring the ticket income (and imposing that all demand should be served by at most two transfers). The customer-oriented objective can be obtained by ignoring the operational cost. In this case, an extra constraint should be considered on the maximum number and length of the lines operated in the line plan. Otherwise, in the end, each OD-pair would be served by its own direct line. The second method is to adjust the parameters or implicit weights associated with the operator cost part and the customer part in our objective. The cost-oriented approach is obtained by increasing the operating costs of the trains while the customer-oriented approach is obtained by increasing the penalty time value.

6 Conclusions

In this paper we tackle the Line Planning Problem (LPP). Instead of using a weighted sum of a cost-oriented objective and a customer-oriented objective, we propose the concept of travel time value in this paper and convert passenger travel time into ticket price and thus operator revenues. This allows to combine the operational cost and the passenger travel time in a single objective. In order to solve this complex problem, this paper presents an iterative approach. In the first stage of optimizing the LPP, an initial line plan is constructed heuristically. Based on the initial line plan, the iterative approach optimizes the line plan by reducing and extending different lines. Then, disturbances to the line plan are considered by removing or inserting an entire line.

We evaluate the performance and efficiency of the algorithms with numerical experiments on a medium and a large scale HSR network in China. Three different passenger scenarios are used. According to the experimental results, the proposed algorithm shows good performance in both examples and it is shown that the operational profit is improved by using the most promising moves calculated during our heuristic evaluation, compared to using random moves. On average, our resulting line plans increases the profit

by more than 20% for the medium and large-scale networks. For the large network, the dedicated modifications obtain better results compared to the random modifications. The proposed iterative approach intends to reduce the long-distance line frequencies and add more short-distance lines to make a trade-off between the operational cost and passenger travel time.

The further research will focus on involving different speeds of high-speed trains and increasing the efficiency of the algorithm to reduce the running time consumption. In addition, a variety of parameter sensitivity analyses will be done using a large amount of numerical experiments.

Acknowledgements
The research was partially supported by the National Key Research and Development Program of China (No. 2017YFB1200700-1), the National Natural Science Foundation of China (No. U1834209), the National Natural Science Foundation of China (No. 61603317) and the Special Fund for Basic Scientific Research of Central Universities (No. 2682017CX023). The author acknowledges the scholarship provided by China Scholarship Council. We are grateful for the contributions made by our colleagues Javier Duran Micco, Evert Vermeir and Guansheng Peng.

Appendix

Algorithm 3: Iterative approach of randomly move without evaluations

Repeat
  Repeat
    Reduction
      Select a e_{red} randomly from E_{red}
      Modify the L_{cur} and TSN_{cur} to obtain L_{nei} and TSN_{nei}
      if network connectivity check = true **do**
        Calculate passenger assignment and the profit of L_{nei}
        if $Z_{nei} > Z_{cur}$ **do**
          - $L_{cur} \leftarrow L_{nei}, TSN_{cur} \leftarrow TSN_{nei}, Z_{cur} \leftarrow Z_{nei}$
          - update E_{red} and E_{ext}
        else do
          - $L_{nei} \leftarrow L_{cur}, TSN_{nei} \leftarrow TSN_{cur}, Z_{nei} \leftarrow Z_{cur}$
          - delete this e_{red} from set E_{red}
      else do
        - roll back to L_{cur} and TSN_{cur}
    Extension method
      Select e_{ext} randomly from E_{ext}
      Modify the L_{cur} and TSN_{cur} to obtain L_{nei} and TSN_{nei}
      Calculate passenger assignment and the profit of L_{nei}
        if $Z_{nei} > Z_{cur}$ **do**
          - $L_{nei} \leftarrow L_{cur}, TSN_{nei} \leftarrow TSN_{cur}, Z_{nei} \leftarrow Z_{cur}$
          - update E_{ext} and E_{red}
        else do
          - $L_{nei} \leftarrow L_{cur}, TSN_{nei} \leftarrow TSN_{cur}, Z_{nei} \leftarrow Z_{cur}$
          - delete this e_{ext} from set E_{ext}
  until maximum number of Reduction and Extension iterations
  Disturb (the same as Algorithm 2)
until maximum number of diversification iterations
Output the line plan and objective profit

References

Borndörfer R, Karbstein M., 2012. "A direct connection approach to integrated line planning and passenger routing", In: *OASIcs-OpenAccess Series in Informatics*, Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Borndörfer, R., Grötschel, M., Pfetsch, M.E., 2007. "A column-generation approach to line planning in public transport", *Transportation Science*, vol. 41, pp. 123–132.

Bussieck, M.R., Kreuzer, P., Meng, L., Zimmermann, U.T., He, Z., 2015. "Optimal lines for railway systems", *European Journal of Operational Research*, vol. 96, pp. 54–63.

Floyd, Robert W., 1962. "Algorithm 97: Shortest path", *Communications of the ACM*, vol. 5(6), pp 345.

Friedrich, M., Hartl, M., Schiewe, A., Schobel, A., 2017. "Integrating Passengers' Assignment in Cost-Optimal Line Planning", In: *17th Workshop on Algorithmic Approaches for Transportation Modelling, Optimization, and Systems (ATMOS 2017)*, Schloss DagstuhlLeibniz-Zentrum fuer Informatik.

Fu, H., Nie, L., Meng, L., Sperry, B.R., He, Z., 2015. "A hierarchical line planning approach for a large-scale high speed rail network: The China case", *Transportation Research Part A: Policy and Practice*, vol. 75, pp. 61–83.

Goerigk M., Schmidt M., 2017. "Line planning with user-optimal route choice", *European Journal of Operational Research*, vol. 259(2): pp. 424-436.

Karbstein M., 2014 "Integrated line planning and passenger routing: connectivity and transfers", *Operations Research Proceedings 2014*. *Springer, Cham*, pp. 263-269.

Liu, X., Li, B., 2018. "A Study on the Scheme Selection of Transfer Activities within the High-Speed Railway Hub", *Railway Transport and Economy*, vol. 11, pp. 32-37.

Nachtigall, K., Jerosch, K., 2008. "Simultaneous network line planning and traffic assignment", In: *8th Workshop on Algorithmic Approaches for Transportation Modeling, Optimization, and Systems (ATMOS'08)*, Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Rosalia, R. (ed.), 2017. *The line planning with simultaneous optimisation of operational cost and travel time*, MS thesis, Norwegian University of Science and Technology.

Schöbel, A., Scholl, S., 2006. "Line Planning with Minimal Traveling Time", In: *5th Workshop on Algorithmic Methods and Models for Optimization of Railways (ATMOS'05)*, Dagstuhl-Leibniz-Zentrum fuer Informatik.

Schöbel, A., 2012. "Line planning in public transportation: models and methods", *OR spectrum*, vol. 34, pp. 491–510.

Yang L., Zhang Y., Li S., Gao, Y., 2016. "A two-stage stochastic optimization model for the transfer activity choice in metro networks", *Transportation Research Part B: Methodological*, vol. 83, pp, 271-297.

Connectivity Reliability on an Urban Rail Transit Network from the Perspective of Passengers' Travel

Jie Liu[1,2], Qiyuan Peng[1,2], Jinqu Chen[1,2] and Yong Yin [1,2*]

[1] Southwest Jiaotong University, Chengdu, China

[2] National United Engineering Laboratory of Integrated and Intelligent Transportation, Chengdu, China

yinyong@home.swjtu.edu.cn

Abstract

In the context of the urbanization and the rapid development of Urban Rail Transit (URT). The reliability of URT network is getting attention. To measure it, three indicators are constructed from passengers' tolerable travel paths, passengers' travel efficiency and passengers' travel realization on URT network, respectively. The tolerable coefficient which is the ratio of passengers' tolerable travel time to shortest possible travel time is proposed and added to indicators. It can reflect passengers' travel paths choice behavior. The ratio of affected passenger volume (RPV) is proposed to identify the important stations. The Automatic Fare Collection (AFC) data, train running time data are used to calculate the passenger volume and the number of passengers' tolerable travel paths in Wuhan subway (China). Finally, the connectivity reliability of Wuhan subway network is analysed through simulating attack stations. The result shows that the important stations identification indicators of Degree Centrality (DC), Betweenness Centrality (BC) and ratio of affected passenger volume (RPV) can effectively identify the important stations on connectivity reliability of Wuhan subway. In particular, the important station identification indicator of RPV can identify the stations effectively which have great influence on passengers' travel realization. In addition, attacking stations has greater impact on passengers' tolerable travel paths than passengers' travel efficiency and passengers' travel realization.

Keywords

Urban Rail Transit network, Connectivity reliability, Passengers' travel, Tolerable travel paths, Travel efficiency, ravel realization

*Corresponding author, email: yinyong@home.swjtu.edu.cn

1 Introduction

In recent years, China is experiencing rapid urbanization. Urban Rail Transit (URT) with large capacity, high speed and environmental protection is constructing rapidly. Now, China has longest URT operation lines in the world. However, some emergencies (such as natural disasters and operation accidents) always cause great harm to the operation of URT. Passengers' travel time will increase and passengers' travel paths will be disrupted because of these emergencies. In addition, URT network is a small world network. Most of the nodes are not connected to each other directly in small world network. It means that the connectivity of URT network is easily affected by emergencies. Therefore, how to evaluate and analyze the connectivity reliability of URT network is of great significance to improve the reliability of URT and ensure the normal operation of URT.

Connectivity reliability was proposed by Mine and Kaiwai (1982) at first [8]. In connectivity reliability research, researchers always studied connectivity reliability of transportation network based on graph theory and topology of traffic network. Such as Bell & Iida (1997), and Wakabayashi &Iida (1992) thought that connectivity reliability is the probability that there is still a connection between a pair of nodes when one or more links are removed [1,11]. Zhang et al. (2015) studied the resilience of seventeen principal networks in terms of connectivity. Some researchers researched the connectivity reliability of transportation network considering passengers [13]. Such as Mattsson and Jenelius (2015) summarized and reviewed the existing research on transportation reliability. They pointed out that it is necessary to consider traffic demand and transport supply to study the reliability of transportation network [7]. Zhang et al. (2009) proposed LOS-based connectivity reliability evaluation model to calculate connectivity reliability of a regional transportation network [12]. Liu et al. (2017) analyzed the connectivity reliability of rail transit with Monte Carlo simulation [5]. Guidotti et al. (2017) proposed two types of indicators to measure connectivity reliability of transport network. The indicators that consider weights of nodes and links are compared with indicators that did not consider weights of nodes and links [2]. Li et al. (2014) proposed four connectivity reliability indicators. Then, the indicators were weighted to establish one indicator. Finally, they analyzed the connectivity reliability of Beijing URT [4]. Reggiani et al. (2015) further deepen the analysis of how resilience and vulnerability can be framed, interpreted and measured, and their relationship with connectivity [9].

In these literatures on connectivity reliability of transport networks. Most researchers assumed that as long as there is at least one connected path between the two nodes, the two nodes are in a connected state. However, this assumption does not entirely consistent with passengers' travel path choice behavior. Actually, passengers inclined to choose the travel path whose travel time is shortest. If the shortest travel path (Corresponding to shortest travel time) can not be used, passengers will choose other path whose travel time is less than they could bear (tolerable travel time). If all connected paths' travel time is longer than their tolerable travel time, although the paths are connected, passengers will not use them. Therefore, tolerable coefficient is put forward to confirm whether a travel path is tolerable. In addition, researchers emphasized on studying the topological connectivity reliability of URT network. The passenger volume should be considered when analyzing the connectivity reliability of URT network.

2 Method

To research the connectivity reliability of URT network, the URT network is defined as a directed graph. Three indicators are put forward to measure the connectivity reliability of URT network from passengers' tolerable travel paths, passengers' travel efficiency and passengers' travel realization on URT network, respectively. Based on the maximum impact of passengers, a new indicator to identify the important stations is put forward. The connectivity reliability of URT network is analysed by simulating destroy stations.

2.1 Network Definition

The URT network can be defined by a directed graph $G(N,E)$. N represents the set of stations. $E \subseteq N \times N$, it represents the set of links between stations. There are multiple connected travel paths from one station to another. Assuming that $path_{od}$ is the set of connected path(s) from station o to station d. $path_{od}^{i}$ is the path i from station o to station d, $path_{od}^{i} \in path_{od}$. $path_{od}^{i}$ includes set of stations $path_{od}^{i,N}$ and set of links $path_{od}^{i,E}$. The set of stations $path_{od}^{i,N}$ in $path_{od}^{i}$ includes sets of transfer station(s) $path_{od}^{i,N_1}$ and non-transfer station(s) $path_{od}^{i,N_2}$.

2.2 Passengers' Tolerable Travel Paths

As mentioned above, not all connected paths is meaningful for passengers. Only the connected path is tolerable travel path, passengers use it. To calculate the number of tolerable travel path from one station to another, tolerable coefficient α is put forward. Assuming that $path_{od}^{i}$ is path i from station o to station d, equation (1) is used to confirm whether $path_{od}^{i}$ is a tolerable travel path.

$$n\left(t_{path_{od}^{i}}, \min(t_{path_{od}})\right) = \begin{cases} 1; t_{path_{od}^{i}} / \min(t_{path_{od}}) \le \alpha \\ 0; otherwise \end{cases}. \quad (1)$$

Where $n\left(t_{path_{od}^{i}}, \min(t_{path_{od}})\right)$ is a 0-1constant. If the path i is a tolerable travel path, then $n\left(t_{path_{od}^{i}}, \min(t_{path_{od}})\right) = 1$, otherwise $n\left(t_{path_{od}^{i}}, \min(t_{path_{od}})\right) = 0$. $path_{od}$ is the set of connected path(s) from station o to station d. $t_{path_{od}^{i}}$ and $t_{path_{od}}$ are the travel time of path i and the travel time set of connected path(s) from station o to station d, respectively. α is the tolerable coefficient. It reflects the relation between tolerable travel time and shortest possible travel time.

$t_{path_{od}^{i}}$ can be calculated with equation (2). It includes the train running time in links (including the train start and stop time), train dwell time, passengers' transfer time and passengers' waiting time.

$$t_{path_{od}^{i}} = \sum_{e \in path_{od}^{i,E}} t_e + \sum_{m \in path_{od}^{i,N_1}} t_m^{dwell} + \sum_{n \in path_{od}^{i,N_2}} t_n + t_{wait}^{o}. \quad (2)$$

Where e and $path_{od}^{i,E}$ are link e and set of link(s) in path i from station o to station d, respectively. t_e is the train running time in link e. m and $path_{od}^{i,N_1}$ are non-transfer station m and set of non-transfer station(s) in path, respectively. t_m^{dwell} is the train dwell time at non-transfer station m. n and $path_{od}^{i,N_2}$ are transfer station n and set of transfer station(s) in path i, respectively. t_n is the transfer time at transfer station n. t_{wait}^{o} is passengers' waiting time at origin station o which can be estimated as half of the headway.

2.3 Evaluating Connectivity Reliability of URT Network

Most of researchers usually adopted the network efficiency (the mean of the reciprocal of

shortest distance between all nodes), maximal connected subgraph and other similar indicators to measure the connectivity of URT network. These indicators lack of information on passenger volume and the number of tolerable travel paths. Therefore, three indicators which consider the passenger volume and tolerable travel paths are proposed to evaluate the connectivity reliability of URT network.

2.3.1 The Number of Tolerable Travel Paths in URT Network

The number of tolerable travel paths from one station to another can reflect the connectivity reliability from one station to another. Sometimes, operational accidents, terrorist attacks and natural disasters will destroy the URT network. They cause some tolerable paths become unavailable. Therefore, the higher the number of tolerable travel paths between station pairs, the higher the probability that passengers can travel between them. Before the URT network is destroyed, the average number of tolerable travel paths for every passenger is used to evaluate the connectivity reliability of URT network. It is represented as equation (3):

$$R_{path}^{0} = \frac{\sum_{o\in N}\sum_{d\in N,o\neq d}\sum_{path_{od}^{i}\in path_{od}} v_{od}\cdot n\left(t_{path_{od}^{i}}, \min(t_{path_{od}})\right)}{V}. \tag{3}$$

Where R_{path}^{0} is average number of tolerable travel paths for every passenger before the network is destroyed. N is the set of stations in URT network. v_{od} is the passenger volume from station o to station d. V is passenger volume in URT network.

After the URT network is damaged, the average number of tolerable travel paths for every passenger will decrease. Assuming that the network is suffered from damage event δ. It causes x number of stations lost their functions. In this situation, the average number of tolerable travel paths for every passenger is $R_{path}^{(\delta,x)}$. The relative number of tolerable travel paths is used to measure the connectivity reliability of URT network after the network is destroyed. It can be calculated with equation (4):

$$R_{path} = \frac{R_{path}^{(\delta,x)}}{R_{path}^{0}}. \tag{4}$$

Where R_{path} is relative number of tolerable travel paths in URT network. It can reflect the connectivity reliability of URT network from passengers' tolerable travel paths when the network is destroyed.

2.3.2 Travel Efficiency of URT Network

The connectivity reliability of URT network is always measured from passengers' shortest travel time (network efficiency). The implicit assumption of network efficiency is that every station plays the same role on the network (the weight of stations is same). However, the functions of stations in URT network are different. The passenger volume between stations varies considerably. Therefore, in order to measure the travel efficiency of URT, the passenger volume is considered. Before the URT network is destroyed, the travel efficiency of URT network can be presented by equation (5):

$$E_{eff}^{0} = \frac{1}{V}\sum_{o\in N}\sum_{d\in N,o\neq d} v_{od}\cdot\frac{1}{\min(t_{path_{od}})}. \tag{5}$$

Where E_{eff}^{0} is travel efficiency of URT network when URT network is not damaged.

The passenger volume is used to calculate the stations' weight. $\left(\sum_{d\in N,o\neq d} v_{od}\right)\Big/V$ is the weight of station o. Assuming that the network is suffered from damage event δ. It causes x

number of stations lost their functions. In this situation, the travel efficiency will decrease to $E_{eff}^{(\delta,x)}$. The relative travel efficiency of URT network can be calculated with equation (6):

$$R_{eff} = \frac{E_{eff}^{(\delta,x)}}{E_{eff}^{0}}. \quad (6)$$

Where R_{eff} is relative travel efficiency of URT network, which can reflect the connectivity reliability of URT network when the network is destroyed.

2.3.3 The Rate of Passengers' Travel Realization on URT Network

In normal operations, most of passengers can travel on URT network successfully. Therefore, Passengers' travel realization rate in URT network is near to 100%. However, if the URT network is destroyed, then some passengers' tolerable travel paths are interrupted. passengers will give up travel on URT network because their tolerable travel paths are interrupted. The Passengers' travel realization rate will decrease. Therefore, the rate of passengers' travel realization on URT network is put forward to measure the connectivity reliability of URT network. Assuming that the network is suffered from damage event δ. It causes x number of stations lost their function. In this situation, the rate of passengers' travel realization on URT network is represented by equation (7):

$$R_{rate} = \frac{V^{(\delta,x)}}{V} = \frac{\sum_{o\in N}\sum_{d\in N,o\neq d} v_{od}\cdot n_{od}^{(\delta,x)}}{V}. \quad (7)$$

Where R_{rate} is the rate of passengers' travel realization on URT network when the network is destroyed. $V^{(\delta,x)}$ is passenger volume that can travel on URT network when URT network is damaged. $n_{od}^{(\delta,x)}$ is a 0-1 constant, if there is at least one tolerable travel path from station o to d ,then $n_{od}^{(\delta,x)}$ is 1, otherwise, $n_{od}^{(\delta,x)}$ is 0.

2.3.4 Identifying Important Stations

Many researchers have done some work on identifying important nodes (Liu et al., 2016; El-Rashidy and Grant-Muller, 2014; Hu et al., 2015) in complex network [6,10,3]. Some indicators had been used to evaluate the importance of nodes, such as Degree Centrality (DC), Betweenness Centrality (BC) and Closeness Centrality (CC). DC emphasizes the number of links linked to the node directly. BC describes the ratio of all shortest paths that passing through the node in the network. CC reflects distances between the node and other nodes. However, these indicators focus on identifying the important nodes from the topology of the network. The passenger volume has not been considered. Therefore, to reflect the importance of nodes to passengers' travel, the ratio of passenger volume (RPV) affected by the station to the total passenger volume is used to measure the importance of the station. Supposing that some passengers' travel is affected by station j. Then, station j can affect passengers whose origin station is j, whose destination station is j and whose travel path includes station j. The importance indicator of station j is represented by equation (8):

$$I_j = \frac{\sum_{o\in N,o\neq j}\sum_{d\in N,o\neq d} v_{od}^{j} + \sum_{d\in N,d\neq j}(v_{jd}+v_{dj})}{V}. \quad (8)$$

Where I_j is the importance indicator of station j. v_{od}^{j} is passenger flow travel from station o to station d via station j. v_{jd} and v_{dj} are passenger volume who travel from station j to station d and travel from station d to station j, respectively. V is passenger volume in URT network.

In equation (8), to calculate the importance indicator of station j, the passenger flow via station j need to be confirmed. A user stochastic equilibrium model is used to calculate the passenger flow in URT network. Then the importance indicators of all stations can be calculated.

3 Implementation

3.1 Data Preparation

Wuhan subway system in China is used to validate effectiveness of the method and indicators used. Figure 1 shows the operation lines, stations' name and numbers for Wuhan subway in September 2018.

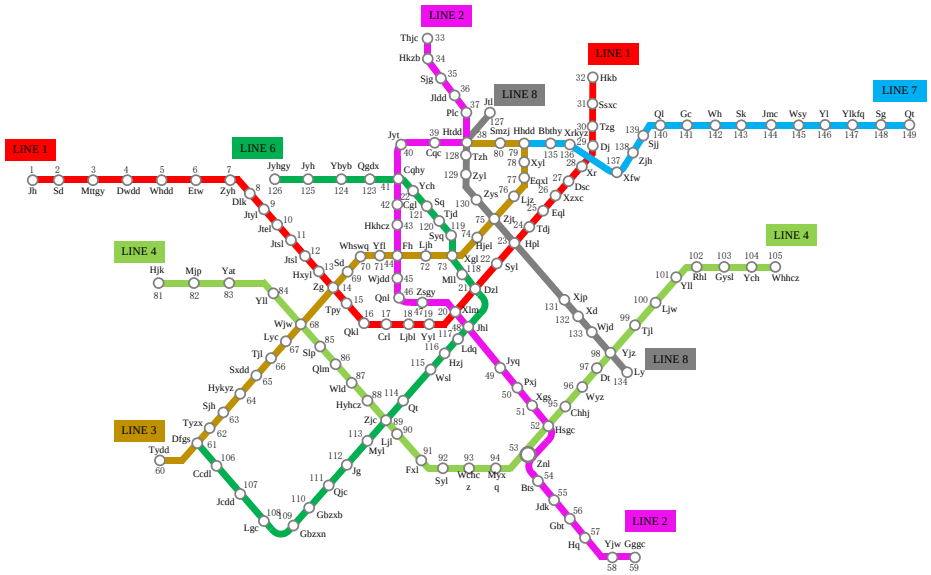


Figure 1: Network of Wuhan subway in September 2018

The Automatic Fare Collection (AFC) data record for 5 working days in September 2018 were obtained from Wuhan subway company. The data recorded the types of the tickets, tickets' number, entry and exit time at stations, stations' names and numbers. the tickets' number is matched to obtain the stations' entry and exit time for passengers. Passenger volume between stations is counted to construct Origin-Destination (OD) matrices during morning peak hours.

To calculate the travel time between stations, the travel time (including train dwell time and train running time) of links, headway of different lines and the transfer time at transfer stations are obtained from Wuhan subway company.

3.2 Important Stations

Degree Centrality (DC) of stations, Betweenness Centrality (BC) of stations, Closeness Centrality (CC) of stations and the ratio of passenger volume affected by the stations (RPV) during

morning peak hours are calculated, respectively. The most important ten stations in Wuhan subway network are listed in Table 1.

Table 1: Ten important stations identified by four identification indicators

Stations' number	DC	Stations' number	BC	Stations' number	CC	Stations' number	RPV
38	5	75	0.026	21	11.05	21	0.030
14	4	73	0.025	73	11.12	48	0.026
20	4	21	0.023	20	11.16	20	0.024
21	4	14	0.021	118	11.29	23	0.022
23	4	48	0.020	22	11.32	73	0.022
41	4	74	0.019	74	11.35	75	0.022
44	4	68	0.019	48	11.43	14	0.020
48	4	23	0.017	72	11.50	44	0.016
68	4	89	0.016	23	11.51	118	0.016
73	4	76	0.015	75	11.54	89	0.016

Table 1 shows that four identification indicators have identified some same important stations. Such as, station 21, station 20, station 23 and station 48. These four stations are transfer stations in Wuhan subway network. To analyse the connectivity reliability of Wuhan subway network and demonstrate the effectiveness of different identification indicators, the influence of important stations' failure on connectivity reliability of Wuhan subway network is analysed.

3.3 Connectivity reliability of Wuhan Subway Network

To analyse the connectivity reliability of Wuhan subway network, MATLAB is used to simulate destroying stations. If the station is destroyed, then the station and the links which are connected to the station directly are removed from the network. The connectivity reliability of Wuhan subway network is reflected by calculating the three indicators (from equation (3) to (7)). Supported by National Key R & D Program of China (2017YFB1200700) and The National Natural Science Foundation of China (No. U1834209), we conducted an in-depth questionnaire survey on characteristics of passengers' travel path choices in Wuhan subway. It is found that passengers' tolerable coefficient in Wuhan subway is 1.38.

3.3.1 The Relative Number of Tolerable Travel Paths in Wuhan Subway Network

The connectivity reliability of Wuhan subway network can be reflected by relative number of tolerable travel paths R_{path} . Attacking stations randomly and attacking important stations deliberately are used to simulate destroying stations. The Figure 2 shows relative number of tolerable travel paths R_{path} simulation outputs under five attack modes. Before the stations are destroyed, the number of tolerable travel paths in Wuhan subway network is 3.49. It means that every passenger in Wuhan subway network has 3.49 tolerable travel paths on average. After the stations are destroyed, it is found that attacking important stations deliberately can make R_{path} decrease quickly. Attacking stations randomly leads to the decease of R_{path} moderately. In addition,

R_{path} is more sensitive to attacking important stations that are identified by identification indicators of CC, BC and RPV. R_{path} decease to 0.37 (average passenger has 1.29 tolerable travel paths) when one station (station 21) identified by indicators of CC and RPV is destroyed. Therefore, the station 21 plays an important role in connecting tolerable travel paths. It also shows that important stations have huge impact on the diversity of passengers' travel paths choices. Attacking one to three important stations that are identified by identification indicators of BC, CC and RPV make R_{path} decease much quickly. When over three important stations are destroyed, the identification indicators of DC, BC and RPV can identify the important stations effectively which can influence R_{path} heavily.

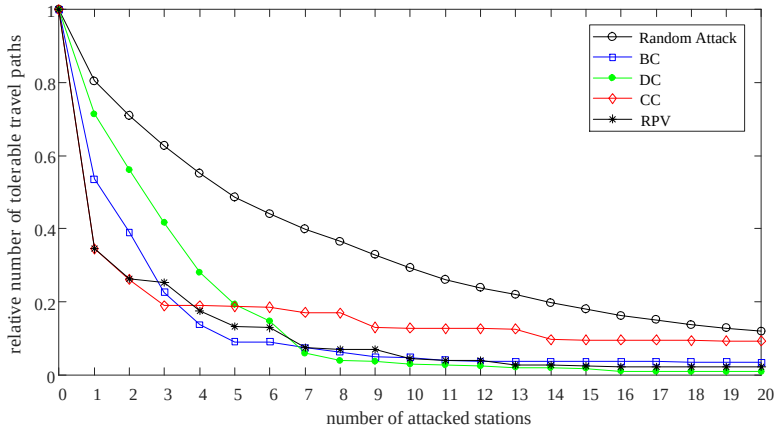


Figure 2: Relative number of tolerable travel paths R_{path} simulation outputs under five attack modes

3.3.2 Relative Travel Efficiency of Wuhan Subway

The shortest travel time between stations will increase when stations are destroyed. Therefore, the connectivity reliability of Wuhan subway network can be reflected by relative travel efficiency R_{eff} of Wuhan subway. The Figure 3 shows relative travel efficiency R_{eff} simulation outputs under five attack modes. Before the stations are destroyed, R_{eff} of Wuhan subway network is 0.037. When the stations are destroyed, attacking important stations deliberately can make R_{eff} decrease quickly. However, compared with relative number of tolerable travel paths (Figure 2), R_{eff} decreases slowly when stations are destroyed. It can bear attacking two stations before R_{eff} drops to 0.8. Therefore, attacking stations has greater impact on passengers' tolerable travel paths than passengers' travel efficiency. In addition, R_{eff} is more sensitive to attacking important stations which are identified by identification indicators of BC, DC and RPV. When less than four stations are destroyed, then the identification indicator of BC can identify the important stations which can influence R_{eff} heavily. When over five important stations are destroyed, identification indicator of DC can identify the important stations which can influence R_{eff} heavily.

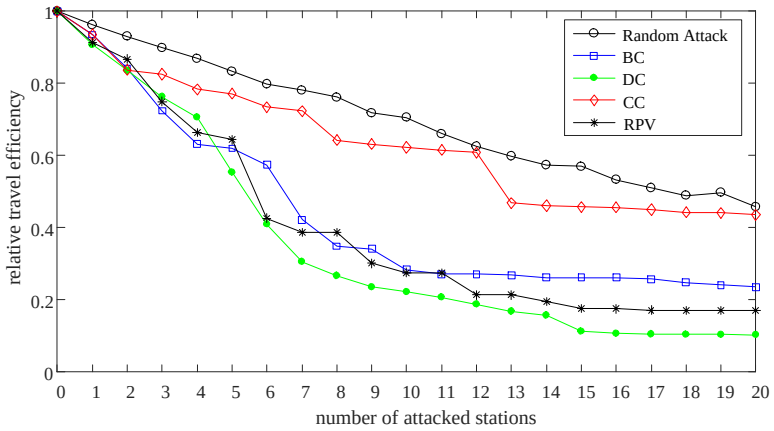


Figure 3: Relative travel efficiency of Wuhan subway simulation outputs under five attack modes

3.3.3 The Rate of Passengers' Travel Realization in Wuhan Subway Network

Some passengers' tolerable travel paths will be unconnected when stations are destroyed in URT network. It causes some passengers' travel can not be realized on URT network. Therefore, the connectivity reliability of Wuhan subway network can be reflected by the rate of passengers' travel realization R_{rate} . The Figure 4 shows the rate of passengers' travel realization R_{rate} simulation outputs under five attack modes. Attacking three important stations which are identified by indicators of BC, DC and RPV causes R_{rate} decrease to less than 0.8. It means that more than 20% of passengers will not travel on Wuhan subway network. Attacking six important stations which are identified by indicators of DC and RPV are destroyed causes R_{rate} decrease to nearly 0.4. Therefore, only 40% of passengers can get tolerable travel paths to travel on Wuhan subway network. The rate of passengers' travel realization decreases more slowly than relative number of tolerable travel paths when stations are destroyed (Figure 2). It also proved that the impact of attacking stations on passengers' tolerable travel paths is greater than passengers' travel realization. In addition, when attacked stations are over three, identification indicator of RPV can identify the important stations which have most influence on R_{rate} . It demonstrates that identification indicator of RPV is effective to identify the important stations that can influence R_{rate} heavily.

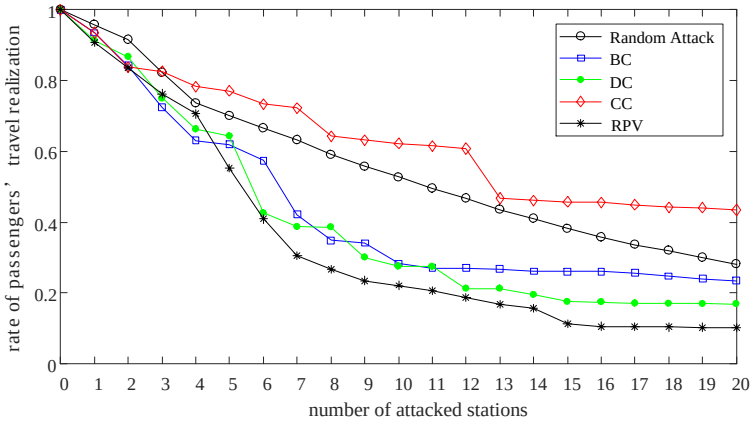


Figure 4: Passengers' travel realization simulation outputs under five attack modes

4 Conclusions

The connectivity reliability of URT network are measured from passengers' tolerable travel paths, passengers' travel efficiency and passengers' travel realization on URT network, respectively. Three indicators which considering passenger volume and passengers' tolerable coefficient are used to analyze the connectivity reliability of Wuhan subway network. A new indicator RPV which can maximize the number of affected passengers is proposed to identify the important stations. The important stations identification indicators of DC, BC, CC and RPV are used to identify the important stations in Wuhan subway network. Combining the above measures together, the connectivity reliability of Wuhan subway network is analyzed through attack stations simulation. Some findings and conclusions are summarized as below:

- The connectivity reliability of Wuhan subway is more sensitive to attacking important stations deliberately than attacking stations randomly. The indicators of DC, BC and RPV can effectively identify the important stations on connectivity reliability of Wuhan subway. The simulation result shows that the different important stations identification indicator can identify the important stations effectively when the connectivity indicators and the number of destroyed stations are different.
- Compared with relative number of tolerable travel paths, the relative travel efficiency and the rate of passengers' travel realization decrease slowly when the stations are destroyed. Before the relative travel efficiency and the rate of passengers' travel realization drop to 0.8, they can bear attacking two stations. Therefore, attacking stations has greater impact on the passengers' tolerable travel paths than passengers' travel efficiency and travel realization.
- The new indicator RPV can identify the important stations on connectivity reliability of Wuhan subway effectively. In particular, it can identify the important stations that can influence the passengers' travel realization on URT network most.

In URT network, identifying important stations effectively is of great importance on the operation of URT network. Since, the connectivity reliability of URT network can be improved by protecting important stations. The three indicators are used to measure the connectivity reliability of URT network comprehensively. Although one URT network is analyzed here, the same indicators and method can be used to other URT networks. Further studies will consider more factors, such as using historical data to confirm the probability of stations failure and considering the passengers' travel quality.

Acknowledgements

The authors thank the Wuhan subway company for providing relevant data. We also acknowledge the support of National Key R & D Program of China(2017YFB1200700) and The National Natural Science Foundation of China (No. U1834209).

References

1. Bell, M.G.H., Iida, Y., 1997. *Transportation Network Analysis*. Wiley, Chichester.
2. Guidotti, R., Gardoni, P., Chen, Y., 2017. “Network reliability analysis with link and nodal weights and auxiliary nodes”, *Structural Safety*, vol. 65, pp. 12-26.
3. Hu, P., Fan, W., & Mei, S., 2015. “Identifying node importance in complex networks”, Physica A: Statistical Mechanics and its Applications, vol. 429, pp. 169-176.
4. Li, M., Jia, L., Wang, Y., 2014. “Research and implementation on connectivity reliability calculation algorithm of urban rail transit network operation”, Proceeding of the 11th World Congress on Intelligent Control and Automation, Shenyang, China.
5. Liu, J., Lu, H., Ma, H., et al., 2017. “Network vulnerability analysis of rail transit plans in Beijng-Tianjin-Hebei region considering connectivity reliability”, *Sustainability* vol.9, pp.1479.
6. Liu, J., Xiong, Q., Shi, W., et al., 2016. “Evaluating the importance of nodes in complex networks”, *Physica A: Statistical Mechanics & Its Applications*, vol. 452, pp. 209-219.
7. Mattsson, L.G., Jenelius, E., 2015. “Vulnerability and resilience of transport systems-A discussion of recent research”, *Transp. Res. Part A,* vol.81, pp. 16-34.
8. Mine, H., Kawai, H., 1982. *Mathematics for reliability analysis*, Tokyo: Asakura-shorten.
9. Reggiani, A., Nijkamp, P., Lanzi, D., 2015. “Transport resilience and vulnerability: the role of connectivity”, *Transp. Res. Part A*, vol. 81, pp. 4-15.
10. Rodriguez-Nunez, E., Garcia-Palomares, J.C., 2014. “Measuring the vulnerability of public transport networks”, *J. Transp. Geogr,* vol. 35, pp.50-63.
11. Wakabayashi, H., Iida, Y., 1992. “Upper and lower bounds of terminal reliability in road networks: an efficient method with Boolean algebra”, J. Nat. Disaster Sci, vol. 14, pp. 29-44.
12. Zhang, X., Jia, L., Dong, H., et al., 2009. “Analysis and evaluation of connectivity reliability for dynamic transportation network”, 2009 Fifth International Joint Conference on INC, IMS and IDC, Seoul, South Korea.
13. Zhang, X., Miller-Hooks, E., Denny, K., 2015. “Assessing the role of network topology in transportation network resilience”, J. Transp. Geogr, Vol. 46, pp. 35-45.

Energy-efficient Metro Train Operation Considering the Regenerative Energy: A Discrete Differential Dynamic Programming Approach

Junjie Lou [a], Xuekai Wang [a,1], Shuai Su [a,2], Tao Tang [a], Yihui Wang [a],
[a] State key lab of railway control and safety, Beijing Jiaotong University
Shangyuancun No. 3, Haidian District, Beijing, China
[1] E-mail: 17120285@bjtu.edu.cn, Phone: +86-13363951131
[2] E-mail: shuaisu@bjtu.edu.cn, Phone: +86-13810879341

Abstract
With the increase of the operating mileage, a large amount of energy consumption generated by metro systems needs to be taken seriously. One of the effective ways to reduce the energy consumption is to collaboratively optimize the driving strategy and train timetable by considering the regenerative energy (RE). However, the dimensionality and computational time will increase accordingly in optimization as the number of operating trains rises. With the intention of tackling this problem by efficiently reducing dimensionality, the energy-efficient metro train operation problem is optimized in this paper by applying the discrete differential dynamic programming (DDDP) approach. Firstly, the model calculating the net energy consumption that takes into account the RE is formulated. Then, the optimization model will be transformed to a discrete decision problem by using Space-Time-Speed (STS) network methodology, and the corresponding solution will be obtained through the DDDP based algorithm. Finally, two case studies will be conducted in a metro network to illustrate the effectiveness of the proposed approach.

Keywords
Energy reduction, Regenerative energy, Space-Time-Speed network methodology, Discrete differential dynamic programming

1 Introduction

Due to the advantages of high efficiency, large capacity and energy-efficient, metro systems are developing rapidly worldwide. However, with the increase of the operating mileage, a large amount of energy consumption generated by metro systems needs to be taken seriously. Furthermore, the traction energy accounts for the most important part in the energy consumption of the system. What is more, utilization of the regenerative energy (RE) provides us a good opportunity to reduce the energy consumption in metro systems. As a result, it is necessary to carry out the research on the optimization and control of the metro train operation by considering the RE.

The optimization and control of train operation are divided into driving strategy optimization and train timetable optimization. These two methods determine the energy consumption of the train operation by influencing the traction energy consumption and the

reused RE. On the one hand, optimization of the driving strategy aims to find the energy-efficient control strategy so that the traction energy consumption is minimized by optimizing the regime sequences and the switching points. Literature on driving strategy optimization can date back to 1960s: Ishikawa (1968) proposed an optimal control model on the assumption that the train runs on a flat track with constant gradient and traction efficiency. Later, Khmelnitsky (2000) presented a complete study on the optimal train control problem, in which variable gradients, variable traction efficiency and arbitrary speed limits were all considered. Liu and Golovitcher (2003) gave an analytical solution to the problem with variable gradients for finding driving strategies for each part of the route. Chang and Sim (2008) applied a genetic algorithm on the train control problem to generate an optimal coast control based on evaluation of the punctuality, riding comfort and energy consumption together. Keskin and Karamancioglu (2017) developed the optimal train operation strategies by using three nature-inspired metaheuristic algorithms: Genetic Simulated Annealing, Firefly, and Big Bang-Big Crunch.

On the other hand, optimization of the train timetable can greatly reduce the traction energy consumption as well as efficiently utilize the RE from the macroscopic views, such as Su et al. (2013) provided an analytical formulation to calculate the optimal speed profile with fixed trip time for each section. He also designed a numerical algorithm to distribute the total trip time among different sections and prove the optimality of the distribution algorithm. Furthermore, Su et al. (2015) proposed a bisection method to solve the optimal departure time for an accelerating train. Rodrigo et al. (2013) used a semi analytical solution that leads to a discretization and to the application of the Langrange multipliers method to solve the optimization of n-tuples of speed. Tian et al. (2017) proposed a multi-train traction power network modelling method to determine the system energy flow of the railway system with regenerating braking trains. Yin et al. (2016) developed a stochastic programming model for metro train rescheduling problem in order to jointly reduce the time delay of affected passengers, their total traveling time and operational costs of trains.

However, in this field, previous research mainly carried on optimization by considering the driving strategy or the train timetable, which can only achieve the local optimization of the energy consumption. In order to get the global optimal solution, some experts proposed a kind of collaborative optimization which focuses on both aspects. In this way, the energy consumption of the metro system will be reduced from the perspective of the system and the performance of optimization will be significantly improved. For example, Bocharnikov et al. (2010) presented a single train speed profile optimization model considering both tractive energy consumption and utilization of RE. Furthermore, the authors performed a multi-train simulation to estimate the benefits and effects of the optimal speed profile on minimizing the net energy consumption. Li and Lo (2014) gave the quantitative analysis of tractive energy consumption, RE utilization and net energy consumption, then they proposed an energy-efficient scheduling and speed control model to minimize the net energy consumption, which assuming all trains run with maximum acceleration, coasting and maximum deceleration in each segment. Ning et al. (2018) proposed a two-stage urban rail transit operation planning approach comprising running time allocation and RE utilization to save energy consumption. Bu et al. (2018) set up a 'time slot and energy grid' model, which can effectively reduce the complexity of analyzing the usage of RE among multiple bidirectional running trains. Based on the model, they designed the energy saving method. Zhou et al. (2018) proposed an integrated optimization model on train control and timetable to minimize the net energy consumption, in which the proposed train control is based on

finding the optimal switching points among the control modes of maximum acceleration, cruising, coasting, and maximum braking to minimize the net energy consumption, while cruising and coasting regimes might be adopted for more than one time.

Nevertheless, among the existing research about integrated optimization, researchers mainly adopted two kind of ways to obtain the optimal solution: Some authors adopted the two-stage method to optimize the driving strategy and the train timetable respectively, this kind of hierarchy optimization can not make full use of real-time RE; Other authors achieved the simultaneous optimization, but they need to specify the transition sequence of control modes artificially. The solution obtained by above methods is still not optimal. As a result, it is necessary to simultaneous optimize the driving strategy and train timetable when the control strategy is uncertain. This paper proposes a integrated optimization approach of the driving strategy and the train timetable. In the approach, the control mode of trains can be arbitrary at each time.

Another challenge encountered in related research is that the problem of multi-train RE utilization is a highly dimensional problem. Because the states of multiple trains need to be considered at the same time during the optimization, "the curse of dimensionality" often exists with the number of trains increasing. Many researchers have tried to overcome similar multi-variable optimization problems by using improved DP algorithms, in which the discrete differential dynamic programming (DDDP) approach is an effective way. It is an iterative method firstly proposed by Heidari et al. (1971) when optimizing the operating policies of multiple unit and multiple purpose water resources systems. It can sharply reduce the computing time as well as the required computer's memory space by decreasing the dimension of the problem. A relative coefficient based on maximum output capacity and an adaptive bias corridor technology are proposed by Li et al. (2014) to improve the DDDP approach in order to get more power generation and enhance the convergence speed. Feng et al. (2017) optimized the operation of hydropower system by proposing a algorithm which combining the merits of DDDP and orthogonal experiment design. In this way, the computing amount can be sharply reduced when the quality of the result is influenced a bit. Tospornsampan et al. (2016) proposed a general operating policy for a multiple reservoir system operation which using the combination of a DDDP and a neural network (NN). The result shows the combination model performs satisfactorily. The previous studies show that DDDP approach is a suitable means to solve the high-dimensional problem and this algorithm has not been applied in the energy conservation optimization of metro systems yet. Therefore, the application of DDDP method in the simultaneous optimization of driving strategy and timetable motivates the study of this paper.

In this paper, we will propose an integrated energy-efficiency optimization model for multi trains which combines the driving strategy and the train timetable. Then a DDDP based algorithm will be designed to solve the proposed model so as to get the global optimal solution with low calculating time and the computer memory requirement. In this way, the energy consumption of the urban rail systems will be reduced from the perspective of the system. The optimal result can be more accurate and the calculation time can be shorter by comparing with the traditional dynamic programming algorithm.

The rest of this paper is organized as follows. In Section 2, the problem is formulated with a coordination optimization model. In Section 3, the solution approach consisting of the STS network methodology and DDDP approach is proposed. In Section 4, the effectiveness and efficiency of the proposed approach are demonstrated in a metro network by comparing with dynamic programming. In Section 5, the main contributions of this paper

are summarized and some future research is discussed.

2 Mathematical Formulations

This paper aims to reduce the energy consumption of train operation by optimizing the train timetable and the driving strategy at the same time. Before introducing the solution approach of this problem, we will create the mathematical model and show the formulations in this section.

2.1 Key parameters

Firstly, for a better understanding of the paper, the key parameters of the model are illustrated in Table 1 and Table 2.

Table 1: Sets and decision variables

Symbol	Description
	Set
N	Set of time sequence
I	Set of trains
K	Set of stations
	Decision variable
$v_i(t_n)$	Speed of train i at time t_n (m/s)
$s_i(t_n)$	Space of train i at time t_n (m)

2.2 Energy consumption without considering RE

In this section, the energy consumption calculation model with considering RE is introduced, in which the net traction energy that defined as the difference between the traction energy and the reused RE is the objective function of this optimization problem.

Specifically, the operating time of trains is divided into many small parts and N is the set of time sequence. By using this method, we will make the calculation of RE transmission process more precise. The amount of utilised RE can be obtained after fixed time to approximate simulate the transmission process of RE. What is more, the total traction energy can also be obtained in this way.

Firstly, for a certain train i, the resistance can be divided into the basic resistance and line resistance. $RB(v_i(t_n))$ is the basic resistance including roll resistance and air resistance, which can be described as

$$RB(v_i(t_n)) = m(a_1 v_i(t_n)^2 + a_2 v_i(t_n) + a_3) \tag{1}$$

What is more, $RC(s_i(t_n))$ is the line resistance caused by track grade, curves and tunnels, which is related to the position of train. Resistance caused by tunnels is always too small compared with resistance caused by track grade and curves, so it can be ignored in the optimization. As a result, $RC(s_i(t_n))$ can be calculated by following equation, in which $\alpha(s_i(t_n))$ and $R(s_i(t_n))$ are the angle of gradient and radius of turning circle, respectively.

$$RC(s_i(t_n)) = mgsin\alpha(s_i(t_n)) + \frac{600}{R(s_i(t_n))} \tag{2}$$

Table 2: Notation

Symbol	Description
n	Time sequence index, $n \in N$
i	Train index, $i \in I$
k	Station index, $k \in K$
t_n	Time stamp
$u_i(t_n)$	Acceleration/deceleration rate of train i at time t_n (m/s^2).
m	Mass of train (kg)
$RB(v_i(t_n))$	Basic resistance of train i at time t_n (N)
$RC(s_i(t_n))$	Line resistance of train i at time t_n (N)
$fr(v_i(t_n), s_i(t_n))$	Total resistance of train i at time t_n
$F_i(t_n)$	Force of train i at time t_n (N)
$FT_i(t_n)$	Tractive effort of train i at time t_n (N)
$ET_i(t_n)$	Traction energy consumption of train i between t_n and t_{n+1} (J)
ET	Total traction energy consumed by trains in whole time range (J)
$FB_{i'}(t_n)$	Braking effort of train i' at time t_n (N)
$EB_{i'}(t_n)$	RE produced by train i' between t_n and t_{n+1} (J)
$w_{i',i}(t_n)$	Factor to measure how much RE transferred from i' to i (J)
$ER_{i',i}(t_n)$	RE allocated from train i' to train i at time t_n(J)
$EL_{i',i}(t_n)$	RE loss during transmission from train $i'$to train i at time t_n (J)
e	RE loss per distance unit (J/m)
$EU_{i',i}(t_n)$	RE absorbed by train i form train i' at time t_n (J)
$EU_i(t_n)$	RE actually utilized by train i (J)
EU	Total RE utilized in whole time range (J)
E	Net energy consumption in whole time range (J)
$d_i(k)$	Departure time for train i at station k (s)
$a_i(k)$	Arrival time for train i at station k (s)
$[\varpi^{min}(k), \varpi^{max}(k)]$	Dwell time threshold for trains at station k (s)
$[\xi^{min}(k), \xi^{max}(k)]$	Trip time threshold between station k and $k+1$ (s)
h^{min}	Minimum headway (s)
$[TC^{min}, TC^{max}]$	Cycle time threshold for trains (s)
v^{max}	Maximum speed (m/s)

In this way, the total resistance can be calculated as

$$fr(v_i(t_n), s_i(t_n)) = RB(v_i(t_n)) + RC(s_i(t_n)) \tag{3}$$

Then, the accelerate rate of train will be solved.

$$u_i(t_n) = \frac{F_i(t_n) - fr(v_i(t_n), s_i(t_n))}{m} \tag{4}$$

The total traction energy between two adjacent time t_n and t_{n+1} can be obtained by Equation (5) and (7), in which the traction force $FT_i(t_n)$ is the larger value between the actual force applied to the train $F_i(t_n)$ and 0.

$$FT_i(t_n) = \max\{u_i \cdot m + fr(v_i(t_n), s_i(t_n)), 0\} \tag{5}$$

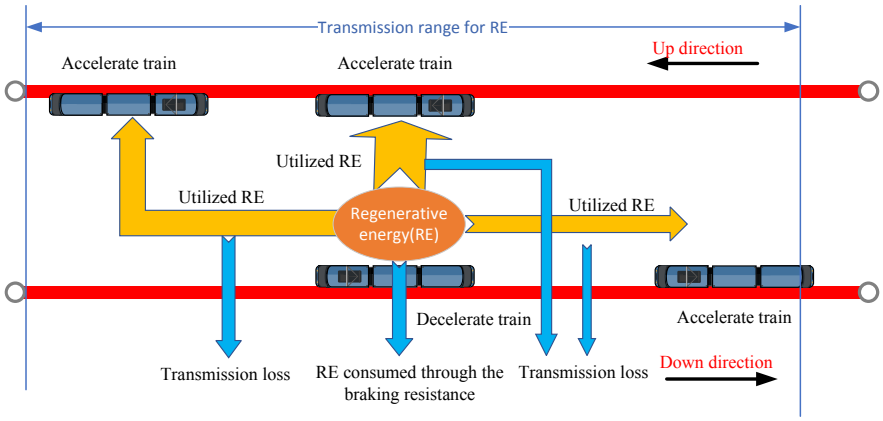


Figure 1: RE transmission schematic diagram

$$ET_i(t_n) = \int_{t_n}^{t_{n+1}} FT_i(t) \cdot v_i(t) dt \tag{6}$$

By adding up the traction energy of all trains during each periods of time, the total traction energy can be obtained by Equation (7).

$$ET = \sum_{i=1}^{I} \sum_{n=1}^{N-1} ET_i(t_n) \tag{7}$$

2.3 Energy consumption with considering RE

In addition to the traction energy during operation, RE produced by braking trains also need to be considered to realize energy efficient operation. Figure 1 is the schematic diagram of the RE transmission principle. It shows the RE produced by decelerate trains will be divided into all the accelerate trains within the energy transfer range. The amount of RE absorbed by each train is related to the distance between two trains as well as the voltage difference in the traction power grid. The energy loss during transmission and rest energy consumed by braking resistance should also be considered. The corresponding model is listed as follows:

The RE produced by train i' can be calculated by following equations. Some of the braking power will be consumed and cannot be transformed to RE. φ is the energy conversion rate in this process.

$$FB_{i'}(t_n) = -\min\{u_{i'} \cdot m + fr(v_{i'}(t_n), s_{i'}(t_n)), 0\} \tag{8}$$

$$EB_{i'}(t_n) = \int_{t=t_n}^{t_{n+1}} \varphi \cdot FB_{i'}(t) \cdot v_{i'}(t) dt \tag{9}$$

According to the actual principle, the RE produced by train i' can be used by all nearby accelerating trains, whether running in the same direction or negative direction. In this

paper, the formulas of RE transmission are assumed as Equation (10), (11), (12). Factor $w_{i',i}(t_n)$ is set to measure how much RE transferred to train i from i' at time t_n. If the distance between two trains is farther than transmission range z, the value of $w_{i',i}(t_n)$ is 0. Otherwise, the value is related to distance and voltage difference between two trains. The larger the distance or the smaller the voltage difference is, the smaller the value of $w_{i',i}(t_n)$ is. Equation (10) is used to model the RE utilization, in which s_0 and E_0 are fixed values.

$$w_{i',i}(t_n) = \begin{cases} 1/\left(\frac{|s_i(t_n)-s_{i'}(t_n)|}{s_0} - \frac{max\{0,\Delta ET_i(t_n)-\Delta EB_{i'}(t)\}}{E_0}\right), & |s_{i'}(t_n) - s_i(t_n)| \leq z; \\ 0, & otherwise. \end{cases} \tag{10}$$

By considering all trains in the line, the RE allocated from train i' to train i can be calculated as

$$ER_{i',i}(t_n) = EB_{i'}(t_n) \times \frac{w_{i',i}(t_n)}{\sum_{j=1}^{I} w_{i',j}(t_n)} \tag{11}$$

What is more, the energy loss during transmission can be calculated with the Equation (12), which is related to the distance between trains. Then, the energy absorbed by train i can be obtained as Equation (13).

$$EL_{i',i}(t_n) = e \times |s_i(t_n) - s_{i'}(t_n)| \tag{12}$$

$$EU_{i',i}(t_n) = ER_{i',i}(t_n) - EL_{i',i}(t_n) \tag{13}$$

Further more, it is possible that not all the RE will be utilized by train i, because the energy absorbed may be more than traction energy required. In this condition, the redundant energy will be consumed by the braking resistance. As a result, the RE actually utilized by train i can be calculated as

$$EU_i(t_n) = \min\{\sum_{j=1}^{I} EU_{j,i}(t_n), ET_i(t_n)\} \tag{14}$$

By adding up the RE utilized by all trains during each periods of time, the total RE utilized during the whole time can be obtained by Equation (15).

$$EU = \sum_{i=1}^{I} \sum_{n=1}^{N-1} EU_i(t_n) \tag{15}$$

Finally, the objective function of this problem can be obtained by following formula, which is the result of total traction power minus total RE.

$$E = ET - EU \tag{16}$$

2.4 Optimization model

The optimization model is formulated as below, which includes the objective function of this problem and two kind of constraints.

Minimize

$$E = ET - EU$$

Subject to

$$\varpi^{min}(k) \leq d_i(k) - a_i(k) \leq \varpi^{max}(k); \quad \forall 1 \leq i \leq I, 1 \leq k \leq K \tag{17}$$

$$\xi^{min}(k) \leq a_i(k+1) - d_i(k) \leq \xi^{max}(k); \quad \forall 1 \leq i \leq I, 1 \leq k \leq K-1 \tag{18}$$

$$d_i(k) - d_{i-1}(k) \geq h^{min}; \quad \forall 2 \leq i \leq I, 1 \leq k \leq K \tag{19}$$

$$a_i(k) - a_{i-1}(k) \geq h^{min}; \quad \forall 2 \leq i \leq I, 1 \leq k \leq K \tag{20}$$

$$TC^{min} \leq d_i(K) - a_i(1) \leq TC^{max}; \quad \forall 1 \leq i \leq I \tag{21}$$

$$FT_i(t_n) \leq FT^{max}; \quad \forall 1 \leq i \leq I, 1 \leq n \leq N-1 \tag{22}$$

$$FB_i(t_n) \leq FB^{max}; \quad \forall 1 \leq i \leq I, 1 \leq n \leq N-1 \tag{23}$$

$$0 \leq v_i(t_n) \leq v^{max}; \quad \forall 1 \leq i \leq I, 1 \leq n \leq N \tag{24}$$

Among above constraints, Formula (17) and Formula (18) restrict the dwell time and travel time of trains. In Formula (17), $\varpi^{min}(k)$ and $\varpi^{max}(k)$ are both determined by the station's condition and the number of passengers. In Formula (18), $\xi^{min}(k)$ is determined by the accelerating and decelerating ability of trains, the length of segments and the speed limits of segments. $\xi^{max}(k)$ is determined by the planned train timetable. As for headway limits, the headway between adjacent trains should be within the given range, which are shown in Formula (19) and Formula (20). What is more, Formula (21) is the constraint for the cycle time of trains during operation. In summary, Formula (17)- (21) are the constraints of the train timetable.

On the other hand, Formula (22) and (23) restrict the maximum traction and braking force of trains, which are represented by FT^{max} and FB^{max}. Formula (24) is the speed limit constraint in order to ensure the safe operation of trains. These three inequations are the constraints of the driving strategy.

3 Solution approach

In order to obtain the optimal value of objective function when all the constraints are satisfied, the STS network methodology is applied to discretize the state variables and the DDDP approach is used to solve this problem. This section will simply introduce the theory of the STS and DDDP method and show how to realize energy efficient optimization by using this method.

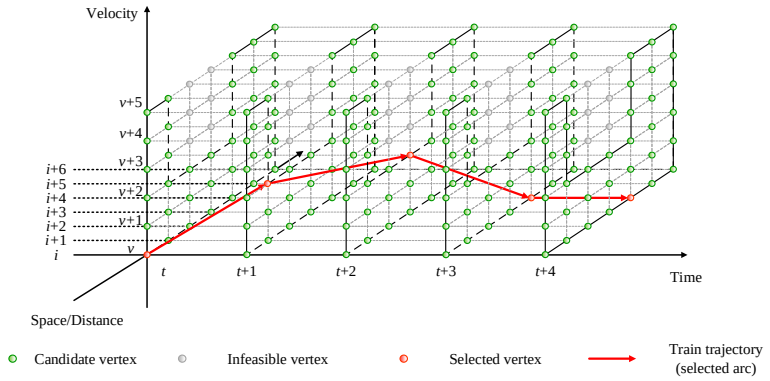


Figure 2: STS network for the integrated train operation problem

3.1 An overview of STS

The theory of STS network methodology is discretizing the space, velocity and operating time of trains to construct a large number of cells as shown in Figure 2, then the optimal operating route of trains can be selected flexibly in the network such as the red curve in the figure(Zhou et al. (2017)). STS network methodology mainly has following advantages: Firstly, it discretizes the problem into a multi-step decision process, which is suitable for solving the problem by dynamic programming and its improved algorithm, such as the DDDP approach in this paper. Secondly, in the STS networks, the shadow of the train route in the Space-Time side is the driving curve of trains and the shadow in the Space-Time side is the train timetable. Finally, the integrated optimization to reduce energy consumption can be realized by systematically incorporating the Space-Time (train timetable) model and the Space-Speed (driving strategy) model into the STS network.

As a result, the STS network methodology has been widely used in the transportation route optimization, various scheduling applications and general dynamic network flow modeling.

3.2 An overview of DDDP

DDDP is an improved DP method to overcome the ”curse of dimensionality” by reducing the computational dimension. The principle of DDDP approach is dividing the solution space of problem into several subspaces and obtain the best local optimal solution in each iteration. By repeating this process, the global optimal solution can be solved. The schematic diagram of DDDP is shown as Figure 3 and the general procedures of DDDP are presented as follows (Heidari et al. (1971), Li et al. (2014), Feng et al. (2017)):

- The initial test trajectory which satisfies the constraints can be obtained by experience or other methods, shown as the red curve in Figure 3.
- In the neighborhood of the test trajectory, the solution space of problem at each stage can be separated into several subspaces and combined to form the corridor,

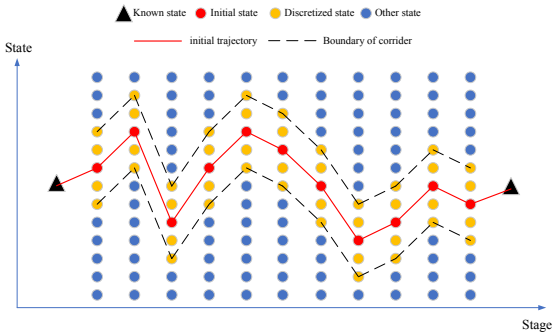


Figure 3: Schematic diagram of DDDP calculating process

- DP recursive equation is used to find an improved trajectory in the corridor,
- The optimal trajectory in the last iteration will be taken as the initial trajectory of the next iteration,
- Repeat iteration until the convergence condition is met.

It can be seen that, compared with the DP, DDDP method does not need to optimize in the entire feasible region of state variables, but only find the optimal solution within the corridor each time, which is a small range compared with the former. In this way, by using DDDP method we can effectively reducing the computational storage and time.

For example, we assume that in a optimization problem there are τ stages, κ state variables and ρ states of each variables in each stage. If we solve the problem by using DP method, the space and time complexity are $\tau\kappa^{\rho}$ and $\tau\kappa^{2\rho}$, respectively. If the number of variables or states is quite some, the computational amount will be too large to calculate. However, by using DDDP method, the space and time complexity can only reach $\tau\lambda^{\rho}$ and $\tau\lambda^{2\rho}$ by setting the number of states in corridor is λ in each stage. In this way, the dimension of problem will be reduced and the computational storage and time can decrease a lot by choosing λ as a small number. Energy efficient problem for multi-trains will have more state variables with the increase of train number, so it can get better effect of dimensionality reduction by using DDDP method.

3.3 Procedures to solve energy-efficient train operation problem

In this section, we will solve the multi-trains energy-efficient problem by using a combination of STS and DDDP method. The solving thought is discretizing the state variables of problem to create a multi-step decision process firstly, then obtain the global optimal solution by repeating solving the best solution in corridor selected. The detail procedures are presented as follows:

(i) Step 1: Discretization-Before optimizing, state variables of train should be discretized and STS network should be constructed firstly by using STS method. In this problem, operating time is the stage which can be represented by (t_a), a is the index of stage. What is more, the state variables in each stage conclude the speed and space of each

train. If there are I trains operating in metro line at the same time, there will be $2I$ state variables in each stage, which can be represented by $\Psi_A(v_I, s_I)$. Meanwhile, for a certain train i, $\Psi_a(v_i, s_i)$ is the speed and space variables in stage a. In this way, we can obtain the best operating trajectory for each train by choosing optimal states in each stage.

(ii) Step 2: Parameters intialization-Set the basic parameters of DDDP method, including time interval between stages, initial speed and space interval in each stage, terminal condition and so on.

(iii) Step 3: Establish the initial trajectory-Randomly generate a feasible solution of this problem which satisfies all the constraints. The initial trajectory can be listed by $\Psi_A(v_I^0, s_I^0)$.

(iv) Step 4: Create optimizing corridor for initial trajectory-Firstly, we use Δv^0 and Δs^0 to represent initial speed interval and space interval. L is the width of corridor which can only be even number. In this way, there will be $L+1$ kinds for each state variable. Then, the corridor for initial trajectory can be constructed as Formula (25) for $i \in I, a \in A$. The best local solution in the first iteration will be selected from the states in this corridor.

$$\begin{gathered}
\Psi_a(v_i^0 - \tfrac{L}{2}\Delta v_0, s_i^0 - \tfrac{L}{2}\Delta s_0) \\
\Psi_a(v_i^0 - (\tfrac{L}{2}-1)\Delta v_0, s_i^0 - (\tfrac{L}{2}-1)\Delta v_0) \\
\cdots \\
\Psi_a(v_i^0, s_i^0) \\
\cdots \\
\Psi_a(v_i^0 + (\tfrac{L}{2}-1)\Delta v_0, s_i^0 + (\tfrac{L}{2}-1)\Delta v_0) \\
\Psi_a(v_i^0 + \tfrac{L}{2}\Delta v_0, s_i^0 + \tfrac{L}{2}\Delta s_0)
\end{gathered} \tag{25}$$

(v) Step 5: Calculation-Search the best local solution in the current corridor by using traditional DP method as following process:

- Determine the driving strategy of arcs. Given the starting point $\Psi_a(v_i, s_i)$ and the end point $\Psi_{a+1}(v_i, s_i)$ of a arc of train i in STS networks, we can determine the driving strategy of the train between these two points by the following means. Firstly, three reference states $\Psi_{a+1}(v_i^{ma}, s_i^{ma})$, $\Psi_{a+1}(v_i^{co}, s_i^{co})$, $\Psi_{a+1}(v_i^{mb}, s_i^{mb})$ in stage $a+1$ which represent the train operates with maximum acceleration, coasting and maximum deceleration between stage a and $a+1$ need to be calculated according to $\Psi_a(v_i, s_i)$. Next, by comparing $\Psi_{a+1}(v_i, s_i)$ and reference states, the driving strategy of arc will be chosen on the basis of Table 3. As shown in the table, there are 5 feasible driving strategy and other conditions will be eliminated. In this way, the equation of each train's curve between adjacent stages will be solved.
- Calculate the cost of each set of arcs. In this problem, the cost of each set of arcs represents the net energy consumption of trains during this process. By given

Table 3: Driving strategy for each kind of arcs

Condition	Driving strategy
$v_i = v_i^{ma}, s_i = s_i^{ma}$	Maximum acceleration
$v_i^{co} < v_i < v_i^{ma}, s_i^{co} < s_i < s_i^{ma}$	Partial acceleration-coasting
$v_i = v_i^{co}, s_i = s_i^{co}$	Coasting
$v_i^{mb} < v_i < v_i^{co}, s_i^{mb} < s_i < s_i^{co}$	Partial deceleration-coasting
$v_i = v_i^{mb}, s_i = s_i^{mb}$	Maximum deceleration

the driving curve of each operating train between continuous stage a and $a+1$ by using above the means, the cost of certain set of arcs $U(\Psi_{a,a+1}(v_I, s_I))$ can be calculated by Formulas (1)-(16). If there are one or more arc is not attainable, the cost of this set will be ∞.

- Choose the best set of arcs and eliminate others. After calculating the cost of the whole sets of arcs between two stages, it is necessary to choose the best set for each states in stage $a+1$ according to Bellman Equation to decrease the calculated amount. Up to stage a, the total cost of all the previous steps is $J(\Psi_a(v_I, s_I))$. Then, $J(\Psi_{a+1}(v_I, s_I))$ can be calculated by Equation (26).

$$J(\Psi_{a+1}(v_I, s_I)) = J(\Psi_a(v_I, s_I)) + U(\Psi_{a,a+1}(v_I, s_I)) \tag{26}$$

Finally, the best set of arcs with lowest cost for each states in stage $a+1$ will be selected according to Equation (27), in which $\Psi_a(v_I^*, s_I^*)$ is the chosen state to match $\Psi_{a+1}(v_I, s_I)$. In addition, other sets of arcs should be eliminated.

$$\Psi_a(v_I^*, s_I^*) = J(\Psi_a(v_I^*, s_I^*)) + U(\Psi_{a,a+1}(v_I, s_I)) \tag{27}$$

- Repeat above steps of DP until all the possible states in corridor have been selected and obtain the best local optimal solution in corridor which can be represented by Υ^ρ for the ρth iteration.

(vi) Judge the best local solution and adjust parameters-Compare Υ^ρ with $\Upsilon^{\rho-1}$ starting with the second iteration and adjust the speed and space intervals in $\rho+1$th iteration according to difference value of costs as Formula (28).

$$\begin{cases} \Delta v^{\rho+1} = \Delta v^\rho - \delta(v), \Delta s^{\rho+1} = \Delta s^\rho - \delta(s); & \Upsilon^\rho - \Upsilon^{\rho-1} \geq \delta(\Upsilon) \; and \; \rho \geq 2 \\ \Delta v^{\rho+1} = \Delta v^\rho, \Delta s^{\rho+1} = \Delta s^\rho; & \Upsilon^\rho - \Upsilon^{\rho-1} < \delta(\Upsilon) \; or \; \rho = 1 \end{cases} \tag{28}$$

(vii) Create new corridor for improved the trajectory. The improved trajectory in the ρth iteration will be the initial trajectory in the $\rho+1$th iteration. Then the corridor in $\rho+1$th iteration can be constructed as Formula (25).

(viii) DDDP iteration. Repeat step (v)-step (vii) until the terminal condition which is listed as Formula (29) is met. $\underline{v}$ is a fixed valve which means the minimal speed interval.

$$\Delta v^{\rho+1} < \underline{v} \tag{29}$$

4 Case study

In order to illustrate the effectiveness of the proposed model and numerical algorithm, two numerical examples are conducted based on a small metro network (3 stations, 2 segments, and 2 turn-back stations). The operation requirement and basic infrastructure data about this metro network is described in Table 4. The trains operate in the network are composed of six cars and three of them are traction units. The length of trains is 114m and the net train mass is 192000kg. As the important parameters to calculate the energy consumption, the mass of trains with passengers is set as 250000kg, energy loss per distance unit e is 100J/m, the conversion rate from braking energy to RE φ is 0.5 and the transmission range of RE z is 1500m. As for the parameters for DDDP method, the initial speed interval Δv^0 and space interval Δs^0 are 20m/s and 20m, respectively. The time interval between two adjacent stages is 15s and the minimal speed interval $\underline{v}$ is 10^6J. In the aspect of computer configuration, the algorithms described in this paper were implemented in MATLAB R2014a. What is more, the operating system is Windows 7 professional, the CPU consists of one Intel Core i7-7700@3.6GHz and the memory size 16 GB.

Table 4: Operation requirement and basic infrastructure data of metro network

Destination	**Distance**	T_{min}	T_{max}
Turn-back station 1	0m	-	-
Station 1	667m	72s	102s
Station 2	1700m	67s	97s
Station 3	3020m	80s	110s
Turn-back station 2	3687m	72s	102s

In Example 1, we apply DDDP approach to optimize the energy-efficient driving strategy as well as train timetable by considering there are 2 trains operate in the metro network and compare the result with DP approach. In Example 2, we apply DDDP approach to the same problem with considering 3 running trains and contrast the solution with DP approach.

4.1 Example 1

In example 1, there are 2 trains whose original locations and directions are shown in Figure 4 operate in the line. In order to test the performance of DDDP approach, we optimize the problem by using traditional DP approach firstly as control group, in which the speed interval and space interval are 10 m/s and 10m, respectively. The Space-Time-Speed diagram after optimization is shown as Figure 5 and the result data is shown in Table 5. Then the energy conservation issue is solved by DDDP approach: The Space-Time-Speed diagram, Space-Time diagram and Speed-Time diagram are presented in Figure 6 and Figure 7, respectively. In the figures, curves with different colors represent operating trajectories of different trains. Space-Time diagram can be regarded as the train timetable and Speed-Time diagram can be seen as the driving strategy of trains. Although the driving strategy of each train as shown in Figure 7 may not be the optimal when only one train is taken into consideration, the actual energy consumption of all trains will be optimal with considering the regenerative energy. What is more, the simulating data by using DDDP method is listed in Table 5.

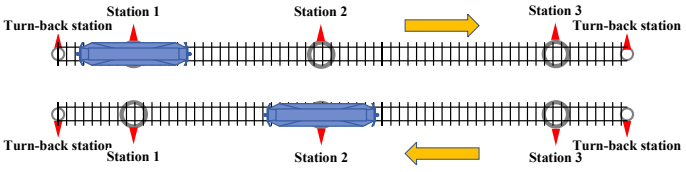


Figure 4: Original location and direction of 2 trains

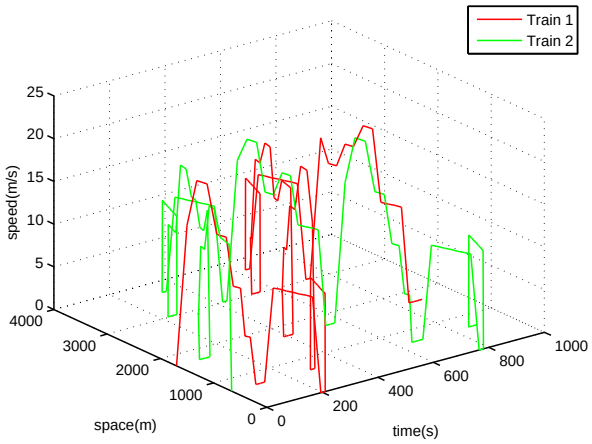


Figure 5: Space-Time-Speed diagram of 2 trains by using DP method

By comparing the data in Table 5, the total energy consumption and net energy consumption are reduced by 7.8% and 12.2%. What is more, the reused RE utilized by trains increases from 9.62 kW·h to 15.55 kW·h, in which the increase rate is 61.6 %. It is because the integrated optimum design can decrease the tractive energy needed and improve the utilization rate of RE at the same time. By using DDDP method, the obtained solution will be more accurate and the energy-saving effect is better. As for computing time, it only costs 95 seconds to calculate the problem by using DDDP approach, which is reduced by 73.9 % compared with DP's time.

More specifically, the contrast picture of the result of DP and DDDP method is shown as Figure 8, in which the computing time is plotted on the horizontal axis, the net energy consumption and variable interval are plotted on the primary vertical axis, the speed and space interval of DDDP are plotted on the secondary vertical axis. From Figure 8, the quality of DDDP's solution is not better than DP's in the first few times iterations. However, the net energy consumption after optimization by using DDDP will be less than DP's since the 16th iteration and the difference will be more and more with the passage of computing time. As a result, we can get better optimal solution in shorter time by applying DDDP approach.

Table 5: Data of energy consumption and computing time of 2 trains

Method	Total consumption	Utilized RE	losing RE	Net consumption	Computing time
DP method	162.91 kW·h	9.62 kW·h	1.82 kW·h	153.29 kW·h	364 s
DDDP method	150.18 kW·h	15.55 kW·h	1.81 kW·h	134.63 kW·h	95 s

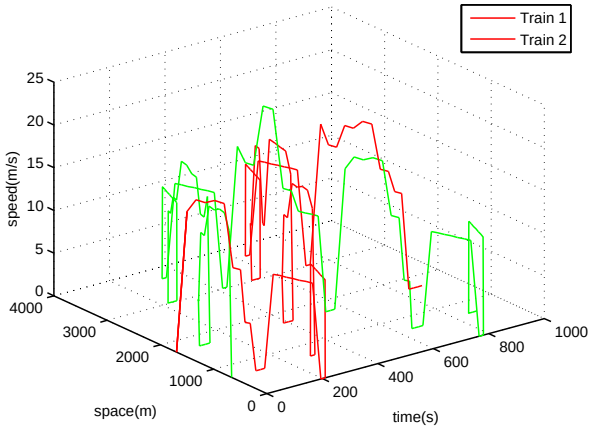


Figure 6: Space-Time-Speed diagram of 2 trains by using DDDP method

4.2 Example 2

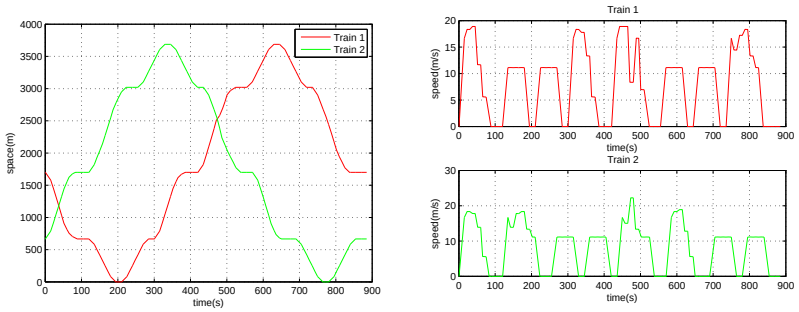
In this case study, DP and DDDP are applied to the same metro network with considering 3 running trains as shown in Figure 9. By using DP approach, the Space-Time-Speed diagram after optimization is shown as Figure 10 and the result data is shown in Table 6. As for the result of solving by DDDP approach, Figure 11 and Figure 12 show the Space-Time-Speed diagram, Space-Time diagram and Speed-Time diagram. The simulating data is also listed in Table 6.

On the one hand, by calculating the total energy consumption and net energy consumption are reduced by 11.8% and 20.3%. Also, the reused RE utilized by trains increases from 14.12 kW·h to 32.96 kW·h, in which the growth rate is 133.4 %. On the other hand, the computing time decreases from 69779 seconds to 14625 seconds. The droop rate is 79.0 %.

Table 6: Data of energy consumption and computing time of 3 trains

Method	Total consumption	Utilized RE	losing RE	Net consumption	Computing time
DP method	257.51 kW·h	14.12 kW·h	12.25 kW·h	243.39 kW·h	69779 s
DDDP method	227.02 kW·h	32.96 kW·h	1.26 kW·h	194.06 kW·h	14625 s

The result contrast picture of DP and DDDP method when considering 3 trains is shown as Figure 13. From the figure, we can find the net energy consumption after optimization by using DDDP will be less than DP's since the 16th iteration and the gap of difference will increase in the later iterations.



(a) Space-Time diagram (b) Speed-Time diagram

Figure 7: Space-Time and Speed-Time diagram of 2 trains by using DDDP method

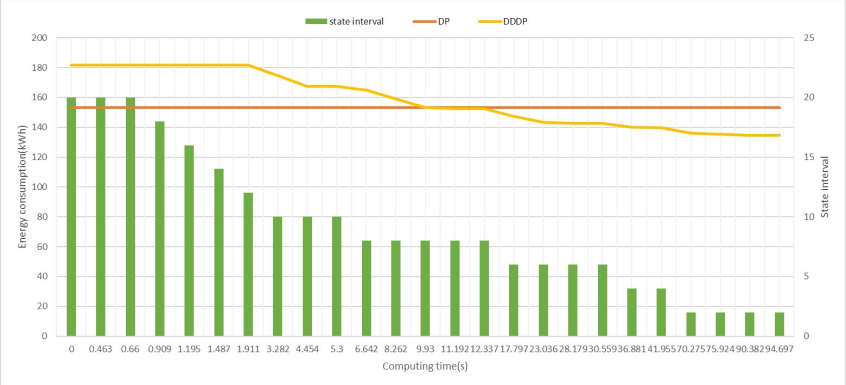


Figure 8: Result contrast picture of DP and DDDP method with considering 2 trains

4.3 Summary of experiment results

According to the above result data, the solution of DDDP can produce the lower energy consumption and make the better use of the RE then DP no matter when 3 trains or 2 trains operate in the line. What is more, the computing time is also much lower because DDDP method reduces the number of feasible states for each variables per phase. In summary, it is accurate as well as efficient to use this method in multi-train energy efficient problem.

We can also discover some information by comparing the results of two examples. Firstly, by using DDDP method, the energy-saving effect and RE utilization effect are more obvious with the increase of train number. It is because DDDP's state intervals of each train are more and more precise with the iterations. In this way, each additional train will increase the overall search accuracy to a greater extent compared to other algorithm. In addition, the decreasing amplitude of computing time compared to DP will be larger when more trains operate in the line. The reason is the calculated amount of each train is fewer by using DDDP approach and the amplitude reduction ratio of the whole computation is much less with the increase of exponent.

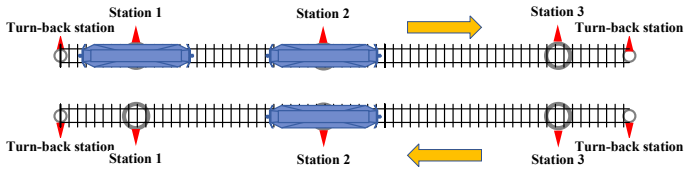


Figure 9: Original location and direction of 3 trains

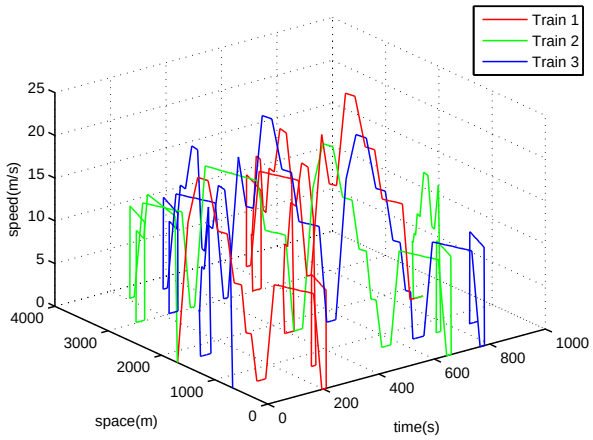


Figure 10: Space-Time-Speed diagram of 3 trains by using DP method

In summary, multi-trains optimization for energy efficiency by using DDDP approach is high-efficiency in terms of the energy-saving effect and the computing time. What is more, the optimizing performance will be better when more trains taken into considered.

5 Conclusions

In this study, a new efficient technique employing DDDP is presented to address the energy-efficient metro train operation problem with considering the RE. Firstly, the objective function and constraints are formulated to construct the mathematical model of the problem. Then, the solution approach is proposed, in which the state variables of trains in the proposed model are discretized by STS network methodology and optimized by DDDP approach. Finally, two numerical experiments are simulated to test the potential ability of DDDP when multi trains operate in the small network. The simulation results show that, the net energy consumption of trains by using DDDP is 12.2 % and 20.3 % lower compared with DP when 2 and 3 trains operate in the line, respectively. Besides, the use ratio of RE increases by 61.6 % and 133.4 %. The computing time of simulation also decreases by 73.9 % and 79.0 % in two cases. As a result, DDDP approach applied in this paper is an effective way to obtain more exact solution in shorter time. What is more, the effect of dimension-

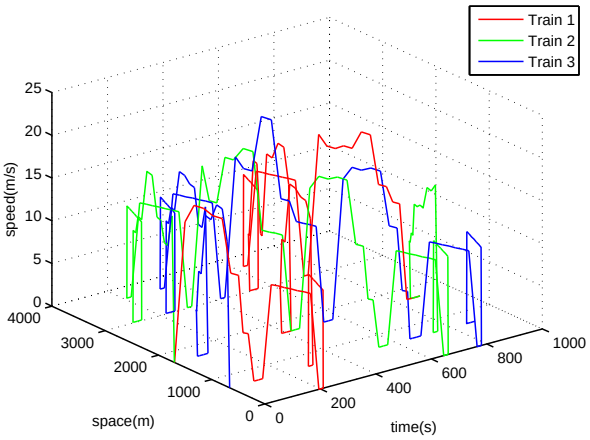
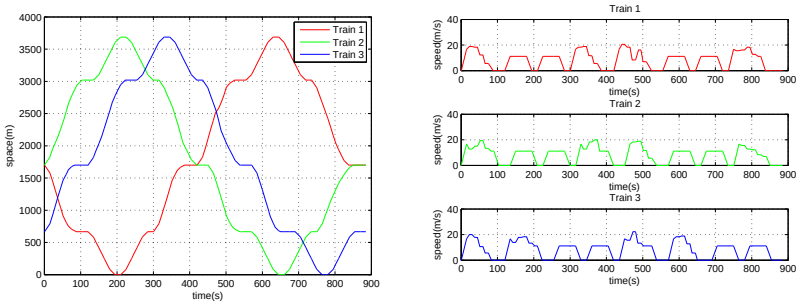


Figure 11: Space-Time-Speed diagram of 3 trains by using DDDP method

ality reduction by using DDDP can be more obvious compared to DP when more trains are considered in the optimization according to the result data, which will further prove the great performance of DDDP in the integrated optimization of the driving strategy and the train timetable.

However. the ability of DDDP approach remains limited in the optimization when large number of trains operate in the metro network and further design of algorithm is still necessary. In the future, we will focus on the reseach of parallel algorithms, in order to redouble reduce the computing time of optimization.



(a) Space-Time diagram (b) Speed-Time diagram

Figure 12: Space-Time and Speed-Time diagram of 3 trains by using DDDP method

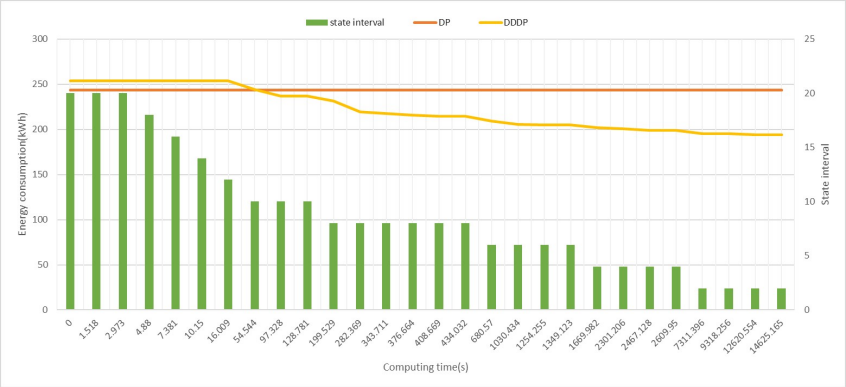


Figure 13: Result contrast picture of DP and DDDP method with considering 3 trains

6 Acknowledgements

This work was supported by Beijing Laboratory of Urban Rail Transit, Beijing Key Laboratory of Urban Rail Transit Automation and Control, by the National Natural Science Foundation of China (No. 61803021, U1734210) and the Beijing Natural Science Foundation "The Joint Rail Transit" (No. L171007).

References

Bocharnikov, J., Tobias, A., Roberts, C., 2010. "Reduction of train and net energy consumption using genetic algorithms for trajectory optimisation", *IET Conference on Railway Traction Systems (RTS 2010)*, Birmingham, UK.

Bu, B., Qin, G., Li , L., Li, G.,2018. "An Energy Efficient Train Dispatch and Control Integrated Method in Urban Rail Transit", *Energies*, vol. 11, no. 5.

Chang, C., Sim, S., 2008. "Optimising train movements through coast control using genetic algorithms", *IEE Proceedings-Electric Power Applications*, vol. 144, no. 1, pp. 65–73.

Chang, C., Sim, S., 2008. "Optimising train movements through coast control using genetic algorithms", *IEE Proceedings-Electric Power Applications*, vol. 144, no. 1, pp. 65–73.

Feng, Z., Niu, W., Cheng, C., Liao, S., 2017. "Hydropower system operation optimization by discrete differential dynamic programming based on orthogonal experiment design", *Energy*, vol. 126, pp. 720–732.

Heidari, M., Chow, V., Kokotovic, P., Meredith, D., 1971. "Discrete differential dynamic programing approach to water resources systems optimization", *Water Resources Research*, vol. 7, pp. 273–282.

Ishikawa, K., 1968. "Application of optimization theory for bounded state variable problems to the operation of trains", *Bulletino of JSME*, vol. 11, no. 47, pp. 857–865.

Keskin, K., Karamancioglu, A., 2017. "Energy-Efficient Train Operation Using Nature-Inspired Algorithms", *Journal of Advanced Transportation*, pp. 1–12.

Khmelnitsky, E., 2000. "On an Optimal Control Problem of Train Operation", *IEEE Trans-*

actions on Automatic Control, vol. 45, no. 7, pp. 1257–1266.

Li, C., Zhou, J., Ouyang, S., DingX, D., Chen, L., 2014. "Improved decomposition coordination and discrete differential dynamic programming for optimization of large scale hydropower system", *Energy Conversion & Management*, vol. 84, pp. 363–373.

Li, X., Lo, H., 2014. "An energy-efficient scheduling and speed control approach for metro rail operations", *Transportation Research Part B*, vol. 64, pp. 73–89.

Liu, R., Golovitcher, I., 2003. "Energy-efficient operation of rail vehicles", *Transportation Research Part A*, vol. 37, pp. 917–932.

Ning, J., Zhou, Y., Long, F., Tao, X., Lund, H., Kaiser, M., 2018. "A synergistic energy-efficient planning approach for urban rail transit operations", *Energy*, vol. 151, pp. 854–863.

Rodrigo, E., Tapia, J., Mera, J., Soler, M., 2013. "Optimizing Electric Rail Energy Consumption Using the Lagrange Multiplier Technique,", *Journal of Transportation Engineering-ASCE*, vol. 139, no. 3, pp. 321–329.

Su, S., Li, X., Tang, T., Gao, Z., 2013. "A Subway Train Timetable Optimization Approach Based on Energy-Efficient Operation Strategy", *IEEE Transactions on Intelligent Transportation Systems*, vol. 14, no. 2, pp. 883–893.

Su, S., Tang, T., Roberts, C., 2015. "A Cooperative Train Control Model for Energy Saving", *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 2, pp. 622–631.

Tian, Z., Weston, N., Zhao, N., Hillmansen, S., Roberts, C., Chen, L., 2017. "System energy optimisation strategies for metros with regeneration", *Transportation Research Part C*, vol. 75, pp. 120–135.

Tospornsampan, J., Kita, I., Ishii, M., Kitamura, Y., 2016. "Discrete differential dynamic programming and neural network on deriving a general operating policy of a multiple reservoir system : a case study in the mae klong system, thailand", *Journal of Rainwater Catchment Systems*, vol. 11, no. 2, pp. 1–9.

Yin, J., Tang, T., Yang, L., Gao, Z., Ran, B., 2016. "Energy-efficient metro train rescheduling with uncertain time-variant passenger demands: An approximate dynamic programming approach", *Transportation Research Part B*, vol. 91, pp. 178–210.

Zhou, L., Tong, L., Chen, J., Tang, J., Zhou, X., 2017. "Joint optimization of high- speed train timetables and speed profiles: A unified modeling approach using space-time-speed grid networks", *Transportation Research Part B*, vol. 97, pp. 157–181.

Zhou, Y., Bai, Y., Li , J., Mao, B., Li, T.,2018. "Integrated Optimization on Train Control and Timetable to Minimize Net Energy Consumption of Metro Lines", *Journal of Advanced Transportation*, pp. 1–19.

Distributed optimization approaches for the integrated problem of real-time railway traffic management and train control

Xiaojie Luan [a,1], Bart De Schutter [b], Ton van den Boom [b], Lingyun Meng [c], Gabriel Lodewijks [d], Francesco Corman [e]

[a] Section Transport Engineering and Logistics, Delft University of Technology
Mekelweg 2, 2628 CD, Delft, the Netherlands
[1] E-mail: x.luan@tudelft.nl, Tel.: +31 (0) 15 27 87294
[b] Delft Center for Systems and Control, Delft University of Technology
[c] State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University
[d] School of Aviation, Faculty of Science, University of New South Wales
[e] Institute for Transport Planning and Systems (IVT), ETH Zürich

Abstract

This paper introduces distributed optimization approaches, with the aim of improving the computational efficiency of an integrated optimization problem for large-scale railway networks. We first propose three decomposition methods to decompose the whole problem into a number of subproblems, namely a geography-based (GEO), a train-based (TRA), and a time-interval-based (TIN) decomposition respectively. As a result of the decomposition, couplings exist among the subproblems, and the presence of these couplings leads to a non-separable structure of the whole problem. To handle this issue, we further introduce three distributed optimization approaches. An Alternating Direction Method of Multipliers (ADMM) algorithm is developed to solve each subproblem through coordination with the other subproblems in an iterative manner. A priority-rule-based (PR) algorithm is proposed to sequentially and iteratively solve the subproblems in a priority order with respect to the solutions of the other subproblems solved with a higher priority. A Cooperative Distributed Robust Safe But Knowledgeable (CDRSBK) algorithm is presented, where four types of couplings are defined and each subproblem is iteratively solved together with its actively coupled subproblems. Experiments are conducted based on the Dutch railway network to comparatively examine the performance of the three proposed algorithms with the three decomposition methods, in terms of feasibility, computational efficiency, solution quality, and estimated optimality gap. Overall, the combinations GEO-ADMM, TRA-ADMM, and TRA-CDRSBK yield better performance. Based on our findings, a feasible solution can be found quickly by using TRA-ADMM, and then a better solution can be potentially obtained by GEO-ADMM or TRA-CDRSBK at the cost of more CPU time.

Keywords

Distributed optimization, Decomposition, Integration of real-time traffic management and train control, Mixed-integer linear programming (MILP), Large-scale

1 Introduction

Real-time traffic management is of great importance to limit the negative consequences caused by perturbations occurring in real-time railway operations. Train control problem reflects the traffic control by defining speed profiles to let the delayed trains reach the stations at the times specified by the traffic management problem. Due to the real-time nature, a solution is required in a very short computation time for dealing with delayed and canceled train services and for evacuating delayed and stranded passengers as quickly as possible.

The real-time traffic management problem has been studied extensively in the literature, and we refer to the review papers by Cacchiani et al. (2014) and Corman and Meng (2015). There are many optimization approaches available for the railway traffic management problem, using different formulation methods, e.g., the alternative-graph-based method by D'Ariano et al. (2007) and the cumulative flow variable based method by Meng and Zhou (2014), and having different focuses, e.g., considering multiple classes of running traffic in Corman et al. (2011) and integrating train control in Luan et al. (2018). These approaches often lead to large and rather complex optimization problems, especially when considering microscopic details or when integrating traffic management with other problems (e.g., train control problem). They mostly have excellent performance on small-scale cases, where optimality can be achieved in a short computation time. However, when enlarging the scale of the case, the computation time for finding a solution or for proving the optimality of a solution increases exponentially in general.

Distributed optimization approaches have gained a lot of attention to face the need for fast and efficient solutions for problems arising in the context of large-scale networks, such as utility maximization problems. We refer to Nedic and Ozdaglar (2010) and Meinel et al. (2014) for more details. The main idea is to solve the problems either serially or in parallel to jointly minimize a separable objective function, usually subject to coupling constraints that force the different problems to exchange information during the optimization process. In the literature, these approaches have been widely studied in many fields. In transportation systems, they have been explored for controlling road traffic (Findler and Stapp, 1992), for managing air traffic (Wangermann and Stengel, 1996), and for railway traffic (Kersbergen et al., 2016). Kersbergen et al. (2016) focused on the railway traffic management problem with macroscopic details and considered a geography-based decomposition. Lamorgese et al. (2016) proposed a Benders'-like decomposition within a master/slave scheme to address the train dispatching problem. The master and the slave problems correspond to a macroscopic and microscopic representation of the railway.

Bad computational efficiency is one limitation that (integrated) optimization approaches have for large-scale networks. Overcoming this limitation will promote the application of such optimization approaches in practice. Thus, we aim at improving the computational efficiency of solving such (integrated) optimization problems by using distributed optimization approaches. The optimization problem that we focus on in this paper is a mixed-integer linear programming (MILP) problem, developed in our previous work (Luan et al., 2018), where the traffic-related variables (i.e., a set of times, orders, and routes to be followed by trains) and the train-related variables (i.e., speed trajectories) are optimized simultaneously.

In this paper, we consider three decomposition methods, namely a geography-based (GEO) decomposition, a train-based (TRA) decomposition, and a time-interval-based (TIN) decomposition. The GEO decomposition consists of first partitioning the whole railway network into many elementary block sections and then clustering these block sections into a

given number of regions. An integer linear optimization approach is proposed to cluster the block sections with the objective of minimizing the total number of train service interconnections among the regions and of balancing the region sizes. Consequently, several subproblems are obtained, and each region corresponds to one subproblem. For the TRA decomposition, we decompose an F-train problem into F subproblems, and each subproblem includes one individual train only. The TIN decomposition makes a division of the time horizon into equal-interval pieces, and each time-interval piece corresponds to one subproblem, which consists of all events (i.e., train departures and arrivals) that are estimated to happen in this time-interval. No matter which decomposition method is used, couplings always exist among subproblems, and the presence of these couplings leads to a non-separable structure of the whole optimization problem. To handle the issue of the couplings, we introduce three distributed optimization approaches. The first one is an Alternating Direction Method of Multipliers (ADMM) algorithm, where each subproblem is solved through coordination with the other subproblems in an iterative manner. The second one is a priority-rule-based (PR) algorithm, where the subproblems are sequentially and iteratively solved in a priority order (based on train delays) with respect to the solutions of the other subproblems that have been solved with a higher priority. The third one is a Cooperative Distributed Robust Safe But Knowledgeable (CDRSBK) algorithm, where four types of couplings are defined and each subproblem is iteratively solved together with its actively coupling subproblems. Experiments are conducted based on the Dutch railway network to comparatively test the performance of the three proposed algorithms with the three decomposition methods, in terms of feasibility, computational efficiency, solution quality, and estimated optimality gap.

The reminder of this paper is organized as follows. In Section 2, we briefly introduce an MILP problem that we focus on in this paper, which addresses the integrated problem of real-time traffic management and train control. Section 3 introduces three decomposition methods, where a number of subproblems are obtained. In Section 4, three distributed optimization approaches are developed for handling the couplings among the resulting subproblems. Section 5 examines the performance of the proposed algorithms and decomposition methods, through experiments on the Dutch railway network. Finally, the conclusions and suggestions for future research are given in Section 6.

2 An MILP Approach for Addressing the Integration of Traffic Management and Train Control

An MILP approach has been developed in our previous work (Luan et al., 2018) for addressing the integrated problem of real-time traffic management and train control. This MILP approach incorporates the representations of microscopic traffic regulations and train speed trajectories into a single MILP optimization problem of the following form:

$$\min_{\lambda} \ \mathcal{Z}(\lambda) = c^{\top} \cdot \lambda \tag{1a}$$

$$\text{s.t. } A \cdot \lambda \leq b \tag{1b}$$

with variable $\lambda \in \mathbb{R}^n$, matrix $A \in \mathbb{R}^{m \times n}$, and vectors $c \in \mathbb{R}^n$, $b \in \mathbb{R}^m$. The objective function $Z(\lambda)$ in (1a) minimizes the weighted sum of the total train delay times at all visited stations and the energy consumption of the train movements. The vector λ contains both the traffic-related variables and train-related variables for describing the train movements on block sections, in particular, the arrival times a, departure times d, train orders θ, incoming speeds v^{in}, cruising speeds v^{cru}, outgoing speeds v^{out}, approach time τ^{approach}, and clear

time τ^{clear}. In (1b), all constraints (inequalities and equalities) are represented for ensuring the train speed limitations, for enforcing the consistency of train transition times and speeds, for guaranteeing the required dwell times, for determining train blocking times, and for respecting the block section capacities. The MILP problem (1a)-(1b) can be solved by a standard MILP solver, e.g., CPLEX or Gurobi. Interested readers are referred to the optimization problem named P_{TSPO} in Luan et al. (2018) for a more detailed description.

3 Problem Decomposition

Three decomposition methods, i.e., the geography-based (GEO), the train-based (TRA), and the time-interval-based (TIN) decomposition, are described in Sections 3.1-3.3 respectively. Section 3.4 discusses the decomposition result, i.e., subproblems and couplings. Figure 1 comparatively illustrates the three decomposition methods in a time-space graph, where black lines indicate train paths and red dashed lines indicate boundaries of subproblems.

3.1 Geography-Based Decomposition

The GEO decomposition partitions the whole railway network into a given number of regions. Consider a railway network composed of a set of block sections E and a set of scheduled trains F traversing this network. We could easily partition the whole network into $|E|$ units, by means of a geography-(i.e., block section)-based decomposition; however, this could result in a large number of subproblems with couplings. In general, a larger number of subproblems implies more couplings among them, which makes coordination difficult and which may affect the overall performance of the system; therefore, we cluster these elementary block sections into a pre-defined number $|R|$ of regions, where $R = \{1, 2, ..., |R|\}$ is the set of regions. Figure 1(b) illustrates a 2-region example of the geography-based decomposition; as shown, the timetable is split in the dimension of space.

To distribute $|E|$ different units into $|R|$ groups, there are $|R|^{|E|}$ ways, e.g., up to 10^6 ways for distributing 20 units into 2 groups only. Thus, in our case, a huge number of the GEO decomposition results are available. To obtain the optimal decomposition result, an integer linear programming (ILP) approach is proposed in Appendix B, with the objective of minimizing the number of couplings among regions (i.e., the total number of train service interconnections) and balancing the region sizes (i.e., the absolute deviation between the number of block sections contained in an individual region and the average value $|E|/|R|$).

For the GEO decomposition with a pre-defined number of regions, there are two impact

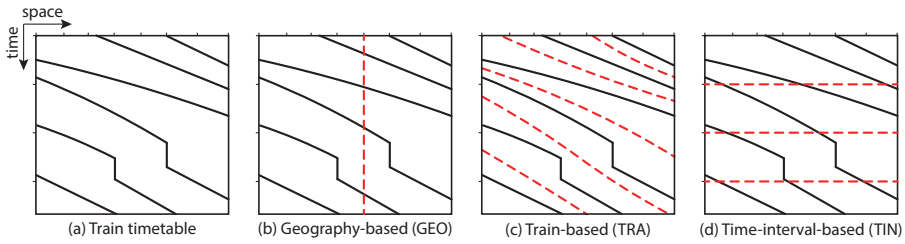


Figure 1: Illustration of the three decomposition methods in a time-space graph

factors: the network layout and the train routes planned in the original timetable. This implies that the optimal decomposition result is same for all delay cases.

When applying the GEO decomposition, some trains may traverse from one region to another region. The time and speed that a train leaves one region should equal the time and speed that the train arrives at the other region. Therefore, the time and speed transition constraints are the complicating constraints for the GEO decomposition, which cause the couplings among regions (i.e., subproblems). The time and speed transition constraints of the MILP problem (1) are formulated in (15a)-(15b) of Appendix A.

3.2 Train-Based Decomposition

The TRA decomposition simply splits a $|F|$-train problem into $|F|$ subproblems, and each subproblem corresponds to a 1-train problem, as illustrated in Figure 1(c). Thus, for a given instance, only one decomposition result is available. The only impact factor of the TRA decomposition is the involved trains. Brännlund et al. (1998) used such train-based decomposition for addressing train timetabling problem by using Lagrangian relaxation.

When applying the TRA decomposition, each train is independently scheduled in each subproblem, so that trains may use the same infrastructure at the same time, resulting in conflicts. Therefore, the capacity constraint is the complicating constraint for the TRA decomposition. The capacity constraint is formulated in (15c)-(15d) of Appendix A.

3.3 Time-Interval-Based Decomposition

The time-interval-based (TIN) decomposition makes a division of a train timetable in the dimension of time, based on a given size of time interval, as illustrated in Figure 1. The TIN decomposition is implemented with consideration of disruptions (delays). We independently schedule all trains by taking disruptions into account, generating an infeasible timetable, where train conflicts exist. With this infeasible timetable, we estimate the times that all events (e.g., train departure and arrival) may happen. Each event is then assigned to one time interval based on its estimated happen time. As a result, the subproblem of each time interval includes all events that are estimated to happen in this time interval. The TIN decomposition result mainly depends on the given size of time interval and the estimated train schedule, which can be different in delay cases.

One train service consists of a set of events indicating the departures and arrivals of the train on block sections. When applying the TIN decomposition, these events may be split into more than one time intervals. Thus, same to the GEO decomposition (where trains may traverse from region to region), the time and speed when a train leaves a time interval should be consistent with those when the train enters the next time interval, i.e., the time and speed transition constraints are complicating constraints, as formulated in (15a)-(15b) of Appendix A. Moreover, as the TIN decomposition is based on an estimated infeasible timetable, an event assigned to time interval t maybe further scheduled to the next time interval $t+1$, causing conflicts with the events in time interval $t+1$. Therefore, the capacity constraint in (15c)-(15d) is also a complicating constraint for the TIN decomposition.

3.4 Subproblems and Couplings

Let us denote S as the set of the $|S|$ resulting subproblems, e.g., $|S| = |R|$ for the GEO decomposition. No matter which decomposition method is used, we can always divide the

constraints of the MILP problem (1) into two categories, i.e., local constraints and complicating constraints. A local constraint is only related to a single subproblem, so that it leads to a separable structure of an optimization problem. A complicating constraint is associated with at least two subproblems, so that it results in a non-separable structure. We thus rewrite (1b) into a general form of the following local and complicating constraints:

$$A^{\text{loc}} \cdot \lambda \leq b^{\text{loc}} \tag{2a}$$

$$A^{\text{cpl}} \cdot \lambda \leq b^{\text{cpl}} \tag{2b}$$

with matrices $A^{\text{loc}} \in \mathbb{R}^{m_1 \times n}$ and $A^{\text{cpl}} \in \mathbb{R}^{m_2 \times n}$ and vectors $b^{\text{loc}} \in \mathbb{R}^{m_1}$ and $b^{\text{cpl}} \in \mathbb{R}^{m_2}$. A detailed explanation of the complicating constraints of the MILP problem (1) is given in Appendix A. Let us denote $Q_p = \left\{q_1, q_2, ..., q_{m_p}\right\}$ as the set of m_p subproblems that have couplings with subproblem p. The subproblem $p \in S$ of the MILP problem (1) is formulated as

$$\min_{\lambda_p} \ \mathcal{Z}_p(\lambda_p) = c_p^\top \cdot \lambda_p \tag{3a}$$

$$\text{s.t. } A_p^{\text{loc}} \cdot \lambda_p \leq b_p^{\text{loc}} \tag{3b}$$

$$A_{p,q}^{\text{cpl}} \cdot \lambda_p + A_{q,p}^{\text{cpl}} \cdot \lambda_q \leq b_{p,q}^{\text{cpl}}, \ \forall q \in Q_p \tag{3c}$$

where $A_{p,q}^{\text{cpl}}$ and $A_{q,p}^{\text{cpl}}$ are selection matrices for selecting the coupling variables between subproblems p and q. Since each coupling constraint in (3c) includes the variables λ_p and λ_q of two subproblems p and q, we cannot explicitly add them to any individual subproblem. Instead we can determine and exchange values of the coupling variables among subproblems in an iterative way. The train(s) of one subproblem p can obtain an agreement through iterations that inform the train(s) of its coupling subproblems $q \in Q_p$ about what subproblem p prefers the values of coupling variables to be. To achieve this agreement, for a single subproblem p, we have to compute the optimal coupling variables (inputs) for its coupling subproblems $q \in Q_p$ as well, rather than only focusing on computing optimal local variables. Moreover, for its coupling subproblems $q \in Q_p$, we need to compute both the optimal local variables and coupling variables (outputs). Through exchanging these desired coupling variables, the values of these outputs and inputs should converge to each other, and a set of local inputs that is overall optimal should be found. Distributed optimization approaches are developed for reaching this agreement in Section 4.

4 Distributed Optimization Approaches

This section introduces three distributed optimization approaches to address the issue of couplings among subproblems, namely the Alternating Direction Method of Multipliers (ADMM) algorithm, the priority-rule-based (PR) algorithm, and the Cooperative Distributed Robust Safe But Knowledgeable (CDRSBK) algorithm, presented in Sections 4.1-4.3 respectively. A key challenge in distributed optimization algorithms is to ensure that the solution generated for a single subproblem leads to feasible solutions that satisfy the complicating constraints with other subproblems.

4.1 Alternating Direction Method of Multipliers Algorithm

The alternating direction method of multipliers (ADMM) algorithm (see e.g., Boyd et al., 2011) solves problems in the following form:

$$\min_{x,z} \quad f(x) + g(z) \tag{4a}$$

$$\text{s.t.} \quad A \cdot x + B \cdot z = b, \tag{4b}$$

with variables $x \in \mathbb{R}^n$ and $z \in \mathbb{R}^m$, matrices $A \in \mathbb{R}^{p \times n}$ and $B \in \mathbb{R}^{p \times m}$, and vector $b \in \mathbb{R}^p$. Assume that the variables x and z can be split into two parts, with the objective function separable across this splitting. We can then form the augmented Lagrangian relaxation as

$$L_\rho(x, z, y) = f(x) + g(z) + y^\top (A \cdot x + B \cdot z - b) + \tfrac{\rho}{2} \cdot \|A \cdot x + B \cdot z - b\|_2^2, \quad (5)$$

where y is the dual variable (Lagrangian multiplier), the parameter $\rho > 0$ indicates the penalty multiplier, and $\|\cdot\|_2$ denotes the Euclidean norm. The augmented Lagrangian function is optimized by minimizing over x and z alternately or sequentially and then evaluating the resulting equality constraint residual. By applying the dual ascent method, the ADMM algorithm consists of the following iterations:

$$x^{i+1} := \arg\min_x L_\rho(x, z^i, y^i), \quad (6a)$$

$$z^{i+1} := \arg\min_z L_\rho(x^{i+1}, z, y^i), \quad (6b)$$

$$y^{i+1} := y^i + \rho(A \cdot x^{i+1} + B \cdot z^{i+1} - b) \quad (6c)$$

where i is the iteration counter. In the ADMM algorithm, the variables x and z are updated in an alternating or sequential fashion, which accounts for the term alternating direction.

The ADMM algorithm can obviously deal with linear equality constraints, but it can also handle linear inequality constraints. The latter can be reduced to linear equality constraints by replacing constraints of the form $A \cdot x \leq b$ by $A \cdot x + s = b$, adding the slack variable s to the set of optimization variables, and setting $\mathcal{Z}(s) = 0$, if $s \geq 0$, otherwise, setting $\mathcal{Z}(s) = \infty$. Alternatively, we can also work with an equivalent reformulation of problem (3), where we replace the complicating constraint (3c) by

$$\mathcal{C}_p(\lambda_p, \lambda_q) = 0 \quad (7)$$

where $\mathcal{C}_p(\lambda_p, \lambda_q) = \max\left\{0, A^{\text{cpl}}_{p,q} \cdot \lambda_p + A^{\text{cpl}}_{q,p} \cdot \lambda_q - b^{\text{cpl}}_{p,q}\right\}$ with component-wise maximum. In such a way, we can transform the inequality constraints into equality constraints.

Now we can apply the ADMM algorithm, and the augmented Lagrangian formulation of the MILP problem (1) can be described as follows:

$$L_\rho = \sum\nolimits_{p \in S} \left[\mathcal{Z}_p(\lambda_p) + \sum\nolimits_{q \in Q_p} \left[y^\top_{p,q} \cdot \mathcal{C}_p(\lambda_p, \lambda_q) + \frac{\rho}{2} \cdot ||\mathcal{C}_p(\lambda_p, \lambda_q)||_2^2 \right] \right] \quad (8)$$

The iterations to compute the solution of the MILP problem (1) based on the augmented Lagrangian formulation (8) include quadratic terms; therefore, the function cannot directly be distributed over subproblems. Inspired by Negenborn et al. (2008), for handling this non-separable issue, the function (8) can be approximated by solving $|S|$ separate problems of the form

$$\min_{\lambda_p} \ \mathcal{Z}_p(\lambda_p) + \sum\nolimits_{q \in Q_p} \mathcal{J}_p(\lambda_q, y_{p,q}) \quad (9)$$

subject to (3b) for the train movements of single subproblem p, where the additional term $\mathcal{J}_p(\cdot)$ deals with coupling variables.

We now define the term $\mathcal{J}_p(\cdot)$ by using a serial implementation. We apply a block coordinate descent approach (Beltran Royoa and Heredia, 2002; Negenborn et al., 2008). The approach minimizes the quadratic term directly in a serial manner. One subproblem after another minimizes its local and coupling variables while the variables of the other subproblems stay fixed. At iteration i, let us denote $\widehat{Q_p^i} \subseteq Q_p$ as the set of those coupling subproblems (of subproblem p) that have been solved before solving subproblem p.

The serial implementation uses the information from both the current iteration i and the last iteration $i-1$. With the information $\bar{\lambda}_q = \lambda_q^{(i)}$ computed in the current iteration i for subproblems $q \in \widehat{Q_p^i}$ and the information $\bar{\lambda}_q = \lambda_q^{(i-1)}$ obtained in the last iteration $i-1$

for the other subproblems $q \in Q_p \backslash \widehat{Q_p^i}$, we can solve (9) for subproblem p by using the following function:

$$\mathcal{J}_p(\bar{\lambda}, y_{p,q}) = y_{p,q}^{\top} \cdot \mathcal{C}_p(\lambda_p, \bar{\lambda}_q) + \frac{\rho}{2} \cdot ||\mathcal{C}_p(\lambda_p, \bar{\lambda}_q)||_2^2 \tag{10}$$

The second term of (10) penalizes the deviation from the coupling variable iterates that were computed for the subproblems before subproblem p in the current iteration i and by the other subproblems during the last iteration $i-1$.

The solution procedure of the ADMM algorithm is described as follows:

The solution procedure of the ADMM Algorithm

Initialization: Set the iteration counter $i := 1$, the penalty multiplier $\rho := 1$, the Lagrange multipliers $y^{(0)} := 0$, and all elements in the latest solution set $\mathcal{S}_{\text{sol}} := \{\bar{\lambda}_p | p \in S\}$ to be empty. Denote the maximum number of iterations as $I^{\max}$.

1: **for** iteration $i := 1, 2, ..., I^{\max}$ **do**
2: Randomly generate the orders of subproblems, denoted as $P_{\text{order}}^{(i)}$.
3: **for** subproblem $j := 1, 2, ..., |S|$ **do**
4: Solve subproblem $p := P_{\text{order}}^{(i)}(j)$, consisting of objective function (9) and constraint (3b), by taking the available solutions in $\mathcal{S}_{\text{sol}}$ for all $q \in \widehat{Q_p^i}$ into account.
5: Denote the obtained solution of subproblem p as $\lambda_p^{(i)}$, and update the latest solution set $\mathcal{S}_{\text{sol}}$ by adding or setting $\bar{\lambda}_p := \lambda_p^{(i)}$.
6: **end for**
7: Update the Lagrange multipliers by $y_{p,q}^{(i)} := y_{p,q}^{(i-1)} + \rho \cdot \mathcal{C}_p(\lambda_p^{(i-1)}, \lambda_q^{(i-1)})$ for all $p \in S$ and $q \in Q_p$.
8: Break the iterations if the difference of the coupling variables at the current iteration step i is less than the expected gap ϵ, i.e., $\|\mathcal{C}\|_\infty \leq \epsilon$, where ϵ is a small positive scalar and $\|\cdot\|_\infty$ denotes the infinity norm.
9: **end for**

By applying the ADMM algorithm, we solve the subproblems $p \in S$ in an iterative manner, with respect to the local constraint (3b) of a single subproblem p and taking the solutions of all coupling subproblems (i.e., the variable $\bar{\lambda}_q$ for $q \in Q_p$ obtained in either the current iteration or the last iteration) into account. In (8), only the local objective $\mathcal{Z}_p$ for a single subproblem p is minimized, not the global objective $\sum_{p \in S} \mathcal{Z}_p$ for all subproblems.

In order to further improve the performance of the ADMM algorithm, we can consider a cost-to-go function $Z_p^{\text{ctg}}(\lambda_p)$ into the objective function of each subproblem, which provides an estimation of the train running to its destination. Then, the objective function (9) for subproblem $p \in S$ can be rewritten as follows:

$$\min_{\lambda_p} \ \mathcal{Z}_p(\lambda_p) + \mathcal{Z}_p^{\text{ctg}}(\lambda_p) + \sum\nolimits_{q \in Q_p} \mathcal{J}_p(\lambda_q, y_{p,q}) \tag{11}$$

For instance, with the GEO decomposition, we can define the cost-to-go function as the deviation between the actual and planned departure time from the block section where a train leaves a region. Thus, an original timetable with more details is needed, where the departure and arrival times are given not only for stations but also for block sections.

4.2 Priority-Rule-Based Algorithm

The ADMM algorithm incorporates the complicating constraint (3c) into the objective function and strives to make the information consistent among subproblems (i.e., each subprob-

lem takes the information of the other subproblems into account) in an iterative manner. However, convergence cannot be guaranteed for non-convex optimization problems, so that a feasible solution may not be available. Therefore, we need to explore other distributed optimization approaches. We next introduce a priority-rule-based (PR) algorithm.

The main idea of the PR algorithm is to optimize train schedules of the subproblems in a sequential manner according to problem priorities, with respect to the solutions of the other subproblems that have already been solved in the current iteration. The problem priorities are determined by the train delay times of the subproblems, e.g., we solve the subproblem with the largest delay time first. Note that the result could be different even with the same problem priorities, as multiple optimal solutions may exist for each subproblem. These different optimal solutions with the same objective value for one subproblem could result in different objective values for the other subproblems.

By applying the PR algorithm, the complicating constraint (3c) for the subproblem $p \in S$ can be rewritten as follows:

$$A^{\text{cpl}}_{p,q} \cdot \lambda_p + A^{\text{cpl}}_{q,p} \cdot \bar{\lambda}_q \leq b^{\text{cpl}}_{p,q},\ \forall q \in Q_p \tag{12}$$

with the solution $\bar{\lambda}_q = \lambda_q^{(i)}$ computed in the current iteration i for all subproblems $q \in \widehat{Q_p^i}$.

The solution procedure of the PR algorithm is described as follows:

The solution procedure of the PR Algorithm

Initialization: Set the iteration counter $i := 1$, the local upper bound $o^{(0)}_{\text{UB}} := M$, and the global upper bound $O^{(0)}_{\text{UB}} := M$, where M is a sufficient large positive number. Initialize the problem priorities $P^{(0)}_{\text{prior}}$ arbitrarily. Denote the maximum number of iterations as $I^{\max}$.

1: **for** iteration $i := 1, 2, ..., I^{\max}$ **do**
2: Sort subproblems in set S in a descending order by their problem priorities $P^{(i-1)}_{\text{prior}}$, denoted as $P^{(i)}_{\text{order}}$.
3: Set the solution set $\mathcal{S}_{\text{sol}} := \{\bar{\lambda}_p | p \in S\}$ to be empty.
4: **for** subproblem $j := 1, 2, ..., |S|$ **do**
5: Solve subproblem $p := P^{(i)}_{\text{order}}(j)$, including objective function (3a) and constraints (3b) and (12), with respect to the available solutions in $\mathcal{S}_{\text{sol}}$ for all $q \in \widehat{Q_p^i}$.
6: Denote the obtained solution of subproblem p as $\lambda_p^{(i)}$, and update the solution set $\mathcal{S}_{\text{sol}}$ by adding $\bar{\lambda}_p := \lambda_p^{(i)}$.
7: **end for**
8: Compute the local upper bound $o^{(i)}_{\text{UB}}$, and update the global upper bound by

$$O^{(i)}_{\text{UB}} := \begin{cases} o^{(i)}_{\text{UB}}, & \text{if } O^{(i-1)}_{\text{UB}} > o^{(i)}_{\text{UB}} \\ O^{(i-1)}_{\text{UB}}, & \text{otherwise} \end{cases}$$

9: Update the problem priorities $P^{(i)}_{\text{prior}}$ by the train delay times of the subproblems.
10: Break the iterations if the global upper bounds are not improved for a given number of iterations κ, i.e., $O^{(i)}_{\text{UB}} = O^{(i-\kappa)}_{\text{UB}}$.
11: **end for**

In the priority-rule-based algorithm, we solve each subproblem $p \in S$ in a sequential manner according to the priorities of the subproblems, with respect to the local constraint (3b) and the outputs $\bar{\lambda}_q$ of the coupling subproblems $q \in Q_p$ in (12). Similar to the ADMM

algorithm, only the local objective Z_p is minimized when solving subproblem p, rather than the global objective $\sum_{p \in R} Z_p$ for all subproblems. Constraint (12) ensures that the coupling variables of subproblem p satisfy those of its coupling subproblems $q \in Q_p$ obtained in the current iteration. For the first solved subproblem in each iteration, the complicating constraint (12) is relaxed.

4.3 Cooperative Distributed Robust Safe but Knowledgeable Algorithm

The third algorithm considered in this paper is the Cooperative distributed robust safe but knowledgeable (CDRSBK) algorithm, introduced by Kuwata and How (2011) to address trajectory planning problems. In the CDRSBK algorithm, four types of couplings among subproblems are defined for a subproblem $p \in S$, as illustrated in Figure 2.

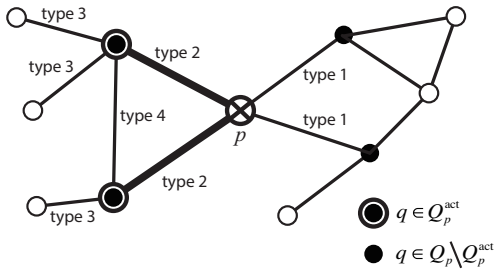


Figure 2: Four types of couplings defined in the CDRSBK algorithm

Type_1 indicates a non-active coupling between subproblem $p \in S$ and its neighbor; Type_2 indicates an active coupling between subproblem p and its neighbor; Type_3 indicates the coupling between the active coupling neighbors of subproblem p and their neighbors; and Type_4 indicates the coupling between two active coupling neighbors of subproblem p. Let us denote Q_p as the set of all coupling neighbors of subproblem p and denote Q_p^{act} as the set of subproblem p's neighbors that have an active coupling with subproblem p. The interpretation of active and non-active couplings can be different for different decomposition methods. We discuss the details regarding their implementations in Section 4.4.

By applying the CDRSBK algorithm, the subproblem $p \in S$ of the MILP problem (3a)-(3c) can be reformulated as

$$\min_{\lambda_p, \xi_q} \quad \mathcal{Z}_p(\lambda_p) + \sum\nolimits_{q \in Q_p^{\text{act}}} \mathcal{Z}_q(\bar{\lambda}_q + T_q \cdot \xi_q) \tag{13a}$$

$$\text{s.t.} \quad A_p \cdot \lambda_p \leq b_p^{\text{loc}} \tag{13b}$$

$$A_q \cdot (\bar{\lambda}_q + T_q \cdot \xi_q) \leq b_q^{\text{loc}}, \forall q \in Q_p^{\text{act}} \tag{13c}$$

$$A_{p,q}^{\text{cpl}} \cdot \lambda_p + A_{q,p}^{\text{cpl}} \cdot \bar{\lambda}_q \leq b_{p,q}^{\text{cpl}}, \forall q \in Q_p \backslash Q_p^{\text{act}} \tag{13d}$$

$$A_{p,q}^{\text{cpl}} \cdot \lambda_p + A_{q,p}^{\text{cpl}} \cdot (\bar{\lambda}_q + T_q \cdot \xi_q) \leq b_{p,q}^{\text{cpl}}, \forall q \in Q_p^{\text{act}} \tag{13e}$$

$$A_{o,q}^{\text{cpl}} \cdot \bar{\lambda}_o + A_{q,o}^{\text{cpl}} \cdot (\bar{\lambda}_q + T_q \cdot \xi_q) \leq b_{o,q}^{\text{cpl}}, \forall o \in Q_q \backslash Q_p^{\text{act}}, q \in Q_p^{\text{act}} \tag{13f}$$

$$A_{q_1,q_2}^{\text{cpl}} \cdot (\bar{\lambda}_{q_1} + T_{q_1} \cdot \xi_{q_1}) + A_{q_2,q_1}^{\text{cpl}} \cdot (\bar{\lambda}_{q_2} + T_{q_2} \cdot \xi_{q_2}) \leq b_{q_1,q_2}^{\text{cpl}}, \quad \forall q_1, q_2 \in Q_p^{\text{act}}, q_2 \in Q_{q_1}, q_1 \in Q_{q_2} \tag{13g}$$

In (13a), the objective function of both subproblem p and its actively coupled subproblems $q \in Q_p^{\text{act}}$ are included. Constraints (13b)-(13c) represent the local constraints of

subproblem p and its actively coupled subproblems $q \in Q_p^{\text{act}}$ respectively. In (13d)-(13g), coupling constraints (3c) are rewritten for the four types of couplings among subproblems respectively. When solving subproblem p, besides the local variable λ_p, the variable ξ_q is also optimized for its actively coupled subproblems $q \in Q_p^{\text{act}}$ on the communicated solution $\bar{\lambda}_q$, as follows:

$$\lambda_q = \bar{\lambda}_q + T_q \cdot \xi_q \tag{14}$$

parameterized with a matrix T_q, which is formed to allow the variable ξ_q to change only the rows corresponding to the active complicating constraints. This can be also interpreted as allowing a change for the constraint that has a non-zero Lagrange multiplier. In (13a), the objectives of a single subproblem p and its actively coupled neighbors $q \in Q_p^{\text{act}}$ are both minimized.

The solution procedure of the CDRSBK algorithm is described as follows:

The solution procedure of the CDRSBK Algorithm

Initialization: Set the iteration counter $i := 1$, the local upper bound $o_{\text{UB}}^{(1)} := M$, and the global upper bound $O_{\text{UB}}^{(1)} := M$, and all elements in the latest solution set $\mathcal{S}_{\text{sol}} := \{\bar{\lambda}_p | p \in S\}$ to be empty. Denote the maximum number of iterations as $I^{\max}$.

1: **for** iteration $i := 1, 2, ..., I^{\max}$ **do**
2: Randomly generate the orders of subproblems, denoted as $P_{\text{order}}^{(i)}$.
3: **for** subproblem $j := 1, 2, ..., |S|$ **do**
4: Solve subproblem $p := P_{\text{order}}^{(i)}(j)$ and its actively coupling subproblems $q \in Q_p^{\text{act}}$, consisting of objective function (13a) and constraints (13b)-(13g), by taking the available solutions in set $\mathcal{S}_{\text{sol}}$ for all $o \in (Q_p \backslash Q_p^{\text{act}}) \cup (Q_q \backslash Q_p^{\text{act}})$ into account.
5: Denote the obtained solutions of subproblem p and its actively coupling subproblems $q \in Q_p^{\text{act}}$ as $\lambda_p^{(i)}$ and $\lambda_q^{(i)}$ (which is obtained by (14)) respectively, and update the latest solution set $\mathcal{S}_{\text{sol}}$ by adding or setting $\bar{\lambda}_p := \lambda_p^{(i)}$ and $\bar{\lambda}_q := \lambda_q^{(i)}$ for all $q \in Q_p^{\text{act}}$.
6: **end for**
7: Compute the local upper bound $o_{\text{UB}}^{(i)}$, and update the global upper bound by

$$O_{\text{UB}}^{(i)} := \begin{cases} o_{\text{UB}}^{(i)}, & \text{if } O_{\text{UB}}^{(i-1)} > o_{\text{UB}}^{(i)} \\ O_{\text{UB}}^{(i-1)}, & \text{otherwise} \end{cases}$$

8: Break the iterations if the global upper bounds are not improved for a given number of iterations κ, i.e., $O_{\text{UB}}^{(i)} = O_{\text{UB}}^{(i-\kappa)}$.
9: **end for**

In each iteration, the CDRSBK algorithm actually solves each subproblem, with additional objectives and coupling constraints that include the changeable (local) variables of its actively coupled subproblems $q \in Q_p^{\text{act}}$. If the variables of its actively coupled subproblems are unchangeable, i.e., $\lambda_q = \bar{\lambda}_q$ when ξ_q has no impact on the variables, the coupling constraints are automatically satisfied and could be omitted.

4.4 Remarks on the Implementation of the Decomposition Methods and Algorithms

Here we give some remarks for the implementation of the proposed decomposition methods and algorithms, e.g., interpreting the active and non-active couplings in the CDRSBK algo-

rithm for different decomposition methods and giving some tips for achieving feasibility.

Remark 1 (Train orders in the ADMM algorithm with the GEO decomposition and the TIN decomposition)**.** It is essential to ensure that train orders in subproblems are feasible, in order to avoid unnecessary iterations and to achieve fast convergence. To do this, we keep a consistency of the train orders that are interrelated, e.g., if two trains cannot overtake on a sequence of block sections, then the train orders of these two trains on these block sections are interrelated and must be same.

Remark 2 (The CDRSBK algorithm & the GEO decomposition)**.** If two regions are connected by tracks, i.e., they are neighbors, then we consider that a coupling exists between the two subproblems of these two regions. A coupling between two subproblems is considered to be active (Type_2) if there is any train traverse between the two regions of the two subproblems; otherwise, the coupling is recognized as non-active coupling (Type_2). For coupling Type_3 and Type_4, we follow their general definitions, i.e., the couplings between an active coupling neighbor and its coupling neighbors are labeled as Type_3 coupling and the coupling between two active coupling neighbor is labeled as Type_4.

Remark 3 (The CDRSBK algorithm & the TRA decomposition)**.** If two trains use the same infrastructure (block section), then we consider that a coupling exists between the two subproblems of these two trains. If a conflict exists between these two trains, then their coupling is recognized as an active coupling; otherwise, their coupling is considered to be non-active. For coupling Type_3 and Type_4, we follow their general definitions. In the TRA decomposition, we often have many trains that use the same infrastructure; but conflicts may never happen among some of them, e.g., a train scheduled in the early morning has little chance to conflict with another train scheduled in the late afternoon. Thus, to further reduce the problem complexity for large-scale networks, we provide two more options for defining coupling Type_1 and Type_3. We denote the option described above as Opt_1. The difference between Opt_1 and Opt_2 is in the definition of coupling Type_3: in Opt_2, we label the couplings between an active coupling neighbor and its *active* coupling neighbor as Type_3. Based on Opt_2, we discard all Type_1 couplings, which results in Opt_3, i.e., when and only when a conflict happens between two trains, a coupling exists between them and is recognized as active coupling (Type_2). However, we still have Type_3 and Type_4 couplings in Opt_3 by following their general definitions. An illustrative example is provided in Appendix C to graphically explain these three options.

Remark 4 (The CDRSBK algorithm & the TIN decomposition)**.** Due to the nature of the TIN decomposition, the relation among subproblems is relatively simple in this case. Couplings exist only between two consecutive subproblems (i.e., two subproblems of two consecutive time intervals t and $t+1$) and are all recognized as active couplings (Type_2). As a result, according to the general definition of the four types of couplings, the couplings between a consecutive subproblem and its consecutive subproblem are considered as Type_3 (e.g., for subproblem t, a Type_3 coupling exists between subproblems $t+1$ and $t+2$), and Type_1 and Type_4 couplings do not exist. Moreover, for guaranteeing a feasible solution in the first iteration, solving subproblems in a time sequence (i.e., for time intervals $t = 1, 2, 3....$ in sequence) is recommended.

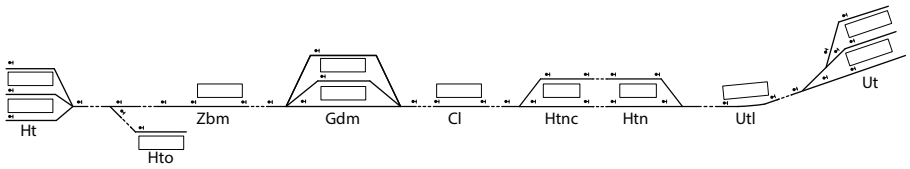


Figure 3: An experimental railway network

5 Case Study

5.1 Set-Up

We consider a line of the Dutch railway network, connecting Utrecht (Ut) to Den Bosch (Ht), of about 50 km length, with 9 stations, as shown in Figure 3. The network comprises 42 nodes and 40 cells. We consider one hour of heterogeneous traffic with 15 trains. Moreover, we considered different numbers of regions for the GEO decomposition, ranging from 2 to 6, and we consider 4 time intervals for the TIN decomposition, i.e., 300s, 600s, 900s, and 1200s. In the result presentation, we present the average result of 15 delay cases with randomly generated primary delays. The maximum number of iterations is set to 200, 100, and 30 for the ADMM, PR, and CDRSBK algorithm respectively. A larger number is set for the ADMM algorithm because it needs some iterations to converge, and a smaller number is set for the CDRSBK algorithm because it often finds a feasible solution very fast and its solution is updated multiple times in one iteration. In the case study, we consider the weight $\zeta = 0.55$ for the ILP problem proposed in Appendix B for the GEO decomposition, which is appropriate for getting a result with an acceptable difference of the size of regions.

We adopt the CPLEX solver version 12.6.3 implemented in the MATLAB (R2018a) TOMLAB toolbox to solve the MILP problems. The experiments are performed on a computer with an Intel® Core™ i7 @ 2.00 GHz processor and 16GB RAM.

5.2 Experimental Results

This section shows the (average) results of 15 delay cases from the viewpoints of feasibility, estimated optimality gap, solution quality, and computational efficiency.

Figure 4 presents the number of cases that we can find feasible solutions within the maximum number of iterations. We can conclude that, for achieving feasibility, the TRA decomposition performs best among the three decomposition methods, and the CDRSBK algorithm is the best among the three algorithms. Considering a larger number of regions for the GEO decomposition or considering a smaller time interval for the TIN decomposition can make feasibility difficult to achieve, as they lead to a larger number of couplings among subproblems.

In Figure 5, an estimated optimality gap for each decomposition method and each algorithm is given. As shown, the estimated optimality gap of the GEO decomposition is 3.52%, the lowest among the three decomposition methods, and the CDRSBK algorithm has the smallest estimated optimality gap (only 1.11%) among the three algorithms. A large estimated optimality gap does not reflect a bad solution quality; it may be caused by a loose lower bound, as in the case of the TRA decomposition.

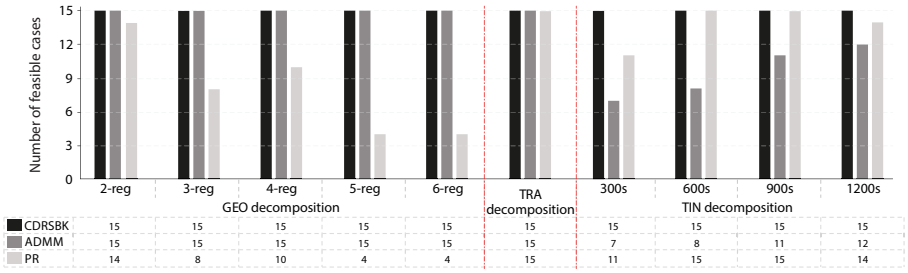


Figure 4: Feasibility of the three decomposition methods and three algorithms

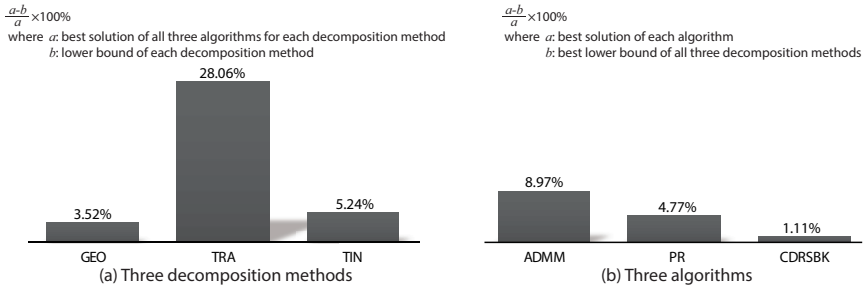


Figure 5: Estimated optimality gap of the three decomposition methods and three algorithms

Figure 6 shows the cumulative computation time (on the X-axis) and the objective value (on the Y-axis). The cumulative computation time is the CPU time consumed for finding the best feasible solution. Dashed circles around symbols indicate that feasible solution(s) can be found for all 15 delay cases by using the corresponding decomposition method and algorithm. When focusing on the three decomposition methods (represented by colors), the GEO decomposition (in pink) leads to a large range in computation time and a small range in objective value. This implies that the GEO decomposition results in small differences in the solution quality, but the computational efficiency is quite different for different algorithms. For the TRA decomposition (in blue) and the TIN decomposition (in green), ranges still exist in the two dimensions, and their results show a general trade-off between solution quality and computational efficiency. Let us now focus on the three algorithms (indicated by symbols). The CDRSBK algorithm (indicated by diamonds) overall yields the best solution quality, and the computation efficiency becomes much better when the TRA decomposition is applied. The performance of the ADMM and PR algorithms is highly variable. For the ADMM algorithm (indicated by circles), the best solution quality is achieved when using the GEO decomposition, and the best computation efficiency is achieved when the TRA decomposition is adopted. The PR algorithm (indicated by triangles) has the best performance on solution quality when the GEO decomposition is used and on computational efficiency when the TIN decomposition is applied. A black dashed circle around a symbol indicates that feasible solution(s) can be found for all 15 delay cases by using the corresponding decomposition method and algorithm. Moreover, the lower bound of the TRA decomposition (indicated by a blue cross symbol) is the loosest, which leads to its large estimated

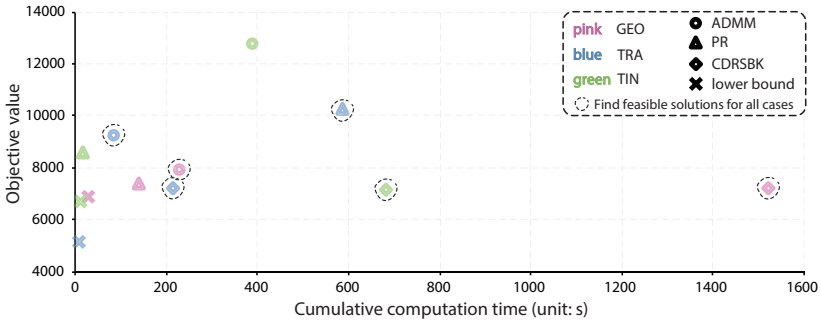


Figure 6: Solution quality and computational efficiency

optimality gap in Figure 5.

Overall, the CDRSBK algorithm with the TRA decomposition, the ADMM algorithm with the GEO decomposition, and the ADMM algorithm with the TRA decomposition have good overall performance. All these three combinations can find feasible solutions for all delay cases. In comparison, the first two combinations have the best performance on solution quality and a satisfactory performance on computational efficiency. The last combination shows the best computational efficiency (roughly half-shorter computation time than the first two combinations) but at the cost of relatively bad solution quality.

Moreover, when using the CDRSBK algorithm together with the TRA decomposition, Opt_3 described in Section 4.4 yields the best performance on both solution quality and computational efficiency. For Opt_1, Opt_2, and Opt_3, the average objective value for the 15 delay cases is 7934.43, 7334.86, and 7217.08 respectively, and the average cumulative computation time is 255.19 seconds, 224.64 seconds, and 104.75 seconds.

6 Conclusions

We have introduced distributed optimization approaches, aiming at improving the computational efficiency of the integrated optimization problem for large-scale railway networks. Three decomposition methods have been presented to split the whole optimization problem into several subproblems, and three distributed optimization approaches have been proposed for dealing with the couplings among subproblems.

The performance of the proposed approaches has been examined in terms of feasibility, estimated optimality gap, solution quality, and computational efficiency. The TRA decomposition and the CDRSBK algorithm have the best performance from the perspective of feasibility. The GEO decomposition and the CDRSBK algorithm yield the smallest estimated optimality gap. The CDRSBK algorithm with the TRA decomposition and the ADMM algorithm with the GEO decomposition achieve the best performance on solution quality and satisfactory performance on computational efficiency. The ADMM algorithm with the TRA decomposition shows the best computational efficiency but gives a relatively bad solution.

For practical applications, a promising two-step procedure can be used: first generate a feasible solution in short time (e.g., by applying the ADMM algorithm) and then improve the solution quality (by using the CDRSBK algorithm) based on that feasible solution if time permits. This leads to one direction of the future research on exploring the interactions

of algorithms and decompositions so that we can play with their advantages, in order to further achieve best overall solution. Moreover, we are going to test the performance of the proposed approaches on larger-scale railway instances.

Acknowledgments

This work is supported by China Scholarship Council under Grant 201507090058.

Appendix A The complicating constraints in the MILP problem (1)

As explained in Section 3, there are some complicating constraints in the MILP optimization problem (1), causing the couplings among subproblems and making a non-separable structure of the whole problem.

When applying the GEO decomposition, the complicating constraints are the time and speed transition constraints, which can be written as follows:

$$d_{f,i,j} = a_{f,j,k}, \forall f \in F, (i,j) \in E_f, (j,k) \in E_f \tag{15a}$$

$$v^{\text{out}}_{f,i,j} = v^{\text{in}}_{f,j,k}, \forall f \in F, (i,j) \in E_f, (j,k) \in E_f \tag{15b}$$

Constraint (15a) enforces the transition time between two adjacent block sections, i.e., the departure time of train f on the preceding block section (i,j) equals the arrival time of train f on the successive block section (j,k), if two adjacent block sections (i,j) and (j,k) are used consecutively by train f. Constraint (15b) ensures the consistency of the train speed between two adjacent block sections, i.e., the incoming speed of train f on block section (j,k) equals to its outgoing speed on the preceding block section (i,j).

When applying the TRA decomposition, the couplings result from the competitive use of infrastructure by trains, i.e., the capacity constraint is the complicating constraint, formulated as follows:

$$\begin{aligned} a_{f',i,j} - \tau^{\text{approach}}_{f',i,j} - \tau^{\text{sig\_set}} + (1-\theta_{f,f',i,j}) \cdot M \geq d_{f,i,j} + \tau^{\text{clear}}_{f,i,j} + \tau^{\text{rel}}, \\ \forall f \in F, f' \in F, f \neq f', \rho_f = \rho_{f'}, (i,j) \in E_f, (i,j) \in E_{f'}, \end{aligned} \tag{15c}$$

$$\begin{aligned} a_{f',j,i} - \tau^{\text{approach}}_{f',j,i} - \tau^{\text{sig\_set}} + (1-\theta_{f,f',i,j}) \cdot M \geq d_{f,i,j} + \tau^{\text{clear}}_{f,i,j} + \tau^{\text{rel}}, \\ \forall f \in F, f' \in F, f \neq f', \rho_f \neq \rho_{f'}, (i,j) \in E_f, (j,i) \in E_{f'}. \end{aligned} \tag{15d}$$

where M is a sufficiently large positive number, $\tau^{\text{sig\_set}}$ is the setup, sight, and reaction time to lock a block section before the arrival of a train, and τ^{rel} is the release time to unlock a block section after the departure time of a train. Constraints (15c) and (15d) ensure that any pair of trains using one block section in the same or different direction respectively are conflict-free, by avoiding the overlap between the block section release time for a preceding train and the block section occupancy time for a successive train.

For the the TIN decomposition, all constraints in (15) can be complicating constraints.

Appendix B An integer linear programming approach for the geography-based decomposition

The set E_f contains the sequence of block sections composing the route of train f, and $|E_f|$ represents the number of block sections along the route of train f. The binary vector β_f indicates whether two consecutive block sections along the route of train f belong to different regions, e.g., if $(\beta_f)_j = 1$, then the j^{th} and $(j+1)^{\text{th}}$ block sections in set E_f

belong to different regions, otherwise, $(\beta_f)_j = 0$. The binary vector α_r indicates the assignment of all block sections for region r, e.g., if $(\alpha_r)_i = 1$, then the i^{th} block section in set E is assigned to region r, otherwise, $(\alpha_r)_i = 0$. The route matrix $B_f \in \mathbb{Z}^{(|E_f|-1)\times|E|}$ indicates that train f traverses a sequence of block sections, e.g., if train f traverses from the 1^{st} block section to the 3^{rd} block section in the set E, then $B_f = \begin{bmatrix} 1 & 0 & -1 & 0 & \dots \end{bmatrix}$. The integer vector $\mu \in (\mathbb{Z}^+)^{|E|}$ indicates the index of regions that each block section $e \in E$ belongs to. We use $\|\cdot\|_1$ to denote the 1-norm. The objective function is formulated as follows:

$$\min_{\alpha,\beta} \left[\zeta \cdot \left(\sum\nolimits_{f\in F} \|\beta_f\|_1 \right) + (1-\zeta) \cdot \left(\sum\nolimits_{r=1}^{|R|} \left| \|\alpha_r\|_1 - \frac{|E|}{|R|} \right| \right) \right], \tag{16}$$

where the weight $\zeta \in [0, 1]$ is used to balance the importance of the two objectives. The first term serves to minimize the train service interconnections among regions, and the second term aims at balancing the region sizes.

We consider four constraints, presented as follows:

$$\frac{\left|(B_f \cdot \mu)_j\right|}{|R|-1} \le (\beta_f)_j, \quad \forall f \in F, j \in \{1, ..., |E_f|-1\}, \tag{17}$$

guarantees that $(\beta_f)_j > 0$ if the two consecutive block sections along the route of train f belong to different regions, i.e., $\left|(B_f \cdot \mu)_j\right| > 0$.

$$\mu_i \in \{1, ..., |R|\}, \quad \forall i \in \{1, ..., |E|\}, \tag{18}$$

enforces that the indices of the resulting regions cannot exceed the pre-defined number of regions, while

$$(\alpha_r)_i \le 1 - \frac{|\mu_i - r|}{|R|-1}, \quad \forall r \in \{1, ..., |R|\}, i \in \{1, ..., |E|\}, \tag{19}$$

and

$$\|\alpha_r\|_1 \ge 1, \quad \forall r \in \{1, ..., |R|\}, \tag{20}$$

are used to avoid solution in which no block section is assigned to some region(s). Specifically, in (19), if the i^{th} block section in set E is assigned to region r, i.e., $\mu_i = r$, then the binary variable $(\alpha_r)_i = 1$; otherwise, $(\alpha_r)_i = 0$. In (20), we ensure that at least one block section is assigned to each region. As a result, (19) and (20) imply that the number of the resulting regions must equal the given number $|R|$. An illustrative example is provided in Appendix C to explain the above formulations.

Appendix C An illustrative example

In this appendix, we use a small instance to explain the proposed decomposition methods and algorithms. As illustrated in Figure 7, the instance includes 4 trains following the pre-

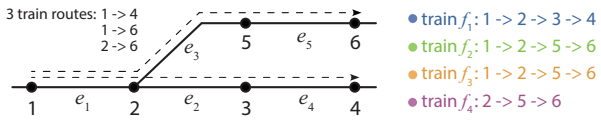


Figure 7: A small instance

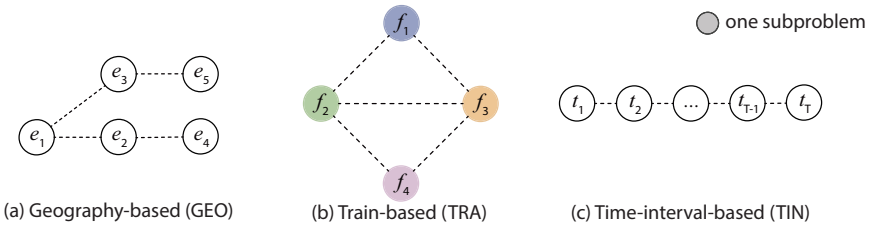


Figure 8: Subproblems and couplings

defined routes, i.e., train $f_1 : 1 \to 2 \to 3 \to 4$, train f_2 and $f_3 : 1 \to 2 \to 5 \to 6$, and train $f_4 : 2 \to 3 \to 4$.

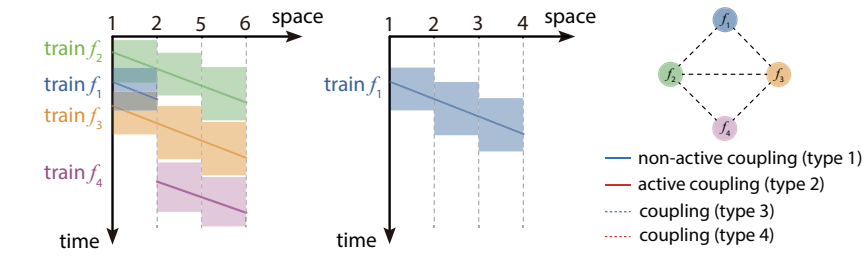
We now illustratively explain the formulation of the ILP problem proposed in Appendix B. We can write the set of block sections as $E = \{e_1, e_2, e_3, e_4, e_5\}$. The route matrix B_{f_1} and the variable vector β_{f_1} for train f_1 and the variable vector μ for block sections can be expressed as

$$B_{f_1} = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 \end{bmatrix}, \ \beta_{f_1} = \begin{bmatrix} (\beta_{f_1})_1 \\ (\beta_{f_1})_2 \end{bmatrix}, \text{ and } \mu = \begin{bmatrix} \mu_1 & \mu_2 & \mu_3 & \mu_4 & \mu_5 \end{bmatrix}^\top.$$

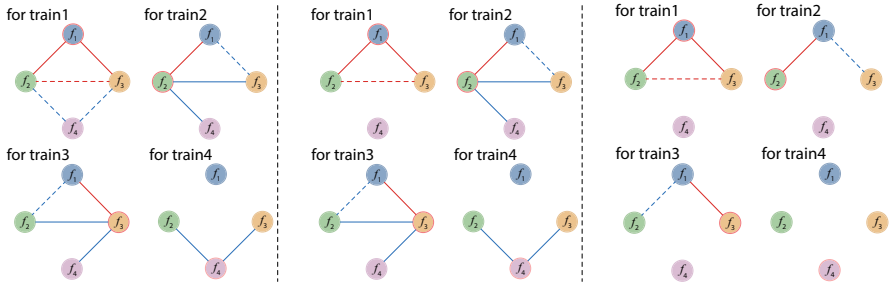
Consider the consecutive block sections e_1 and e_2 in the route of train f_1; the (17) results in the inequality $\frac{|\mu_1 - \mu_2|}{|R|-1} \leq (\beta_{f_1})_1$. If the two block sections belong to the same region, i.e., $\mu_1 = \mu_2$, then we will have $(\beta_{f_1})_1 = 0$ (as we are solving a minimization problem). If block sections e_1 and e_2 belong to different regions, i.e., $\mu_1 \neq \mu_2$, then we will have $(\beta_{f_1})_1 = 1$, as the left-hand side of the inequality is strictly in range $[0, 1)$ and B_{f_1} is an integer matrix. Constraints (19)-(20) are used to avoid the solutions like $\mu = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \end{bmatrix}^\top$.

We now illustrate the three decomposition methods. Let us assume $|R| = 5$, i.e., 5 regions and each region contains only one block section, and denote T as the number of subproblems for the TIN decomposition. By applying the three propose decomposition methods, the resulting subproblems and (primary) couplings can be shown in Figure 8. As illustrated, the GEO decomposition results in 5 subproblems, corresponding to 5 block sections respectively; the TRA decomposition leads to 4 subproblems, corresponding to 4 trains respectively; and the TIN decomposition gives T subproblems connected in an order of time horizon.

We now illustrate the three options for defining the four types of couplings in the CDRSBK algorithm with the TRA decomposition. Let us assume an infeasible timetable shown in Figure 9(a), which can be generated by independently scheduling trains one-by-one without considering their couplings. The three options are illustrated in Figure 9(b)-Figure 9(d) respectively. Let us now focus on train f_1 (i.e., subproblem f_1) to explain. In Opt_1, couplings between f_1 and f_2 is recognized as active coupling (Type_2), because train f_1 has conflict with train f_2 in the timetable shown in Figure 9(a). Both f_2 and f_3 are actively coupling subproblem of f_1; so a Type_3 coupling exists between f_2 and f_3. Train f_1 and train f_4 use completely different block sections. So subproblem f_4 only has couplings with f_2 and f_3, and their couplings are recognized as a Type 3 coupling for subproblem f_1. Train f_2 uses same block sections with all the other trains, but only has conflict with train f_1; therefore, when we focus on train f_2, the coupling between f_2 and f_1 is considered to be Type_2 and the coupling between f_2 and f_3 (and f_4) is recognized as Type_1. In Opt_2, still focusing on subproblem f_1, as the coupling between f_2 and f_4 is a non-active



(a) An infeasible timetable with some conflicts



(b) Opt_1 (c) Opt_2 (d) Opt_3

Figure 9: Three options of the CDRSBK algorithm with the TRA decomposition

coupling (Type_1, when focusing on subproblem f_2 or f_4), we consider the Type 3 coupling between f_2 and f_4 do not exist, as same as the Type_3 coupling between f_3 and f_4. In Opt_3, we consider no coupling if no conflict, which can be simply explained as removing all Type_1 couplings based on the coupling architecture of Opt_2. However, Type_3 and Type_4 couplings are generally defined, same to Opt_1 (and Opt_2).

References

Beltran Royoa, C., Heredia, F. J., 2002. "Unit commitment by augmented Lagrangian relaxation: Testing two decomposition approaches". *Journal of Optimization Theory and Applications*, *112*, 295–314.

Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J., 2011. "Distributed optimization and statistical learning via the alternating direction method of multipliers". *Foundations and Trends in Machine Learning*, *3*, 1–122.

Brännlund, U., Lindberg, P. O., Nou, A., Nilsson, J.-E., 1998. "Railway timetabling using Lagrangian relaxation". *Transportation science*, *32*, 358–369.

Cacchiani, V., Huisman, D., Kidd, M., Kroon, L., Toth, P., Veelenturf, L., Wagenaar, J., 2014. "An overview of recovery models and algorithms for real-time railway rescheduling". *Transportation Research Part B: Methodological*, *63*, 15–37.

Corman, F., D'Ariano, A., Hansen, I. A., Pacciarelli, D., 2011. "Optimal multi-class rescheduling of railway traffic". *Journal of Rail Transport Planning and Management*, *1*, 14–24.

Corman, F., Meng, L., 2015. "A review of online dynamic models and algorithms for railway traffic management". *IEEE Transactions on Intelligent Transportation Systems*, *16*, 1274–1284.

D'Ariano, A., Pacciarelli, D., Pranzo, M., 2007. "A branch and bound algorithm for scheduling trains in a railway network". *European Journal of Operational Research*, *183*, 643–657.

Findler, N. V., Stapp, J., 1992. "Distributed approach to optimized control of street traffic signals". *Journal of Transportation Engineering*, *118*, 99–110.

Kersbergen, B., van den Boom, T., De Schutter, B., 2016. "Distributed model predictive control for railway traffic management". *Transportation Research Part C: Emerging Technologies*, *68*, 462–489.

Kuwata, Y., How, J. P., 2011. "Cooperative distributed robust trajectory optimization using receding horizon MILP". *IEEE Transactions on Control Systems Technology*, *19*, 423–431.

Lamorgese, L., Mannino, C., Piacentini, M., 2016. "Optimal train dispatching by benders'-like reformulation". *Transportation Science*, *50*, 910–925.

Luan, X., Wang, Y., De Schutter, B., Meng, L., Lodewijks, G., Corman, F., 2018. "Integration of real-time traffic management and train control for rail networks-Part 1: Optimization problems and solution approaches". *Transportation Research Part B: Methodological*, *115*, 41–71.

Meinel, M., Ulbrich, M., Albrecht, S., 2014. "A class of distributed optimization methods with event-triggered communication". *Computational Optimization and Applications*, *57*, 517–553.

Meng, L., Zhou, X., 2014. "Simultaneous train rerouting and rescheduling on an N-track network: A model reformulation with network-based cumulative flow variables". *Transportation Research Part B: Methodological*, *67*, 208–234.

Nedic, A., Ozdaglar, A., 2010. "Cooperative distributed multi-agent optimization". *Convex Optimization in Signal Processing and Communications*, *340*.

Negenborn, R. R., De Schutter, B., Hellendoorn, J., 2008. "Multi-agent model predictive control for transportation networks: Serial versus parallel schemes". *Engineering Applications of Artificial Intelligence*, *21*, 353–366.

Wangermann, J. P., Stengel, R. F., 1996. "Distributed optimization and principled negotiation for advanced air traffic management". In *Proceedings of the IEEE International Symposium on Intelligent Control* (pp. 156–161).

Reducing the Adaptation Costs of a Rolling Stock Schedule with Adaptive Solution: the Case of Demand Changes

Rémi Lucas [a,b,1], Zacharie Ales [b], Sourour Elloumi [b], François Ramond [a]
[a] SNCF Innovation & Recherche, St-Denis, France
[b] Unité de Mathématiques Appliquées (UMA), ENSTA ParisTech, Palaiseau, France
[1] E-mail: remi.lucas@sncf.fr

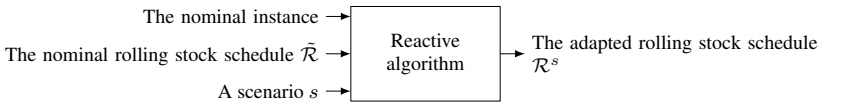
Abstract
In railway scheduling, a nominal traffic schedule is established well in advance for the main resources: train-paths, rolling stock and crew. However, it has to be adapted each time a change in the input data occurs. In this paper, we focus on the costs in the adaptation phase. We introduce the concept of *adaptive nominal solution* which minimizes adaptation costs with respect to a given set of potential changes. We illustrate this framework with the rolling stock scheduling problem with *scenarios* corresponding to increasing demand in terms of rolling stock units. We define adaptation costs for a rolling stock schedule and propose two MILPs. The first one adapts, at minimal cost, an existing rolling stock schedule with respect to a given scenario. The second MILP considers a set of given scenarios and computes an adaptive nominal rolling stock schedule together with an adapted solution to each scenario, again while minimizing adaptation costs. We illustrate our models with computational experiments on realistic SNCF instances.

Keywords
Rolling Stock, Adaptive Solution, Discrete Optimization.

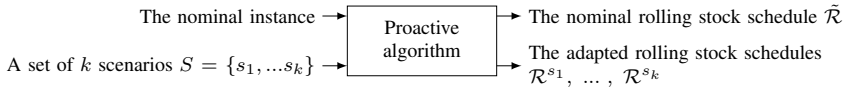
1 Introduction

Railway scheduling is generally divided into different problems, which are solved sequentially. The *line planning* problem computes train lines based on the existing rail network, defining a list of stations and an associated frequency for each line. The *timetabling* problem defines a set of trains with departure and arrival times for each station of the considered lines, with respect to the frequency, providing a complete feasible timetable. The *rolling stock scheduling* problem defines compositions for each train, assigning physical rolling stock units to the given input timetable. The *crew scheduling* problem operates in a similar manner, assigning crew members (*e.g.* train drivers) to each train and each station with respect to specific legal constraints. Finally, the *platforming* problem is solved for each station, assigning a track to each train stopping by or passing through it during the planning horizon.

Ideally, we would like to solve most of these problems together in an integrated manner a few days before the date of operations; in practice, these problems are solved sequentially several years or months in advance for historical, legal and practical reasons. Thus, a complete *nominal* schedule is built some months in advance for each railway resource: train-paths, rolling stock and crew.



(a) The Rolling Stock Adaptation Problem



(b) The Adaptive Rolling Stock Scheduling Problem

Figure 1: The two different methods described in this paper

However, changes may occur afterwards, either prior or during operations. They may either concern the availability of resources, such as infrastructure blockades or rolling stock failures, or some new requirements, such as additional trains to schedule or some changes in their required compositions.

Whenever changes occur, the schedules must be updated. We focus on midterm changes during the *adaptation* phase, which corresponds to rescheduling of resources a few weeks or months before operational time. A schedule can be adapted many times if changes during the adaptation phase are frequent. We describe the changes with the notion of *scenario*, corresponding to a modification in the input data. An *adapted* schedule with regards to this scenario is then computed, in order to satisfy the changes. Our main objective is to reduce the total cost of the adaptation phase.

The rest of this paper is organized as follows. Section 2 is dedicated to a literature review on the main issues discussed in this paper, with a focus on rolling stock resource. We describe in Section 3 the *adaptation costs* in general railway scheduling, and propose a new approach to assessing adaptation costs for the *rolling stock* resource. Besides considering the performance of the new schedule, we also consider *structural* adaptation costs to assess the differences between the nominal and the adapted schedules. We introduce in Section 4 the *Rolling Stock Adaptation Problem* with respect to a given *scenario* in the case of *demand* changes. This is the *reactive* problem appearing during the adaptation phase where a given scenario is revealed (see Figure 1a). A MILP formulation based on the literature review is proposed. In Section 5, we define the notion of *adaptive* nominal solution with regards to a set of scenarios. A nominal solution in the conception phase is said to be adaptive with respect to a set of scenarios if its adaptation cost to each of these scenarios is low. The corresponding *proactive* problem (see Figure 1b) appears in the conception phase where information is available about the probability of occurrence of certain possible scenarios. The *Adaptive Rolling Stock Scheduling Problem* is introduced and a MILP is proposed to solve it. We present in Section 6 computational experiments with realistic instances of SNCF, the major French train operating company. Finally, Section 7 concludes and highlights future perspectives.

2 Literature Review

Models for Rolling Stock Scheduling

There exist a lot of models to schedule rolling stock, with different assumptions.

Fioole *et al.* (2006) introduce the Rolling Stock Circulation Problem, defining a MILP with variables affecting a unique *composition* to each trip. They define a dedicated event graph and obtain a flow formulation. Each trip has one or two successor trips defined as input, which is a strong assumption because it restricts the possibilities.

Cacchiani *et al.* (2010) introduce the *Train Unit Assignment Problem*. They define a graph where each node corresponds to a trip. The authors solve a flow problem with a path formulation, and propose some improvements for the linear relaxation by describing the convex hull of a set of constraints. In this paper, we use a graph similar to this one in order to formulate the MILPs. However, we use a flow formulation, which is more relevant to model adaptation costs.

Giacco *et al* (2014) introduce a rolling stock scheduling problem integrating the maintenance requirements and the empty moves possibility. They define a dedicated graph and propose a MILP to compute a set of hamiltonian paths respecting the maintenance constraints.

Borndörfer *et al.* (2016) introduce a novel approach to schedule railway vehicle rotations. They define a generic *hypergraph* where each train has a departure and an arrival node for each possible composition. Oriented hyperarcs are defined between two set of nodes of two different trains, and indicate the possibility to cover these trains with the same rolling stock units. A MILP formulation is proposed with additional maintenance requirements. It is solved with a dedicated algorithm using column generation and rapid branching heuristics.

Adaptation Costs of a Rolling Stock Schedule

Many papers address real-time disruption management of rolling stock. These *rescheduling* models generally use models deriving from those presented above for rolling stock scheduling. They mostly model the rolling stock adaptation costs by assessing the new performance of the adapted rolling stock schedule and try to minimize the new shunting operations.

Nielsen *et al.* (2012) propose a generic framework for rolling stock rescheduling with rolling horizon approach based on the Rolling Stock Circulation Problem of Fioole *et al.* (2006). They assume major disruptions (infrastructure blockades) and try to reschedule the rolling stock with a dedicated real-time heuristic. Their main objective is to minimize the cancelled trips because of a lack of rolling stock. More recently, Wagenaar *et al.* (2017) propose a MILP formulation based on Fioole's model for the Rolling Stock Rescheduling Problem, while considering dead-head trips (empty moves) possibility and dynamic passenger demands. It allows respectively to decrease the number of cancelled trips and to capture the fact that a cancelled trip will have influence on the passenger demand for the next trip with the same origin and destination. Lusby *et al.* (2017) propose an original approach to solve a rescheduling problem with a dedicated Branch&Price framework. It is based on a path formulation with specific constraints representing operational requirements.

Some papers deal with changing circumstances in the short-term planning stage. Ben-Khedher *et al.* (1998) describe the *Capacity Adjustment Problem*: considering the number of reservations for each train and some forecasts of a yield management system, they try to adjust the compositions of the scheduled trains in order to maximize the expected profit.

The model computes a feasible schedule with these new compositions, but adaptation costs are not explicitly taken into account.

Lingaya *et al.* (2002) propose a MILP model to schedule locomotives and carriages a few days before operations. They consider a changing (static) demand in terms of cars and specific operational constraints in their problem such as maintenance requirements or minimum connection times. They try to modify the current rolling stock schedule to fit these demand changes and operational constraints. They do not explicitly focus on structural adaptation costs, but they consider it implicitly: they only accept to make changes in the *car cycles*, and do not modify the *locomotive* schedules. Thus, changes are limited, and structural adaptation costs are restricted.

Budai *et al.* (2010) address the rolling stock rebalancing problem. They suppose a lack of units at certain stations at the end-of-day and a surplus at other stations, and try to reduce these off-balances by rolling stock rescheduling. Adaptation costs correspond to the classical nominal performance costs and the changes in shunting plans.

More recently, Borndörfer *et al.* (2017) introduce the re-optimization of rolling stock rotation while considering a reference rotation. They use a hypergraph and define a *template* as a set of trips in the reference rotation such that they are covered by the same rolling stock units. They try to keep these templates unchanged in the adapted rotations. The objective function introduces the notion of *deviation* from the reference rotation. Our definition of adaptation costs is quite similar but is based on a simpler model with an adaptive version that is easier to solve. The authors use different scenarios corresponding to infrastructure constructions where timetables slightly change, which implies to reschedule the rolling stock rotations. In this paper, we focus on demand changes and do not suppose any modification of the timetables.

Adjustable Robustness and Recoverable Robustness

The concept of adaptive solutions is closely related to the concepts of adjustable and recoverable robustness.

The concept of *adjustable* optimization was originally introduced by Ben-Tal *et al.* (2004). Following the context of bi-level stochastic optimization, they consider uncertainty set for some parameters, and solve a mathematical program with two types of variables:

- *here-and-now* variables x must be fixed at the early stage of the optimization process;
- *wait-and-see* variables y must be fixed once a scenario is revealed.

The problem is to assign values to the x variables such there exists y values with (x, y) feasible for any realization of the uncertainty set. For this purpose, the authors introduce variables $y(\xi)$ for each ξ in the uncertainty set, and show that this problem is untractable in the general case.

If we consider wait-and-see variables y corresponding to a *recourse* of the here-and-now variables x, we obtain the concept of recoverable robustness, originally introduced by Liebchen *et al.* (2009). The authors consider an uncertainty set with finite support such that it corresponds to a finite set of *scenarios* S. They describe the recourse variables y as a *recourse algorithm* $\mathcal{A}$. If we consider the generic mathematical program minimizing $f(x)$ subject to a feasibility set for vector x, the associated recoverable robust problem aims to find a solution x and an algorithm $\mathcal{A}$ such that $y = \mathcal{A}(x, s)$ is feasible for each scenario $s \in S$. Algorithm $\mathcal{A}$ must be chosen in a class of algorithms and can have a certain recovery cost to be added to the objective function. Cicerone *et al.* (2009) describe such class for $\mathcal{A}$. For example, $\mathcal{A}$ has to run within a maximal time limit. Our model presented Section 5 is a

recoverable robust model where algorithm $\mathcal{A}$ is a MILP, with recourse costs corresponding to the differencies between solutions x and y.

Recoverable robustness was originally applied in railway scheduling. Recoverable robust timetabling was introduced by Liebchen *et al.* (2009) and Cicerone *et al.* (2009), where the uncertainty concerns minimal required time between several pairs of arrival and departure times of a train. The authors compute nominal recoverable robust schedules and propose different classes of algorithms to reschedule trains. These authors also consider applications to platforming and shunting yard problems.

A recoverable robust rolling stock scheduling problem is addressed by Cacchiani *et al.* (2012) with uncertainties corresponding to infrastructure blockade. The authors propose a large MILP based of the model of Fioole *et al.* (2006). They duplicate the nominal variables for each scenario, and minimize both the performance of the nominal solution and the maximal recovery costs among the scenarios. The recovery costs for a given scenario are described with the cancelled trips, the off-balanced units at end-of-days, and the new shunting operations. The authors use Benders decomposition to compute optimal solutions for the relaxed problem, and develop a dedicated Benders heuristic to compute integer solutions to the recoverable robust problem. Our model is based on a different formulation and does not suppose operational disruptions, focusing on demand changes in the adaptation phase. Moreover, our adaptation costs allow to maximize the similarities between the nominal and adapted rolling stock schedules. Another difference is that we minimize the expected adaptation cost instead of the worse one among the scenarios.

3 The Adaptation Costs

In this section, we describe in more details the adaptation phase in railway scheduling. We identify different performance criteria to evaluate the quality of an adaptation, and propose a simple way to evaluate the performance of a rolling stock schedule adaptation.

3.1 Adaptation Costs in General Railway Scheduling

We identify three types of "costs" in the adaptation phase for any railway resource.

1. *Performance cost*
   An adapted schedule has to be assessed with regards to the classical performance criteria. For example, if there is a change in the timetable, the adapted timetable must maximize the passenger satisfaction. However, finding an optimal solution is not crucial in a rescheduling process: we generally look for an acceptable schedule. These performance costs only depend on the adapted schedule, and we can compute them without any information about the nominal one.
2. *Direct costs during the adaptation phase*
   During the adaptation phase, each request of change may impact several departments. They have to look for a new acceptable schedule compatible with the new requirements. It can be difficult and impact different resources, and it implies communication between the departments, which can be interpreted as a direct adaptation cost.
3. *Indirect operational costs*
   Each schedule is generally repeated with a specific horizon (an hour, a day or a week). An adaptation concerns some periods where the schedules are quite different. Thus, an adaptation can have operational consequences. The more different the adapted

schedule from the nominal one, the higher the risk of human error at operational time, which would lead to bad performance, or an increased risks of incidents. We can interpret this as an indirect operational cost of an adaptation.

Let us observe that reducing the first type of cost can lead to an increasing of the two others. Indeed, if we want to have a good performance cost for the adapted schedule, we have to consider the rescheduling of a higher number of resources. It implies a lot of communication between the departments and is responsible for a higher direct cost during the adaptation phase. Moreover, the adapted schedule will probably be very different from the nominal one, implying a higher risk of operational errors at operational time and an increase in the indirect operational costs.

Furthermore, if we force the adapted schedule to be *similar* to the nominal one, we find that it reduces indirect operational costs, but it has also a strong positive impact on the direct cost during the adaptation phase. Indeed, if we want the adapted schedule to be similar to the nominal one, we have to look for an adapted schedule in a smaller solution search space, and it reduces both the number of implied departments and the communication between them. Thus, this notion of similarity between the schedules is the relevant criterion to maximize, or in other words, minimizing the changes between the schedules captures both the direct costs in the adaptation phase and the indirect operational costs. Consequently, we define two complementary types of cost in the adaptation phase:

- the *performance* adaptation costs to evaluate the quality of the adapted schedule with regards to classical nominal performance criteria;
- the *structural* adaptation costs to evaluate the similarities and the differences between the adapted schedule and the nominal one.

3.2 Adaptation of a Rolling Stock Schedule

Performance Adaptation Costs

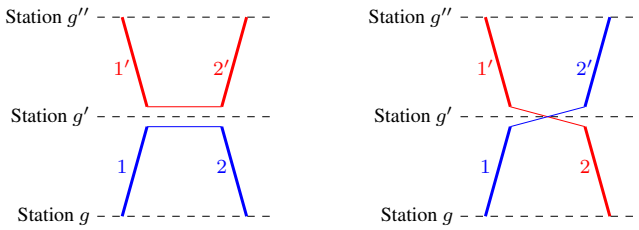
As previously mentioned, the non-optimality of an adapted rolling stock schedule for the classical nominal performance criteria is a first type of adaptation costs. Concerning the rolling stock resource, one generally has to minimize the following criteria:

- The total *lack* of rolling stock units: it is sometimes impossible to propose a schedule with a sufficient number of units for all trains, and we try to minimize the number of missing units;
- The number of *engaged rolling stock units*;
- The number of kilometers of *dead-head trips* for each unit, which correspond to trips between two stations without any passenger (empty moves);
- The number of kilometers of *over-compositions* for each unit, which correspond to trips with higher number of units than required.

Structural Adaptation Costs

In the literature structural adaptation costs are generally defined as the modifications in the shunting plans: if two additional trains have to be combined in the adapted shunting plan, it has indeed a certain operational cost.

We introduce a new definition of structural adaptation costs via the notion of *successions* between trains. Suppose there is a rolling stock unit of type m that covers Train 1, and then covers Train 2 without any train between 1 and 2. In particular, it implies that the arrival



(a) Nominal schedule (b) Adapted schedule

Figure 2: An example of changes in the successors. In the nominal schedule Figure 2a, Train 2 is the successor of Train 1 and Train 2' is the successor of Train 1'. In the adapted schedule Figure 2b, Train 2 is now the successor of Train 1' and Train 2' the successor of Train 1.

station of Train 1 is the same as the departure station of Train 2. Then, Train 2 is a *successor* of Train 1 for type m, and the *succession* 1-2 exists for this unit type.

More precisely, for each couple of trains i and j and for each unit type m, we define the binary value

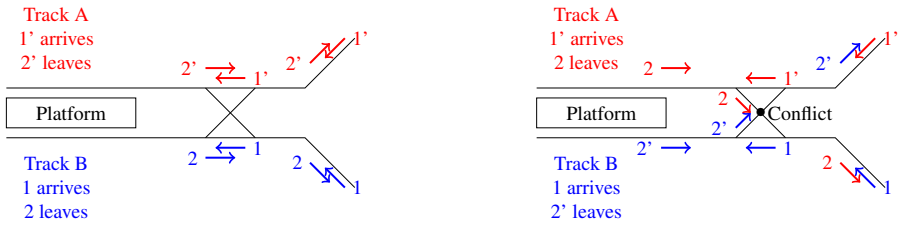
$$Succession(i,j,m) = \begin{cases} 1 & \text{if at least one unit of type } m \text{ is affected to} \\ & i \text{ and } j \text{ without any train between } i \text{ and } j \\ 0 & \text{otherwise} \end{cases} . \quad (1)$$

Let us consider the example of Figure 2 with four trains: 1, 2, 1' and 2'. Suppose Train 2 is the unique successor of Train 1 and Train 2' is the unique successor of Train 1' in the nominal schedule, as shown in Figure 2a. If Train 2 is not anymore a successor of Train 1 in the adapted schedule but a successor of Train 1', as shown in Figure 2b, it will change the *structure* of the rolling stock schedule, and it implies several adaptations.

First, it may change the track-occupation diagram for Trains 1, 2, 1' and/or Train 2'. Trains 1 and 2' (*resp.* 1' and 2) *must* now be scheduled on the same track if there is not enough time to make a shunting movement. Thus, it impacts the passenger information in stations, and is a potential source of bad operational performance. Moreover, it could be difficult to find a new track-occupation diagram with associated paths compatible with the new successions, as shown in Figure 3. It implies more rescheduling effort and is responsible for an increase in adaptation costs.

Second, it possibly modifies the driver schedules in the crew scheduling problem, because they strongly depend on the successors in the rolling stock schedule. If Train 2 is the successor of Train 1, it is convenient that the same driver is assigned to these two trains. Otherwise, the solution is less robust. Indeed, the example of Figure 4 shows that having different successions for the rolling stock and the drivers can lead to a higher number of impacted trains in case of a primary delay, and thus an increase in indirect operational adaptation costs. It is possible to avoid this by changing the drivers schedule, but it increases direct adaptation costs because the rescheduling effort is more important.

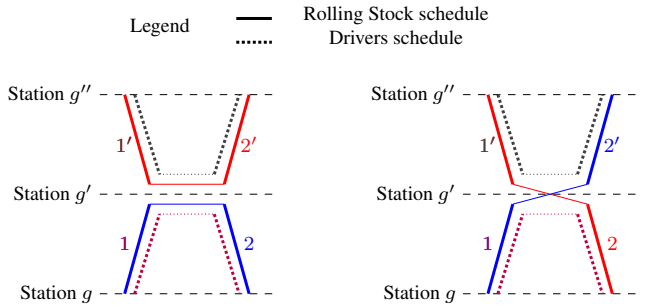
And third, a change in successors may involve changes in the shunting plan, as shown Figure 5. Indeed, if a train has $n \geq 1$ different successors (*resp.* predecessors), it is necessary to make $n-1$ combinations (*resp.* splits) after it arrives (*resp.* before it leaves). Thus, a change in the successors can impact the number of splits and combinations.



(a) Nominal track-occupation diagram with associated conflict-free paths

(b) A simple adapted track-occupation diagram with a path-conflict

Figure 3: An example of conflict in the track-occupation diagram in station g' after the change in the successors in Figure 2. Trains 1 and 2' (*resp.* Trains 1' and 2) have to be scheduled on the same track in the adapted case 3b. If Trains 2 and 2' leave g' at the same time, it is impossible to adapt the track-occupation diagram, because the paths for Trains 2 and 2' are incompatible.



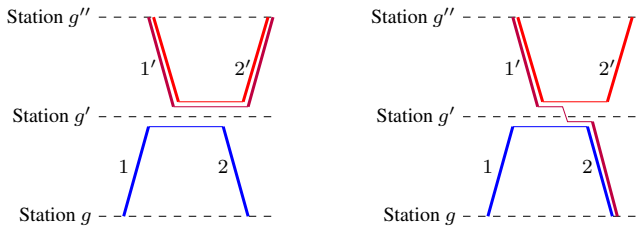
(a) Nominal rolling stock schedule and drivers schedule

(b) Adapted rolling stock schedule and unchanged drivers schedule

Figure 4: An example of 4 trains with different successions for the rolling stock and the drivers schedules. In Figure 4b, a primary delay of Train 1 may imply a delay propagation to Train 2 because the two trains have the same rolling stock unit. In the adapted schedule Figure 4b, a primary delay of Train 1 may imply both a delay propagation to Train 2 (because they have the same driver) and to Train 2'(because they have the same unit). It is possible to avoid this by changing the driver schedule, which is a source of additional adaptation costs.

Following these observations, we define structural adaptation costs to move from one rolling stock schedule to another as the differences in the successions between them:

$$StructuralAdaptationCosts = \sum_{(i,j,m)} \sum \sum \left| Succession^{adapted}(i,j,m) - Succession^{nominal}(i,j,m) \right| \quad (2)$$



(a) Nominal schedule (b) An adapted schedule

Figure 5: In the nominal schedule Figure 5a, there is not any split or combination. In the adapted schedule Figure 5b, there is a split after the arrival of Train 1' and a combination before the departure of Train 2. These new shunting operations correspond to the new succession 1'-2.

4 The Rolling Stock Adaptation Problem with respect to a Given Scenario

In this section, we introduce the Rolling Stock Adaptation Problem more precisely, with a detailed description in the case of demand changes, and propose a mathematical formulation to solve it with a MILP.

4.1 Problem Description

In this paper, we focus on one of the main causes for which the rolling stock schedules have to be adapted: the demand changes. Whether it is passenger or freight transportation, there is always an uncertainty about the minimal demand of the trains. Thus, the number of required units for a given train can change during the adaptation phase.

In passenger railway transportation, a forecasted passenger demand is computed in the conception phase for each train. However, if the number of reserved seats is closely monitored, it is possible to update this forecast some weeks or months before the departure of the train and adjust the number of units depending on the evolution of the forecast. In freight railway transportation, the quantity of goods that need to be transported varies slightly from week to week, because of a more or less favourable economical context. Thus, freight transportation is also concerned by the need to adapt the rolling stock schedules because of demand changes.

Let us introduce some notations. The input of the Rolling Stock Adaptation Problem with respect to a given scenario is:

- a set $\mathcal{G}$ of stations where combinations and splits may be allowed: for each station $g \in \mathcal{G}$, we define the parameter $CS(g) \in \{0; 1\}$ with value 1 if splits and combinations are allowed in station g, and 0 otherwise;
- a set $\mathcal{M}$ of unit types and, for each $m \in \mathcal{M}$, $K_m \in \mathbb{N}$ is the number of available units of type m;
- a set of trains T^{train}. A train $i \in T^{train}$ is defined by:
  - fixed departure and arrival times;
  - fixed departure and arrival stations in $\mathcal{G}$;

  - $\tilde{D}_i > 0$, the nominal demand in terms of rolling stock units, *i.e.* the desired number of rolling stock units for the train i;
  - $D_i^{max} \geq \tilde{D}_i$, the maximal number of rolling stock units for the train i;
  - $\mathcal{M}_i$, the list of unit types compatible with train i.
- a time horizon H expressed in days, numbered $0, 1, \ldots H-1$;
- a nominal rolling stock schedule $\tilde{\mathcal{R}}$;
- a *scenario* s, corresponding to a set of updated demands $D_i^s \in [\![\tilde{D}_i\,; D_i^{max}]\!]$ in terms of rolling stock units for each train i.

The Rolling Stock Adaptation Problem is described in Figure 1a. Considering an input instance as described above, find a *feasible* rolling stock schedule $\mathcal{R}^s$ which minimizes both the performance and structural adaptation costs defined in Section 3. A rolling stock schedule is said to be feasible if it respects some particular constraints we will describe below.

4.2 Mathematical Formulation

The Rolling Stock Adaptation Problem with respect to a given scenario is formulated as a multicommodity flow problem in a graph G similar to the one used by Cacchiani *et al.* (2010), where each unit type m has a corresponding flow in the multiflow.

Description of the Graph
We define $G = (T, A)$ as a directed graph in which the nodes correspond to *tasks* that can be performed (see example in Figure 6), which is directly inspired by the graph of Löbel (1998) in vehicle scheduling. The tasks can be decomposed as follows:

$$T = T^{train} \cup T^{depots} \cup \{\alpha,\, \omega\}\,, \tag{3}$$

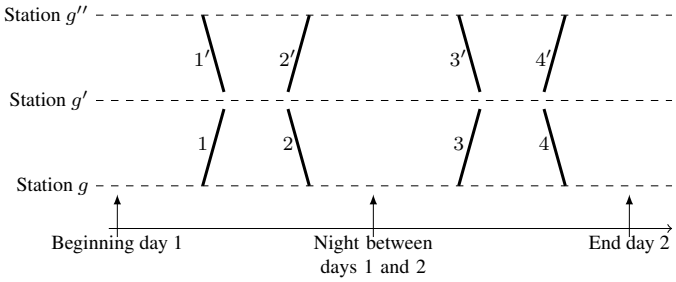
where:

- a node $i \in T^{train}$ corresponds to a train, as defined above;
- a node $i \in T^{depot}$ corresponds to a depot task, which is characterized by a station and two consecutive days. Performing this task means to put some units into the depot during the corresponding night. Consequently, if there is a physical depot at a given station g, we define $H+1$ nodes for it, with labels g^0, $g^1 \ldots g^H$. They respectively correspond to the depot in station g in the morning of day 0, during the night between day 0 and day 1, ... and finally in the evening of day $H-1$;
- node α is the *source* node, and node ω is the *sink* node.

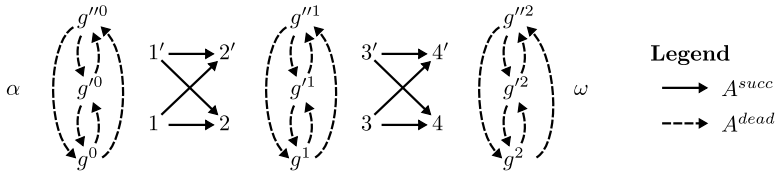
The introduction of the set T^{depot} allows to reduce the number of arcs in the graph and thus the complexity of our formulation.

The arc set A contains several types of arcs:

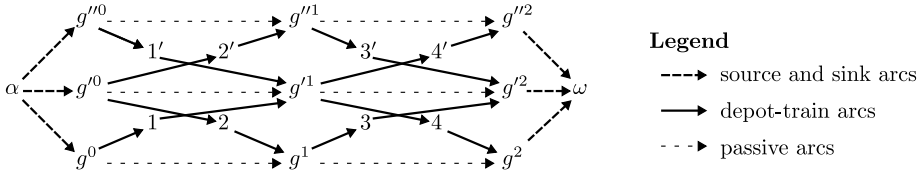
- arcs A^{succ} are the most important arcs, corresponding to *successions* between two trains as defined in Section 3.2.
- arcs A^{dead} between two depot nodes g^0 and g'^0 for two different stations g and g', corresponding to dead-head trips between g and g' during the night. Note that it is impossible to make dead-head trips during a day, or equivalently between two trains operating on the same day;
- arcs between the source node α and a depot node g^0 for a given station g. The value of these arcs in the flow of unit type m corresponds to the number of units of type m



(a) The input timetable



(b) Subgraph with arcs from A^{succ} and A^{dead}



(c) Subgraph with the other arcs, except the fictive sink-source arc (ω, α)

Figure 6: Graph construction for an example with 8 trains and 3 stations over a scheduling horizon of 2 days

starting from the associated depot at the beginning of the horizon. We also define arcs between a depot node g^H for a given station g and the sink node ω, corresponding to units at this physical depot at the end of the horizon;

- arcs between a depot node g^d for a given station g and a train leaving g on day d. The flow value for type m corresponds to the number of units starting with this train on day d. We also define arcs between a train leaving a station g on day d and the depot node g^{d+1}, corresponding to units going into the depot after covering the train;
- passive arcs between two depot nodes of the same station g with consecutive days (g^d,g^{d+1}). They correspond to units staying at the depot during a whole day, without covering any train during this day;
- a fictive arc $(\omega,\ \alpha)$, which is not represented in Figure 6.

In the rest of this paper, we extend the previous definition of $\mathcal{M}_i$ initially defined for all $i \in T^{train}$ to all the nodes $i \in T$, and denote by $\mathcal{M}_i$ the set of unit types compatible with the node $i \in T$. Moreover, for each arc $(i, j) \in A$, we define the set $\mathcal{M}_{ij}$ as $\mathcal{M}_i \cap \mathcal{M}_j$.

Variables Defining a Rolling Stock Schedule

A rolling stock schedule $\mathcal{R}$ can be described with the only integer decision variables x_{ijm} representing flow value of type m for each arc $(i,j) \in A$, and for each unit type $m \in \mathcal{M}_{ij}$.

We introduce the following auxiliary variables in order to get a linear formulation:

- l_i which counts the lack of units for train i if its demand is D_i. More formally,

$$l_i = \max\ (0, D_i - \sum_{\substack{j \in T \\ (i,j) \in A}} \sum_{m \in \mathcal{M}} x_{ijm}); \tag{4}$$

- δ_{ijm}, a binary variable equal to 1 if and only if $x_{ijm} \geq 1$, and 0 otherwise
- δ'_{ij}, a binary variable equal to 1 if and only if at least one variable δ_{ijm} is equal to 1 for all the unit types m.

Let us remark that from variables (x, l, δ, δ') one can complete the description of a rolling stock schedule through path decomposition. However, this partial description is sufficient to describe performance and structural adaptation costs.

Basic Feasibility for a Rolling Stock Schedule

A rolling stock schedule $\mathcal{R} = (x, l, \delta, \delta')$ is said to be *basic-feasible* for the demand vector D if it is feasible with respect to the common strong constraints defined by the following set of inequalities $\mathcal{F}_D$:

$$\sum_{\substack{i \in T \\ (i,h) \in A : m \in \mathcal{M}_i}} x_{ihm} = \sum_{\substack{j \in T \\ (h,j) \in A : m \in \mathcal{M}_j}} x_{hjm} \qquad h \in T,\ m \in \mathcal{M}_h \tag{5}$$

$$\sum_{\substack{j \in T \\ (i,j) \in A}} \sum_{m \in \mathcal{M}_{ij}} x_{ijm} \geq D_i - l_i \qquad i \in T^{train} \tag{6}$$

$$\sum_{\substack{j \in T \\ (i,j) \in A}} \sum_{m \in \mathcal{M}_{ij}} x_{ijm} \leq D_i^{max} \qquad i \in T^{train} \tag{7}$$

$$x_{\omega\alpha m} \leq K_m \qquad m \in \mathcal{M} \tag{8}$$

$$\sum_{\substack{j \in T \\ (i,j) \in A}} \delta'_{ij} \leq 1 \qquad i \in T^{train} | CS(G^{arr}(i)) = 0 \tag{9}$$

$$\sum_{\substack{j \in T \\ (j,i) \in A}} \delta'_{ji} \leq 1 \qquad i \in T^{train} | CS(G^{dep}(i)) = 0 \tag{10}$$

$$x_{ijm} \geq \delta_{ijm} \qquad (i,j) \in A^{succ},\ m \in \mathcal{M}_{ij} \tag{11}$$

$$x_{ijm} \leq M \cdot \delta_{ijm} \qquad (i,j) \in A^{succ},\ m \in \mathcal{M}_{ij} \tag{12}$$

$$\sum_{m \in \mathcal{M}_{ij}} \delta_{ijm} \geq \delta'_{ij} \qquad (i,j) \in A^{succ} \tag{13}$$

$$\sum_{m \in \mathcal{M}_{ij}} \delta_{ijm} \leq M \cdot \delta'_{ij} \qquad (i,j) \in A^{succ} \tag{14}$$

$$x_{ijm} \in \mathbb{N} \qquad (i,j) \in A,\ m \in \mathcal{M}_{ij} \tag{15}$$

$$\delta_{ijm} \in \{0;1\} \qquad (i,j) \in A^{succ},\ m \in \mathcal{M}_{ij} \tag{16}$$

$$\delta'_{ij} \in \{0;1\} \qquad (i,j) \in A^{succ} \tag{17}$$

$$l_i \in \mathbb{R}^+ \qquad i \in T^{train} \tag{18}$$

Constraints (5) are conservative flow constraints for all nodes and all unit types. Constraints (6) force each train to be covered by a sufficient number of rolling stock units or ensure that the variable l_i has the correct value. Constraints (7) prevent a train from being covered by a number of units exceeding the capacity of the train. Constraints (8) check that the number of available rolling stock units for each unit type is respected. If $CS(g) = 0$, Constraints (9) ensure that there is no split at station g. For each train i arriving at g, i must have a unique successor train, with the possibility to have successors of different unit types. For example, a train i can be covered by two units of different unit types but will have a unique successor j, covered by the same units. Constraints (10) ensure that there is no combination after a train j leaves a station g with $CS(g) = 0$ in a similar manner. Constraints (11) – (14) ensure the correct value for the variables δ_{ijm} and δ'_{ij}, where the big-M constant M is arbitrary large. Finally, constraints (15) – (18) restrict the definition set of the variables.

Nominal Feasibility

The input nominal rolling stock schedule $\tilde{\mathcal{R}}$ can be described with the variables $(\tilde{x}, \tilde{l}, \tilde{\delta}, \tilde{\delta}')$ and is basic-feasible for the set $\mathcal{F}_{\tilde{D}}$. Moreover, a nominal feasible schedule has to respect the nominal *cyclicity* constraints:

$$\tilde{x}_{\alpha g^0 m} = \tilde{x}_{g^H \omega m} \qquad g \in \mathcal{G},\, m \in \mathcal{M}, \tag{19}$$

to ensure that such a schedule can be followed or preceded by itself. Thus, if there is $\tilde{x}_{g^H \omega m}$ units of a type m at depot node g^H in the schedule, the same number of units is required at depot node g^0.

Feasibility of an Adapted Schedule for Scenario s

A feasible solution for the Rolling Stock Adaptation Problem with respect to a given scenario s is a basic-feasible rolling stock schedule $\mathcal{R}^s = (x^s, l^s, \delta^s, \delta'^s)$ for the demand D^s such that:

- $\mathcal{R}^s \in \mathcal{F}_{D^s}$
- $\mathcal{R}^s$ respects the following constraints:

$$l_i^s \leq \max(0, D_i^s - (\tilde{D}_i - \tilde{l}_i)) \qquad i \in T^{train} \tag{20}$$

$$x^s_{\alpha g^0 m} \geq \tilde{x}_{g^H \omega m} \qquad g \in \mathcal{G},\, m \in \mathcal{M} \tag{21}$$

$$x^s_{g^H \omega m} \geq \tilde{x}_{\alpha g^0 m} \qquad g \in \mathcal{G},\, m \in \mathcal{M} \tag{22}$$

Constraints (20) deal with the quality of service. It bounds the variables l_i^s: if 2 units were affected to a train i in the nominal rolling stock schedule and if the demand is 3 in scenario s, the variable l_i^s cannot exceed the value 1=3-2, because it is unreasonable to reduce the number of units when the demand increases.

Constraints (21) and (22) are side constraints very similar to cyclicity, which are less restrictive. In practice, an adapted schedule is never followed or preceded by itself, because such adaptations are usually limited in time. Thus, it must be preceded or followed by a nominal schedule. If there is $\tilde{x}_{g^H \omega m}$ units of a type m at a depot g^H in the nominal schedule $\tilde{\mathcal{R}}$, there must be at least as many units at depot g^0 in the adapted schedule. This is the purpose of Constraints (21). Constraints (22) are similar and deal with the number of rolling stock units at the end of the horizon in the adapted schedule.

Objective Function

The objective of the Rolling Stock Adaptation Problem represents adaptation costs to move from $\tilde{\mathcal{R}}$ to the adapted rolling stock schedule $\mathcal{R}^s$. As seen in Section 3, it corresponds to the performance and structural adaptation costs.

Performance adaptation costs of a basic-feasible rolling stock schedule $\mathcal{R} = (x, l, \delta, \delta')$ can be described with the following expressions:

- the total lack of rolling stock:

$$Lack(\mathcal{R}) \triangleq \sum_{i \in T^{train}} l_i; \tag{23}$$

- the total number of engaged units:

$$Units(\mathcal{R}) \triangleq \sum_{m \in \mathcal{M}} x_{\omega\alpha m}; \tag{24}$$

- the number of dead-head trips:

$$Dead(\mathcal{R}) \triangleq \left(\sum_{(i,j) \in A^{dead}} \sum_{m \in \mathcal{M}} x_{ijm} \right); \tag{25}$$

- the number of over-compositions:

$$Over(\mathcal{R}) \triangleq \sum_{i \in T^{train}} \left(\left(\sum_{\substack{j \in T \\ (i,j) \in A}} \sum_{m \in \mathcal{M}_{ij}} x_{ijm} \right) - D_i \right). \tag{26}$$

Let us remark that the objectives $Dead(\mathcal{R})$ and $Over(\mathcal{R})$ in Equations (25) and (26) can be easily weighted with the travelled distance in kilometers.

Structural adaptation costs to move from a feasible nominal rolling stock $\tilde{\mathcal{R}}$ to a feasible adapted rolling stock schedule $\mathcal{R}^s$ are defined with the following equation:

$$Struct(\tilde{\mathcal{R}}, \mathcal{R}^s) \triangleq \sum_{\substack{(i,j) \in A^{succ} \\ i,j \in T^{train}}} \sum \left((1 - 2\tilde{\delta}_{ijm}) \cdot \delta^s_{ijm} + \tilde{\delta}_{ijm} \right), \tag{27}$$

where the expression inside the sum corresponds to a rewriting of $\left| \delta^s_{ijm} - \tilde{\delta}_{ijm} \right|$, which is true because δ are binary variables.

The objective of the Rolling Stock Adaptation Problem can be written as a sum of the different previous objectives with relevant coefficients $(\beta, \Gamma, \Delta, \zeta, \eta)$:

$$\begin{aligned} \min_{\mathcal{R}^s} \quad & \beta \cdot Lack(\mathcal{R}^s) + \Gamma \cdot Units(\mathcal{R}^s) + \Delta \cdot Dead(\mathcal{R}^s) \\ & + \zeta \cdot Over(\mathcal{R}^s) + \eta \cdot Struct(\tilde{\mathcal{R}}, \mathcal{R}^s). \end{aligned} \tag{28}$$

5 The Adaptive Rolling Stock Problem with respect to a Given Set of Scenarios

5.1 Problem Description

As described in Figure 1b, the Adaptive Rolling Stock Scheduling Problem is a recoverable robust problem for rolling stock scheduling.

A solution is said to be *adaptive* for a set of scenarios S if its expected adaptation costs are low for that set of scenarios. In the following, we suppose without lost of generality that the scenarios have the same probability of occurrence.

5.2 Mathematical Formulation

Our mathematical formulation for the Adaptive Rolling Stock Scheduling Problem is based on that of the Rolling Stock Adaptation Problem in Section 4. Following Cacchiani *et al.* (2012) in a different setting, we duplicate the nominal variables $(\tilde{x}, \tilde{l}, \tilde{\delta}, \tilde{\delta}')$ for every scenario $s \in S$ and obtain a MILP with a higher dimension.

Feasible Solution
A feasible solution for the adaptive Rolling Stock Scheduling Problem is composed of:

- a nominal feasible rolling stock schedule $\tilde{\mathcal{R}} = (\tilde{x}, \tilde{l}, \tilde{\delta}, \tilde{\delta}') \in \mathcal{F}_{\tilde{D}}$ that satisfies Equations (19);
- a collection of adapted feasible schedules $(\mathcal{R}^s)_{s\in S} = (x^s, l^s, \delta^s, \delta'^s)_{s\in S}$, where each $\mathcal{R}^s$ is in $\mathcal{F}_{D^s}$ and satisfies Constraints (20) – (22) for scenario s.

Thus, the corresponding MILP contains variables $(\tilde{x}, \tilde{l}, \tilde{\delta}, \tilde{\delta}')$ for the nominal rolling stock schedule $\tilde{\mathcal{R}}$ and variables $(\mathcal{R}^s)_{s\in S} = (x^s, l^s, \delta^s, \delta'^s)_{s\in S}$ for each adapted schedule $\mathcal{R}^s$ to scenario s.

Objective Function
The main objective of the adaptive Rolling Stock Scheduling Problem is to minimize expected adaptation costs of the adapted solution $\mathcal{R}^s$, which corresponds to the following objective function:

$$\min_{\tilde{\mathcal{R}}, \mathcal{R}^s\, s\in S} \quad \beta \cdot \sum_{s\in S} Lack(\mathcal{R}^s) + \Gamma \cdot \sum_{s\in S} Units(\mathcal{R}^s) + \Delta \cdot \sum_{s\in S} Dead(\mathcal{R}^s) + \zeta \cdot \sum_{s\in S} Over(\mathcal{R}^s) + \eta \cdot \sum_{s\in S} Struct(\tilde{\mathcal{R}}, \mathcal{R}^s), \quad (29)$$

where the objective $Struct$ is now quadratic and can be rewritten with a simple linearization.

Controlling the Nominal Performance
We have to ensure a good performance for the rolling stock schedule $\tilde{\mathcal{R}}$. For this purpose, one may minimize the following additional objective, which corresponds to the performance criterion for the nominal rolling stock schedule:

$$\min \quad \tilde{\beta} \cdot Lack(\tilde{\mathcal{R}}) + \tilde{\Gamma} \cdot Units(\tilde{\mathcal{R}}) + \tilde{\Delta} \cdot Dead(\tilde{\mathcal{R}}) + \tilde{\zeta} \cdot Over(\tilde{\mathcal{R}}). \quad (30)$$

In practice, we have no information about the rate of time periods H without any demand changes. We just know these demand changes are quite rare. In other words, although we know the occurrence probability of a scenario s, we have no information about the probability $\tilde{p}$ of a fictive scenario without any demand changes. Thus, we prefer to introduce the nominal performance criteria in the constraints of our MILP:

$$Lack(\tilde{\mathcal{R}}) \leq (1 + \varepsilon^{Lack}) \cdot Lack^* \tag{31}$$

$$Units(\tilde{\mathcal{R}}) \leq (1 + \varepsilon^{Units}) \cdot Units^* \tag{32}$$

$$Dead(\tilde{\mathcal{R}}) \leq (1 + \varepsilon^{Dead}) \cdot Dead^* \tag{33}$$

$$Over(\tilde{\mathcal{R}}) \leq (1 + \varepsilon^{Over}) \cdot Over^*. \tag{34}$$

Parameters $Lack^*$, $Units^*$, $Dead^*$ and $Over^*$ correspond to the optimal associated performance cost for a (non adaptive) nominal schedule without any scenario. They can be computed by solving a MILP, looking for a schedule $\tilde{\mathcal{R}} \in \mathcal{F}_{\tilde{D}}$ respecting Constraints (19) while minimizing the objective (30).

Parameters ε^{Lack}, ε^{Units}, ε^{Dead} and ε^{Over} are non-negative and control the optimality gap between the adaptive nominal rolling stock $\tilde{\mathcal{R}}$ and an optimal non adaptive rolling stock schedule. The larger ε is, the more the adaptive nominal schedule is allowed to degrade the performance criteria. However, a larger value for any ε implies a larger solution space for the rolling stock schedules $\tilde{\mathcal{R}}$ and $(\mathcal{R}^s)_{s \in S}$, and thus to reduce adaptation costs of the objective function (29).

6 Computational Experiments

Description of the instances

We illustrate the relevance and efficiency of the adaptive model by computational experiments on two realistic nominal instances inspired by SNCF instances.

Table 1: Characteristics of the two instances

	Instance 1	**Instance 2**
Context	Passengers	Freight
Horizon	7 days	7 days
Trains	819	339
Stations	9	29
Number of unit types	1	3
Total number of units	23	35
Nominal demands	1	between 1 and 2
Maximal demands	2	between 1 and 4
Split/combination restrictions	no	yes

Table 1 shows the main characteristics of the two instances. The first one is derived from a set of regional trains, while the second one represents a pool of freight trains. The first instance has a lot of trains but has a simple structure, with few stations, only one unit type, homogeneous demands and no split and combinations restrictions. On the other hand, the second instance has less trains but is more complex, with a lot of stations and three unit types.

We complete each of these 2 instances with a set of 3 arbitrarily generated scenarios $S = \{s_1, s_2, s_3\}$ with equal probabilities. Each $s \in S$ has an updated demand $D_i^s = \tilde{D}_i + 1$ for about 10% of randomly chosen trains i. The demand of the other trains is unchanged.

Parameters $(\beta, \Gamma, \Delta, \zeta, \eta)$ for the objective function
From an industrial point of view, the lack of rolling stock units is the main objective to minimize. The second one is the number of engaged units, the third one concerns the dead-head trips and the fourth one the over-compositions. Thus, we can use our formulation with $\beta \gg \Gamma \gg \Delta \gg \eta$, which corresponds to a lexicographical order.

With regard to structural adaptation costs, we assume that they are more important than those of over-compositions, but less important than those of dead-head trips. Indeed, scheduling additional dead-head trips often has an impact on drivers schedules. Thus, it is not reasonable to schedule unnecessary dead-head trips to reduce the adaptation costs. Moreover, a surplus in over-compositions has no impact on drivers schedules and it seems reasonable to increase them to reduce structural adaptation costs.

The MILP formulation with these 5 parameters is not suitable for an optimization tool as typical values of the objective function are too large which may lead to floating errors. Thus, the MILPs are solved for the first objective, after which a constraint is added so that this objective does not exceed its obtained value, and the second criterion is minimized. We proceed in the same way for each of the objectives. This process also enables to understand which of the objectives are the most difficult.

Comparison with the traditional approach
We want to compare the efficiency of the *traditional approach* used at SNCF and an *adaptive* process based on the problem that we described in Section 5. The traditional approach is simulated using the two following steps:

1. We solve the MILP (5)–(19) with objective (30) and we obtain the nominal rolling stock schedule $\tilde{\mathcal{R}}^{tr} = (\tilde{x}^{tr}, \tilde{l}^{tr}, \tilde{\delta}^{tr}, \tilde{\delta}'^{tr}) \in \mathcal{F}_{\tilde{D}}$;

2. For each scenario $s \in S$, we solve a Rolling Stock Adaptation Problem and obtain mean adaptation costs to move from $\tilde{\mathcal{R}}^{tr}$ to $\tilde{\mathcal{R}}^{tr,s}$.

We test three different adaptive processes with different values for parameters ε^{Lack}, ε^{Units}, ε^{Dead} and ε^{Over}. We set ε^{Lack} and ε^{Units} to 0 to prevent any deterioration of these objectives and test three different values $\varepsilon \in \{0, 0.1, 0.25\}$ for $\varepsilon^{Dead} = \varepsilon^{Over}$.

In the following, we use the notation

$$\overline{Lack}(\mathcal{R}^S) \triangleq \frac{1}{\text{card}(S)} \cdot \sum_{s \in S} Lack(\mathcal{R}^s) \tag{35}$$

to represent the mean expected adaptation costs for objective (23) with regard to S, and we proceed in the same way for the other objectives. The objectives $Dead$ and $Over$ are expressed in kilometers.

First instance
The performance of the nominal rolling stock schedule is summarized in Table 2. As expected, when $\varepsilon = 0$, the nominal adaptive solution has exactly the same (optimal) performance as the solution in the traditional approach. The same applies when $\varepsilon = 0.1$, but when $\varepsilon = 0.25$ the numbers of dead-head trips kilometers and over-compositions kilometers are no more optimal.

This nominal near-optimality allows to reduce the mean expected adaptation costs as represented in Table 3, especially when it comes to the structural adaptation costs, passing

Table 2: Performance costs of $\tilde{\mathcal{R}}$ for Instance 1

	$Lack(\tilde{\mathcal{R}})$	$Units(\tilde{\mathcal{R}})$	$Dead(\tilde{\mathcal{R}})$	$Over(\tilde{\mathcal{R}})$
Traditional approach	0	22	228	678
Adaptive process with:				
$\varepsilon = 0$	0	22	228	678
$\varepsilon = 0.1$	0	22	228	678
$\varepsilon = 0.25$	0	22	265	833

Table 3: Mean adaptation costs to move from $\tilde{\mathcal{R}}$ to an adapted schedule for Instance 1

	$\overline{Lack}(\mathcal{R}^S)$	$\overline{Units}(\mathcal{R}^S)$	$\overline{Dead}(\mathcal{R}^S)$	$\overline{Struct}(\tilde{\mathcal{R}}, \mathcal{R}^S)$	$\overline{Over}(\mathcal{R}^S)$
Traditional approach	1.6	23	69.3	30	1985.6
Adaptive process with:					
$\varepsilon = 0$	1.6	23	69.3	5.3	4490.3
$\varepsilon = 0.1$	1.6	23	69.3	5.3	4490.3
$\varepsilon = 0.25$	1.6	23	69.3	4.6	5021

from 30 in the traditional approach to about 5 in the adaptive processes, even if $\varepsilon = 0$. This means that the nominal solution of the traditional approach has a very high mean expected structural adaptation costs which can be reduced without any deterioration of the performances. However, the mean expected number of kilometers for the over-compositions is increased by a factor of 2.5 which represents a big price to pay in term of energy consumption.

All the MILPs are solved to optimality within a few minutes, except those concerning the objectives $Struct$ and $Over$ during the adaptive processes. After an hour of computations, MILPs with objective $Struct$ have an optimality gap between 7% (for $\varepsilon = 0$) and 44% (for $\varepsilon = 0.25$), while those for $Over$ have an optimality gap of about 60%. These large gaps could explain why the objective $Over$ has very high values compared to the optimal values of the traditional approach.

Second instance

Table 4 summarizes the performance costs of the nominal rolling stock schedule. The objectives $Dead$ and $Over$ are not optimal for $\varepsilon = 0.1$ or $\varepsilon = 0.25$. They have significantly higher values than in Instance 1 as there are significantly more stations which makes it more difficult to respect the cyclicity constraints without doing dead-head trips and over-compositions. The objective $Over$ is better with $\varepsilon = 0.25$ than with $\varepsilon = 0.1$, since the objective $Dead$ is optimized before and has a larger value with $\varepsilon = 0.25$.

Table 5 summarizes the mean expected adaptation costs for the 3 scenarios. Resolution times are similar to those of Instance 1, and the optimality gaps after an hour of computa-

Table 4: Performance costs of $\tilde{\mathcal{R}}$ for Instance 2

	$Lack(\tilde{\mathcal{R}})$	$Units(\tilde{\mathcal{R}})$	$Dead(\tilde{\mathcal{R}})$	$Over(\tilde{\mathcal{R}})$
Traditional approach	0	34	1529	15226
Adaptive process with:				
$\varepsilon = 0$	0	34	1529	15226
$\varepsilon = 0.1$	0	34	1653	16616
$\varepsilon = 0.25$	0	34	1851	16383

Table 5: Mean adaptation costs to move from $\tilde{\mathcal{R}}$ to an adapted schedule for Instance 2

	$\overline{Lack}(\mathcal{R}^S)$	$\overline{Units}(\mathcal{R}^S)$	$\overline{Dead}(\mathcal{R}^S)$	$\overline{Struct}(\tilde{\mathcal{R}}, \mathcal{R}^S)$	$\overline{Over}(\mathcal{R}^S)$
Traditional approach	1.3	35	2359.3	25.6	11913.3
Adaptive process with:					
$\varepsilon = 0$	1.3	35	2067.6	11	11843.3
$\varepsilon = 0.1$	1.3	35	2017	5.3	11791
$\varepsilon = 0.25$	1.3	35	2017	5	12036

tion reach about 70% for the objective $Struct$ and 10% for $Over$ in the adaptive processes. However, except the objective $Over$ for $\varepsilon = 0.25$ with a small increase of about 100 kilometers, all the objectives have a better value in the adaptive processes. These results can be explained by the fact that Instance 2 is much more complicated than Instance 1. As a consequence, any demand change is hard to satisfy if it has not been properly anticipated which is precisely the main interest of an adaptive process.

7 Conclusion and Perspectives

In this paper, we developed a new way to model the adaptation costs in rolling stock railway scheduling. We introduced the concept of adaptive solution to reduce the adaptation costs of a rolling stock schedule. Two MILPs were proposed, the first one is solved in the adaptation phase while the second one is designed to compute adaptive solution in the conception phase. Our first results on realistic instances are promising. They show that the adaptation costs can be significantly reduced with an adaptive process while keeping good performance criteria for the nominal solution, especially for instances with complex structures.

In future research, we want to address this issue with a more sophisticated multi-objective optimization. We want to find adaptive solutions with balanced structural and performance costs. In addition, we want to improve the resolution of the MILPs with decomposition techniques.

References

Ben-Khedher, N., Kintanar, J., Queille, C., Stripling, W., 1998. "Schedule Optimization at SNCF: From Conception to Day of Departure", *Interfaces*, **28**(1), 6–23.

Ben-Tal, A., Goryashko, A., Guslitzer, E., Nemirovski, A., 2004. "Adjustable Robust Solutions of Uncertain Linear Programs", *Mathematical Programming*, **99**(2), 351.

Borndörfer, R., Grimm, B., Reuther, M., Schlechte, T., 2017. "Template-based re-optimization of rolling stock rotations", *Public Transport*, **9**(1-2), 365–383.

Borndörfer, R., Reuther, M., Schlechte, T., Waas, K., Weider, S., 2016. "Integrated Optimization of Rolling Stock Rotations for Intercity Railways", *Transportation Science*, **50**(3), 863–877.

Budai, G., Maróti, G., Dekker, R., Huisman, D., Kroon, L. G., 2010. "Rescheduling in passenger railways: the rolling stock rebalancing problem", *Journal of Scheduling*, **13**, 281–297.

Cacchiani, V., Caprara, A., Toth, P., 2010. "Solving a Real-world Train-unit Assignment Problem", *Mathematical Programming*, **124**(1-2), 207–231.

Cacchiani, V., Caprara, A., Galli, L., Kroon, L., Maróti, G., Toth, P., 2012. "Railway Rolling

Stock Planning: Robustness Against Large Disruptions", *Transportation Science*, **46**(2), 217–232.

Cicerone, S., D'Angelo, G., Di Stefano, G., Frigioni, D., Navarra, A., Schachtebeck, M., Schöbel, A., 2009. "Recoverable Robustness in Shunting and Timetabling". *Pages 28–60 of:* Ahuja, R. K., Möhring, R. H., Zaroliagis, C. D. (eds), *Robust and Online Large-Scale Optimization*. Lecture Notes in Computer Science, vol. 5868. Springer.

Fioole, P.-J., Kroon, L., Maróti, G., Schrijver, A., 2006. "A rolling stock circulation model for combining and splitting of passenger trains", *European Journal of Operational Research*, **174**, 1281–1297.

Giacco, G. L., D'Ariano, A., Pacciarelli, D., 2014. "Rolling stock rostering optimization under maintenance constraints", *Journal of Intelligent Transportation Systems: Technology, Planning, and Operations*, **18**(1), 95–105.

Liebchen, C., Lübbecke, M., Möhring, R., Stiller, S., 2009. *The Concept of Recoverable Robustness, Linear Programming Recovery, and Railway Applications*. Berlin, Heidelberg: Springer Berlin Heidelberg. Pages 1–27.

Lingaya, N., Cordeau, J.-F., Desaulniers, G., Desrosiers, J., Soumis, F., 2002. "Operational car assignment at VIA Rail Canada", *Transportation Research Part B: Methodological*, **36**(9), 755–778.

Löbel, A., 1998. "Vehicle Scheduling in Public Transit and Lagrangean Pricing", *Management Science*, **44**(12), 1637–1649.

Lusby, R. M., Haahr, J. T., Larsen, J., Pisinger, D., 2017. "A Branch-and-Price algorithm for railway rolling stock rescheduling", *Transportation Research Part B: Methodological*, **99**, 228–250.

Nielsen, L. K., Kroon, L., Gábor, M., 2012. "A rolling horizon approach for disruption management of railway rolling stock", *European Journal of Operational Research*, **220**(2), 496 – 509.

Wagenaar, J., Kroon, L., Fragkos, I., 2017. "Rolling stock rescheduling in passenger railway transportation using dead-heading trips and adjusted passenger demand", *Transportation Research Part B: Methodological*, **101**, 140–161.

A new Constraint Based Scheduling model for real-time Railway Traffic Management Problem using conditional Time-Intervals

Grégory Marlière [a,1], Sonia Sobieraj Richard[a],
Paola Pellegrini [b], Joaquin Rodriguez [a]
[a] Univ. Lille Nord de France, F-59000 Lille, IFSTTAR, COSYS, ESTAS,
F-59650 Villeneuve d'Ascq, France
[b] Univ. Lille Nord de France, F-59000 Lille, IFSTTAR, COSYS, LEOST,
F-59650 Villeneuve d'Ascq, France
[1] E-mail: gregory.marliere@ifsttar.fr, Phone: +33(0)320438496

Abstract
This paper tackles the real-time Railway Traffic Management Problem (rtRTMP). It is the problem of finding an optimal choice for the train schedules and routes to reduce the delays of trains due to conflicts. We present a new formulation of the rtRTMP. This new formulation is based on a previously proposed one that models railway traffic at a microscopic level with optional activities using a Constraint Based Scheduling (CBS) approach. To ease the modelling of optional activities, a new concept based on a tree data structure and a specific filtering algorithm was extended through the introduction of conditional time-interval variables in Ilog CP-optimizer library. The new formulation of the rtRTMP presented in this paper exploits the conditional time-interval variables. The formulation has been validated with experiments on a large set of instances. The experimental results demonstrate the effectiveness of this new CBS model and show its good performance compared with the state-of-the art RECIFE-MILP algorithm.

Keywords
Real Time Traffic Management, Train Dispatching Problem, Re-routing and re-scheduling trains, Minimize secondary delays, Constraint Propagation

1 Introduction

The design of railway services is a complex process in which the planning of the schedule of trains and the necessary resources can lead to conflicts at the operational level. These conflicts are due to unforeseen perturbation events. The main consequence of these conflicts is the delays suffered by trains and, consequently, the increase of passenger travel time. Delays due to conflicts between two trains are called "secondary" delays. Railway operators try to limit secondary delays inserting time allowances in the timetable design phase. Nevertheless, time allowance is not always sufficient to avoid conflicts or even their propagation to other trains in a snowball (or domino) effect. To limit this propagation, the dispatcher in charge of traffic management can change the dwell times at scheduled stops, the train orders at stations or junctions, or the routes assignment. The problem of finding an optimal choice for the train schedules and routes is defined as the real-time Railway Traffic

Management Problem (rtRTMP) (Pellegrini et al., 2014). A rich literature exists on formulations and methods for solving the rtRTMP, the reader is referred to Lusby et al. (2011), Cacchiani et al. (2014), Fang et al. (2015) for recent literature surveys.

The surveys point out that integer programming (IP) and mixed-integer programming (MIP) models are the most popular approaches along with graph models, while constraint programming (CP) ones are more seldom used. Nevertheless, CP models have some undeniable merits which make them interesting for this problem. In particular, they are able to generate feasible solutions for some hard problems in a short computation time. As an example, to generate the cyclic timetables of the Dutch network (Kroon et al., 2009), the method of the CADANS module to solve the Periodic Event Scheduling Problem (PESP) formulation is based on CP techniques (Schrijver and Steenbeek, 1994). We can also mention that the PESP instances of the whole inter city nertwork of Germany and the south and east subnetworks have been solved with a SAT-solver (Großmann et al., 2012), which uses specific CP techniques for variables with boolean domains.

For a given problem instance, CP models typically have fewer variables and constraints than the other approaches, and therefore requires less memory for the instances formulation. It is also worthwhile mentioning that despite the diversity of models and solutions methods, very few publications compare and analyse their relative performances and advantages.

Since our first proposal of a CP model in (Rodriguez, 2007), new features of CP and Constraint Based Scheduling (CBS) have been developed. CBS extends CP to get stronger propagation algorithms for specific constraints to solve scheduling problems. One feature is the ability to model optional activities along with powerful propagation algorithms (Vilím et al., 2005). In addition, exact algorithms that use hybrid methods (i.e. CP and Linear programming) and provide optimality proofs have been developped (Laborie and Rogerie, 2016). In this research, we aim to deeply investigate some CP and CBS modelling possibilities in the light of the new features developed in the last decade and we initiate a comparison of the achievable performance with the ones of other algorithms.

To do so, in this paper, we present a new CBS formulation of the rtRTMP that has been validated with experiments on a large set of instances. The performances of the heuristic resolution method for this new CBS formulation has been compared with the one of the state-of-the art RECIFE-MILP heuristic (Pellegrini et al., 2014).

2 CBS formulation

2.1 Scheduling theory

The basic idea of the CBS model of the rtRTMP is that a train passing through a control area is a job. According to scheduling theory, the concept of job is a set of activities linked by a set of temporal constraints. The rtRTMP can be viewed as a joint problem of allocating resources (the infrastructure broken down into track sections) to some activities sequences (the movement of a train).

In a CBS formulation, temporal constraints connect the temporal variables concerning activities (e.g., start, end or duration variables) according to principles which are specific to each application. The resource constraints are linked to the use and sharing of the resources by activities. Resources are divided into consumable or renewable resources, with the latter being either of limited capacity or with limited states. By sharing resources, indirect links between the temporal activity variables are generated by capacity or state resource

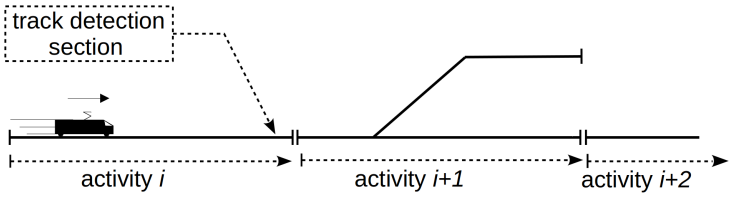


Figure 1: Train movement as a sequence of activities.

constraints.

This modelling approach for train scheduling was first proposed by Spzigel (1973) and then formulates the train scheduling problem on a single track line as a job-shop scheduling problem. Trains are jobs and their traveling through the single track connecting consecutive stations are activities.

In the remaining part of this paper, the formulation of the rtRTMP by Rodriguez (2007) is named RECIFE-CP1 and the new formulation presented in this paper is named RECIFE-CP2. We will refer to RECIFE-CP when we consider common parts to both formulations.

2.2 Microscroscopic model of the rtRTMP

The overall approach named RECIFE-CP is based on a microscroscopic model of the rtRTMP where train movements are controlled with a fixed block signalling system. The first modeling principle characterizing it considers a detailed decomposition of a train journey into a sequence of activities. Each activity is an elementary movement of the train through a track detection section (tds), as illustrated in Figure 1. A track detection section allows the detection of the occupation of a part of the railway infrastructure by a train. Tds's correspond in many railway infrastructures to electric devices named "track circuits" and are part of the block signalling system that ensure the safe movements of trains.

During normal operation, most of the time only one train should be detected by a tds at any point in time. Hence, tds's are modelled as unary resources. A unary resource is a resource allowing only one activity to use it at any point in time. However, an exception occurrs if a train set is splitted to operate two trains or, conversely, two train sets are joined to operate one train. This exception must be taken into account for the tds's corresponding to station platforms where split and join operations are performed.

2.3 Temporal constraints

A second modeling principle of RECIFE-CP consists in the consideration of detailed temporal constraints between activities. They allow modeling some characteristics of the block signalling system[1] such as: the length of trains, the number of signalling aspects, the watching time (e.g. running time of the sight distance), the sectional route release of the interlocking system.

A brief overview of the temporal constraints between activities is illustrated by a time over distance diagram in Figure 2. Along the horizontal axis of this diagram, we have the

[1] Additional characteristics are ommitted , e.g., time for clearing signal or release time, to simplify the presentation

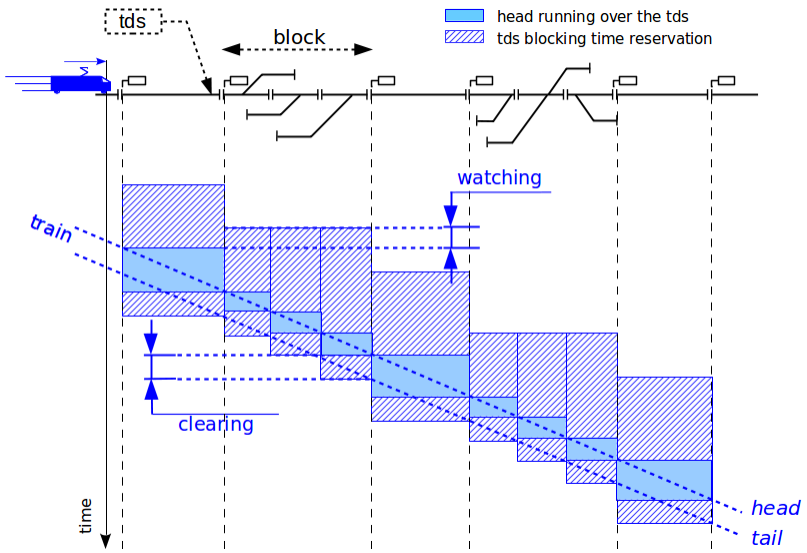


Figure 2: Head running activity and tds blocking time reservation

sequence of tds's that the "blue" train runs through. The line is broken down into blocks that are bounded by signals providing driving information to the train driver. A block can have one or more tds's depending on the configuration of the line. In the diagram, blue dashed lines report the position of the head and the tail of the train.

In RECIFE-CP1, we consider for each tds only one activity. The temporal constraints to maintain safe headway with a preceeding train (i.e., the blocking time theory constraints (Hansen, 2008)) are expressed according the start and end of these activities (Rodriguez, 2007). Instead, in RECIFE-CP2 we define two nested activities for each tds. The first one is the running of the head of the train through the tds. The sequence of the head running activities are shown with filled blue rectangles in Figure 2. Each activity is linked by a "start at end" constraint with the precedent activity. The second activity associated to a tds includes the first one and is extended to contain the reservation time to comply with the blocking time theory constraints. This second type of activities are shown with striped rectangles in Figure 2 for the case of 3-aspect block signalling sytem. All tds's of a block must be reserved when the train reaches the watching distance point of the previous block. The additional detection time due to the length of the train (called clearing time) is shown by the extended striped rectangle. It lasts until the tail of the train is no more detected by the tds. When there are switches within a block, separated tds's for each switch allow the interlocking system to release and set earlier the sequence of incompatible routes and then safely optimise the traffic. The sectional route release of the interlocking system is modelled in RECIFE-CP with separed activities for each tds of a block section (c.f. Figure 2).

Mascis and Pacciarelli (2002) showed that these temporal constraints have the same properties as the ones of a job-shop scheduling problem with blocking and no-wait constraints of the classic scheduling theory.

2.4 Alternative route choices

A third modeling principle of RECIFE-CP consists in considering alternative routes. Therefore, decision variables are defined to select one from the set of alternative routes for each train to avoid conflicts or to reduce secondary delays. The ability to model optional activities (Vilím et al., 2005; Laborie and Rogerie, 2008) has substantially changed the formulation of the model for the re-routing decisions.

In RECIFE-CP1 a train run is modelled with only *one sequence of activities* whatever the chosen route. Hence, each activity does not necessarily correspond to a real train movement through a tds as not all routes have the same tds sequence length. On the other hand, in RECIFE-CP2 a train run is modelled with *as many sequences of activities as route choices*. The sequence of head running activities length is equal to the tds sequence length of the chosen route. It should be noted that each blocking time reservation activity covers a head running activity, thus there is a sequence of blocking time reservation activities "enveloping" the sequence of head running activities. To illustrate both types of model, let us first consider the example of train that has two alternative routes r_1 and r_2 in Figure 3.

In RECIFE-CP1, a single sequence of six activities is defined, Figure 4 gives the different tds's that can be used by each activity according to the route choices r_1 and r_2. If r_2 is chosen, a tds is assigned to each activity which models the time to block and to run through it. If r_1 is chosen, a "dummy" tds, named tds^*, is added in the sequence of r_1 for an additional "fictive" activity as r_1 have only five tds's. tds^* can be added at any position within the sequence. In the example of Figure 4, it is added in the fourth position, therefore a_4 is the additional fictive activity. In a similar way, more than one tds^* can be added and put at any position within the sequence. The duration of the fictive activities cannot be zero due to the temporal constraints that link the sequence of activities: they are both equal to the clearing time of the previous activity. Therefore, tds^* can be viewed as a tds with zero length. Many activities can require tds^* as fictive elementary runs at any point in time, then the resource tds^* has an infinite capacity to satisfy all these capacity requirements. Adding tds^* with the former properties allows the definition of the same kind of resource and temporal constraints to all the activities of the sequence.

In RECIFE-CP2, we have two sequences of activities for this example. A sequence with five activities for r_1 and a sequence with six activities for r_2. To reduce the number of variables and improve the constraint propagation algorithm, the activities of two routes that have the same tds sequence with same running times are merged. After merging the equivalent activities, we obtain a graph of activities such that a path from the first tds activity to the last tds activity gives a sequence of activities for r_1 and a different sequence activities for r_2. In the example, the activities for the elementary run through tds_1 and tds_7 are merged as they have the same characteristics for r_1 and r_2. Conversely, for the elementary runs through tds_2 and tds_6, two activities are kept separated because the minimal running time for r_1 is different from the one for r_2. If r_1 is chosen, the activities a^{r_2,tds_2}, a^{r_2,tds_3}, a^{r_2,tds_5}, a^{r_2,tds_6} are "non-executed" and all related constraints and variables are useless. Similarly if r_1 is chosen, a^{r_1,tds_2}, a^{r_1,tds_4}, a^{r_1,tds_6} are non-executed. When all route assignments are done, we have to get only one sequence of executed activities for each train coherent with the route choice.

To improve the constraint propagation and hence the solution method, we create in RECIFE-CP2 a hierarchical model with new global constraints on groups of activities. These global constraints allow the encapsulation of a group of activities into one high-level

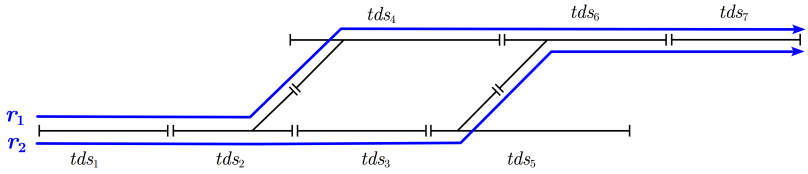


Figure 3: Example of tds route sequences

RECIFE-CP1 model		
sequence of activities	$a_1 \longrightarrow a_2 \longrightarrow a_3 \longrightarrow a_4 \longrightarrow a_5 \longrightarrow a_6$	
tds sequences	r_1	$(tds_1, tds_2, tds_4, tds^*, tds_6, tds_7)$
	r_2	$(tds_1, tds_2, tds_3, tds_5, tds_6, tds_7)$
RECIFE-CP2 model		
tds sequence	r_1	$(tds_1, tds_2, tds_4, tds_6, tds_7)$
graph of head running activities		$a^{r_1,r_2}_{tds_1} \rightarrow a^{r_1}_{tds_2} \rightarrow a^{r_1}_{tds_4} \rightarrow a^{r_1}_{tds_6} \rightarrow a^{r_1,r_2}_{tds_7}$; $a^{r_1,r_2}_{tds_1} \rightarrow a^{r_2}_{tds_2} \rightarrow a^{r_2}_{tds_3} \rightarrow a^{r_2}_{tds_5} \rightarrow a^{r_2}_{tds_6} \rightarrow a^{r_1,r_2}_{tds_7}$
tds sequence	r_2	$(tds_1, tds_2, tds_3 tds_5, tds_6, tds_7)$

Figure 4: Sequences of activities and tds's for the two RECIFE-CP models

activity. Derived high-level activities can be used with any temporal constraint in the same way as low-level ones.

In the example of Figure 3, we can notice that each activity of the group $G_1 = \{a^{r_1}_{tds_2}, a^{r_2}_{tds_2}\}$ always starts after the end of activity $a^{r_1,r_2}_{tds_1}$. In the same way, activity $a^{r_1,r_2}_{tds_7}$ always starts after the end of each of activity of the group $G_2 = \{a^{r_1}_{tds_6}, a^{r_2}_{tds_6}\}$. Let a_{G_1} (resp. a_{G_2}) the high-level activity linked by a "group constraint" to the group G_1 (resp. G_2), then we can state the precedence constraints $a^{r_1,r_2}_{tds_1} \prec a_{G_1}$ and $a_{G_2} \prec a^{r_1,r_2}_{tds_7}$.

More generally, we define a precedence constraint between a pair of high-level activities $(a_{G_{prec}}, a_{G_{succ}})$ such that each high-level activity is linked by a "group constraint" to a group of activities G_{prec} and G_{succ} respectively: each activity of G_{prec} precedes an activity of G_{succ} and conversely each activity of G_{succ} follows an activity of G_{prec}.

Two group constraints are used: `span`$(a_G, a_1, \ldots, a_n)$ states that activity a_G, if executed, spans over all executed activities of the set $\{a_1, \ldots, a_n\}$; `alternative`$(a_G, a_1, \ldots, a_n)$ states that if activity a_G is executed then exactly only one of activities $\{a_1, \ldots, a_n\}$ is executed and a_G starts and ends together with this chosen one. Activity a_G is non-executed if

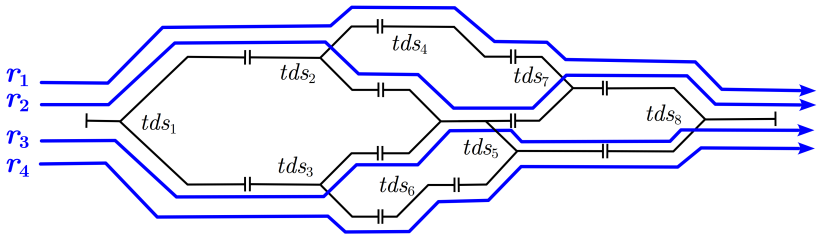


Figure 5: Example of tds sequences to illustrate group constraints

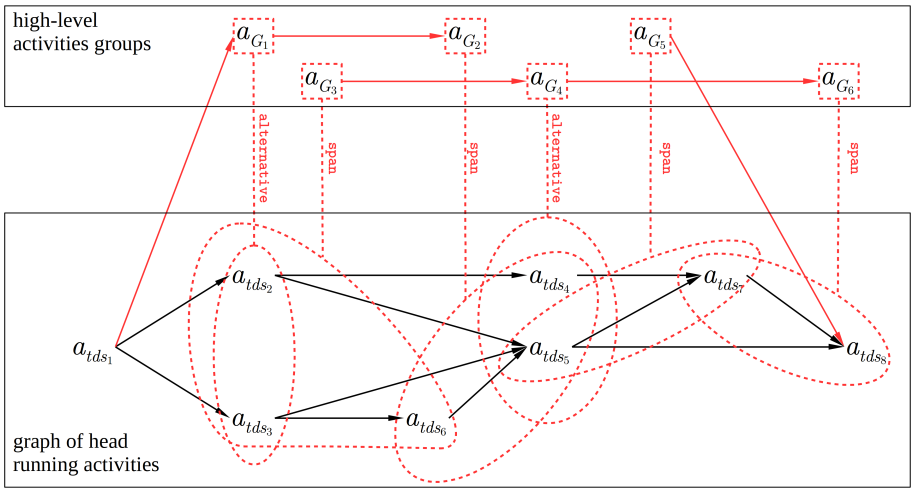


Figure 6: Graph of activities for the example of Figure 5

and only if none of activities $\{a_1, \ldots, a_n\}$ is executed (Laborie and Rogerie, 2016).

To illustrate the definition of high-level activities and the group constraints, let us consider another example, depicted in Figure 5. To simplify the presentation, we consider that the elementary runs through a tds have the same characteristics whatever the route considered. Therefore all the activities corresponding to runs through a common tds are merged into one activity which is not indexed by routes. The lower part of the Figure 6 shows the graph of head running activities. Remark that contrary to the example of Figure 3 not all paths of the precedence graph correspond to a tds sequence activities for a route. The activities of each group are shown with red dotted shapes linked with a red dotted line to the corresponding group activity. The links are named with the group constraint used. The first hierarchical activity a_{G_1} is linked by an `alternative` constraint to the set of activities $\{a_{tds_2}, a_{tds_3}\}$ to state the precedence constraint $a_{tds_1} \prec a_{G_1}$. The group $G_2 = \{a_{tds_4}, a_{tds_5}, a_{tds_6}\}$ is an example in which the activity a_{G_2} cannot be linked with an `alternative` constraint because a_{tds_6} precedes a_{tds_5} thus the activities of G_2 are linked

with a **span** constraint to a_{G_2}. This **span** constraint with a_{G_2} allows to state the precedence constraint $a_{G_1} \prec a_{G_2}$. All in all, the group activities of this example add six high-level activities and five precedence constraints (red arrows).

2.5 Conditional time-interval variables

In many works of scheduling theory, the main decisions are assigning resources to activities and scheduling activities. However in industrial applications, it is also necessary to consider the choice of activities that will be executed in the final schedule, for example when there are alternative production processes in response to an order. As in Artificial Intelligence Planning which requires to choose a sequence of actions to achieve a goal, recent developements in scheduling consider problems involving the choice of whether to execute or not some activities. This translates into the introduction of *optional activities*.

Vilím et al. (2005) introduce a tree data structure and a specific constraint propagation algorithm to model optional activities. This was later extended by the introduction of *conditional time-interval variables* in Ilog CP-optimizer library (Laborie and Rogerie, 2008).

A conditional time-interval variable (or time-interval variable for the sake of simplicity), noted a, represents an interval of time of interest in a schedule. In many cases, as in the problem modelled here, a time-interval variable is the time interval in which an activity is executed.

Let $\perp$ a value meaning the interval of time of interest is not present in the solution schedule or an activity is non-executed. The domain of a time-interval variable is a subset of $\{\perp\} \cup \{[s, e) | s, e \in \mathbb{Z}, s \leqslant e\}$. Like any other variable in a constraint satisfaction problem, a time-interval variable is said to be fixed if its domain is reduced to a singleton. Let $\underline{a}$ denotes a fixed time-interval variable, then $\underline{a} = \perp$ means that the activity is non-executed (not present in the solution schedule) or $\underline{a} = [s, e)$ means that the activity is executed (present in the solution schedule). The values s and e are respectively the start and end time of the activity. A time-interval variable is said to be non-executed if it is not considered by any constraint or expression it is involved in, said in a different way, it is as if they were deleted. An execution or presence status noted $pres(\underline{a})$ is equal to 1 if the activity is executed and 0 if it is non-executed.

When a is linked by a precedence constraint to another time-interval variable b and a is non-executed, then the precedence constraint impacts b. More generally all constraint definitions (i.e. propagation algorithms) must specify how they manage non-executed time-interval variables.

The conditional time-interval variables are linked by two kinds of constraints : the logical constraints and the temporal constraints.

The logical constraints link the execution status of the time-interval variables. These constraints are aggregated in a 2-SAT (2-satisfiability) constraint network. For example, the execution status of the time-interval variables for two alternative tds's that correspond to two route choices will be linked by a clause with an $\vee$ operator.

The temporal constraints state the different temporal positions of the start and end events of the time-interval variables, i.e., "start before start" or "start at end". These constraints are aggregated in a Simple Temporal Network (STN) extended to the presence statuses. The temporal constraints are "hybrid" in the sense that they combine the logical aspect of activities (i.e. "executed" or "non-executed") and the temporal aspect (i.e., it represents an activity with a start, end and duration).

Beside the expressiveness of the time-interval variables, the 2-SAT and STN constraint networks ensure a strong constraint propagation and therefore an efficient search for the optimization engine.

3 Formulation

For the formulation, we use a notation close the one introduced by Pellegrini et al. (2014) and then as follows:

T, R, TDS	set of trains, routes and track detection sections, respectively,
$R_t \subseteq R$	set of routes that can be used by train t,
TDS^r	set of track detection sections composing route r,
ty_t	type corresponding to train t (indicating characteristics as weight, length, engine power, etc.),
$TDS_t \subseteq TDS$	set of track detection sections that can be used by train t ($TDS_t = \bigcup_{r \in R_t} TDS^r$),
$PL \subset TDS$	set of track detection sections corresponding to platforms (if the control area includes a station),
$PL_{t,t'} \subset PL$	set of track detection sections corresponding to the possible departure platforms of a train t' which uses the same rolling stock as train t and results from the turnaround of train t,
$bs_{r,tds}$	block section including track detection section tds along route r,
$p_{r,tds}$	track detection sections preceding tds along route r,
$ref_{r,tds}$	reference track detection section for the blocking time reservation of tds along route r: first track detection section of the $n - 2^{nd}$ block section preceding $bs_{r,tds}$, with n number of aspects characterizing the signaling system,
$rt_{ty,r,tds}$	running time of track detection section tds along route r for a train of type ty,
$ct_{ty,r,tds}$	clearing time of track detection section tds along route r for a train of type ty,
for_{bs}, rel_{bs}	formation and release time for block section bs, respectively,
$init_t$	earliest time at which train t can be operated: either expected arrival in the control area or expected departure from a platform within the control area,
$exit_t$	earliest time at which train t can reach its destination given $init_t$, the route assigned to t in the timetable and the intermediate stops,
$i(t, t')$	indicator function: 1 if trains t and t' use the same rolling stock and t' results from the turnaround of train t, 0 otherwise,
$ms_{t,t'}$	minimum separation between the arrival of a train t and the departure of another train t' using the same rolling stock,
$S_t, TDS_{t,s}$	set of stations where train t has a scheduled stop and set of track detection sections that can be used by t for stopping at station s,
$arr_{t,s}$	scheduled arrival times for train t at station s.

3.1 Decision variables

We define following decision time-interval variables :
for all triplets of $t \in T$, $r \in R_t$ and $tds \in TDS^r$:

$a^{t,r}_{tds,h}$: optional time-interval variable which represents the running time activity of t's head through tds along r,

$a^{t,r}_{tds,b}$: optional time-interval variable which represents the blocking time reservation activity of tds for t along r,

for all $t \in T$:

D^{arr}_t, D^{exit}_t : delay suffered by train t at station arrivals (cumulated) and at the exit from the control area.

Moreover, we define binary variables for the route choices:
for all pairs of train $t \in T$ and route $r \in R_t$:

$$x_{t,r} = \begin{cases} 1 & \text{if } t \text{ uses } r, \\ 0 & \text{otherwise, not;} \end{cases}$$

The objective is the minimization of the total secondary delays suffered by trains at their departure from stations and exit from the control area:

$$\min \sum_{t \in T} (D^{arr}_t + D^{exit}_t) \tag{1}$$

To define the constraints, let us consider the following additional notations :

$s(a), e(a), d(a), pres(a)$ the start, end, duration and presence status for time-interval variable a, respectively,

$first(a^{t,r}_{tds,h}), last(a^{t,r}_{tds,h})$ boolean functions that return true if $a^{t,r}_{tds,h}$ is the first, respectively the last, head running activity of train t through the tds sequence for route r,

$\{(G^t_{prec}, G^t_{succ})\}$ set of pairs of groups of tds of train t $G^t_{prec} \in \mathcal{P}(TDS_t)$[2], $G^t_{succ} \in \mathcal{P}(TDS_t)$ with the folllowing property : each head running activity through a $tds \in G^t_{prec}$ (resp. $tds \in G^t_{succ}$) precedes (resp. follows) at least one head running activity through a $tds \in G^t_{succ}$ (resp. $tds \in G^t_{prec}$),

$prec(G)$ boolean function that returns true if $\exists (tds, tds') \in G$ such that the head running activities through tds and tds' are linked with a precedence constraint, otherwise false is returned.

The constraints are :

$$\sum_{r \in R_t} x_{t,r} = 1 \, \forall t \in T, \tag{2}$$

$$if (x_{t,r} = 1) \Rightarrow pres(a^{t,r}_{tds,h}) = 1 \forall t \in T, r \in R_t, tds \in TDS^r, \tag{3}$$

$$if (x_{t,r} = 1) \Rightarrow pres(a^{t,r}_{tds,b}) = 1 \forall t \in T, r \in R_t, tds \in TDS^r, \tag{4}$$

$$s(a^{t,r}_{tds,h}) \geqslant init_t \forall t \in T, r \in R_t, tds \in TDS^r, \tag{5}$$

[2]We use the notation $\mathcal{P}(S)$ to denote the power set of a set S.

$$d(a^{t,r}_{tds,h}) \geqslant rt_{ty,r,tds} \forall t \in T, r \in R_t, tds \in TDS^r, \tag{6}$$

$$s(a^{t,r}_{tds,h}) = e(a^{t,r}_{p_{r,tds},h}) \forall t \in T, r \in R_t, tds \in TDS^r, \tag{7}$$

$$e(a^{t,r}_{tds,b}) = e(a^{t,r}_{tds,h}) + ct_{ty,r,tds} + rel_{bs_{r,tds}} \forall t \in T, r \in R_t, tds \in TDS^r, \tag{8}$$

$$s(a^{t,r}_{tds,b}) = a^{t,r}_{ref_{r,tds},h} - for_{bs_{r,ref_{r,tds}}} \forall t \in T, r \in R_t, tds \in TDS^r, \tag{9}$$

$$\begin{aligned} &\texttt{alternative}(a^t_G, a^{t,r_1}_{tds_1,h}, \ldots, a^{t,r_n}_{tds_n,h}), G = \{a^{t,r_1}_{tds_1,h}, \ldots, a^{t,r_n}_{tds_n,h}\} \\ &\forall t \in T, G \in (G^t_{prec} \cup G^t_{succ}) : \neg prec(G) \end{aligned} \tag{10}$$

$$\begin{aligned} &\texttt{span}(a^t_G, a^{t,r_1}_{tds_1,h}, \ldots, a^{t,r_n}_{tds_n,h}), G = \{a^{t,r_1}_{tds_1,h}, \ldots, a^{t,r_n}_{tds_n,h}\} \\ &\forall t \in T, G \in (G^t_{prec} \cup G^t_{succ}) : prec(G) \end{aligned} \tag{11}$$

$$\begin{aligned} e(a^t_G) &= s(a^t_{G'}) \\ \forall t \in T, (G, G') &\in \{(G^t_{prec}, G^t_{succ})\} \end{aligned} \tag{12}$$

$$\begin{aligned} &pres(a^{t',r'}_{tds,h}) = pres(a^{t,r}_{tds,h}) \\ &\forall t, t' \in T, r \in R_t, r' \in R_{t'} : i(t,t') = 1 \wedge tds \in PL_{t,t'} \end{aligned} \tag{13}$$

$$\begin{aligned} s(a^{t',r'}_{tds,h}) \geqslant e(a^{t,r}_{tds,h}) + ms_{t,t'} &\forall t, t' \in T, r \in R_t, r' \in R_{t'} : \\ i(t,t') = 1 \wedge last(a^{t,r}_{tds,h}) \wedge first(a^{t',r'}_{tds,h}) &\wedge tds \in PL_{t,t'} \end{aligned} \tag{14}$$

$$\begin{aligned} s(a^{t',r'}_{tds,b}) = e(a^{t,r}_{tds,b}) &\forall t, t' \in T, r \in R_t, r' \in R_{t'} : \\ i(t,t') = 1 \wedge last(a^{t,r}_{tds,h}) \wedge first(a^{t',r'}_{tds,h}) &\wedge tds \in PL_{t,t'} \end{aligned} \tag{15}$$

$$\texttt{noOverlap}(a^{t,r}_{tds,b}) \forall t \in T, r \in R_t : tds \in TDS \tag{16}$$

$$D^{exit}_t = \sum_{\substack{r \in R_t, tds \in TDS_r: \\ last(a^{t,r}_{tds,h})}} e(a^{t,r}_{tds,h}) - exit_t \forall t \in T \tag{17}$$

$$D^{arr}_t = \sum_{r \in R_t} \sum_{\substack{s \in S_t, \\ tds \in TDS_{t,s}}} (s(a^{t,r}_{tds,h}) + rt_{ty,r,tds} - dep_{t,s} x_{t,r}) \forall t \in T \tag{18}$$

Constraints (2) ensure that exactly one route is used by each train.
Constraints (3) and (4) link the choice of a route r and the presence of the corresponding activities, i.e., if route r is chosen all the activities must be executed (be present in the solution schedule) .
Constraints (5) state that trains cannot be operated earlier than $init_t$.
Constraints (6) impose that the duration of the running time head activities are greater than the running time of track detection section tds along route r for a train of type ty .
Constraints (7) impose a precedence constraints between running time head activities of a train.
For constraints (8), the blocking time reservation lasts after the tail of the train clears tds, which corresponds to the start of the head running plus a clearing time for the type of train ty plus the block section release time.
Constraints (9) state that the blocking time reservation activity is synchronized with the time the head of the train is detected by the reference track detection section $ref_{r,tds}$ minus the route formation time.
Constraints (10) and (11) link a group of activities $G = \{a^{t,r_1}_{tds_1,h}, \ldots, a^{t,r_n}_{tds_n,h}\}$ into a high-

level activity a_G^t according to the presence of precedence constraints between low-level activities. High-level activities are linked to low-level activities by `span` or `alternative` constraints.
Constraints (12) state the precedence constraints between high-level activities.
Constraints (13) ensure local coherence: trains using the same rolling stock must use the same platform where they turnaround.
Constraints (14) ensure that a minimum separation time must separate the arrival and departure of trains using the same rolling stock for a turnaround.
Constraints (15) ensure the tds where the turnaround takes place is utilized for the whole time between t''s arrival and t's departure. Thus, the first activity blocking time reservation of t' starts when the last activity blocking time reservation of t ends.
Constraints (16) ensure that the blocking time activities of a shared tds do not overlap.
Constaints (17) and (18) state that the values of the delays D_t^{exit} and D_t^{arr} of a train t is the difference between the actual and the scheduled times at the exit of the infrastructure, respectively at the arrival at stop stations.

4 Solution method

The solution method uses the algorithm of Vilím et al. (2015) for scheduling problems which combines a Failure-Directed Search (FDS) with Self-Adapting Large Neighborhood Search (SA-LNS).

First, SA-LNS (Laborie and Godard, 2007) aims to find a good quality solution quickly. It is an iterative improvement method with following steps:

1. Start with an existing solution (heuristic or CP search)
2. Select a Large Neighborhood (LN) and a Completion Strategy (CS)
3. Apply LN to relax part of the solution and fix the rest
4. Apply CS to improve solution using a limited search tree
5. If time limit is reached then stop else go to 2

SA-LNS uses the following components to improve the search:

- *Constraint propagation algorithms* for the logical and the precedence constraints networks (Vilím et al., 2005),
- Enhanced selection of LN and CS: *machine learning techniques* to portfolios of LN and CS that quickly converge on good solutions (Laborie and Godard, 2007),
- *Temporal Linear Relaxation*: use CPLEX's LP solver for a solution to a relaxed version of the problem to guide heuristics (Laborie and Rogerie, 2016).

FDS is activated when the search space seems to be small enough, and SA-LNS has difficulties improving the current solution. It builds a complete search tree and it drives the search into conflicts in order to prove that the current branch is infeasible. It uses a *restart scheme with nogoods*.

		Junction	Line	Terminal stations	
		# 1 Gonesse	*# 2 MLJ-Rouen*	*# 3 Lille*	*# 4 StLazare*
Infrastructure	*Length (km)*	15	80	7	4.5
	Routes	37	187	2409	84
	Blocks	79	157	829	197
	Track Circuits	89	236	299	212
	Stations	0	13	1	4
	Platforms	0	33	17	51
Timetable	*Trains/Day*	336	237	589	1212
	Routes alternatives/Train	5-13	1-24	1-71	1-9
	Turnarounds	0	6	298	606

Table 1: Case-studies characteristics

5 Experiments

5.1 Case-studies

In the experimental analysis, we tested our formulation on perturbations of real instances representing four French control areas with different characteristics: a junction with mixed traffic, a line with intermediate stops, and two passenger terminal stations with high density traffic. Namely, they cover the Gonesse junction north of Paris (# 1), the line between Mante-La-Jolie and Rouen-Rive-Droite (# 2) and the Lille-Flandres (# 3) and Paris–Saint-Lazare (# 4) stations. Their characteristics are detailed in Table 1 and their layout in Appendix A. Notes that the second line of Table 1 gives the values of parameter R and the heighth line gives the bounds of the number of routes per train (bounds of $|R_t|$).

5.2 Experiments settings

The experiments involve RECIFE-CP2 (named CP) and RECIFE-MILP (named MILP) in order to compare their performances in various cases.
Both algorithms are configured to perform a two-step approach:

- in the first step, a maximum of 10 seconds CPU time is allocated for "fixed-route" solution, which means that the route fixed in the timetable is used for each train,
- in the second step, the best solution of the previous step is used for initializing the "all-route" resolution, which means that all possible routes are used.

A limit of 180s CPU time is imposed for the resolution of these two steps on an Intel(R) Xeon(R) CPU E5-2643 v4 @ 3.40GHz, 24 cores, 128go RAM.
For each control area, we used two methods to increase incrementally instance difficulties:

- Horizon size variation,
- Perturbation rate variation.

5.3 Horizon size variation

For each of the 4 control areas, we generates 30 disruption scenarios: starting from the original timetable, 20% of randomly selected trains are delayed with a value in the interval between 5 and 15 minutes.

To cover a variety of instances difficulty, for each disruption scenario, we have selected 12 morning time intervals starting at 8 am with duration from 10 minutes to 120 minutes with 10 minute step.

In this experimental set, 1440 problem instances are solved by each algorithm.

5.4 Perturbation rate variation

For each of the 4 control areas, one hour horizon scenario starting at 8 am is considered. Starting from the original timetable, the rate of randomly selected delayed trains for assigning the 5 to 15 minutes delay varies from 10% to 60% . The percentage is increased at a 10% step. We generates 30 scenarios for each of the perturbation percentage value.

In this experimental set, 720 problem instances are solved by each algorithm.

6 Results

On Figures 7 and 8, we introduce three types of graphs in order to explain the results, separately for each case study and as a function of horizon size and perturbation rate:

(a) in the first column, the curves indicate the mean frequency at which an algorithm returns the best solution among those found by both CP and MILP. The green squares (respectively the red circles) presents the mean performance of CP (respectively MILP). For example, in Figure 7, for case study #1, CP and MILP provide 100 % of the best solutions for 10 minutes horizon instance set: they always return solutions with the same values. In the same figure, CP, respectively MILP, provides 70 %, respectively 53 %, of the best solutions for the 120 minutes horizon instance set.

(b) in the second column, reports the boxplots show the distribution of the differences of objective values between the two algorithms. Observing these distributions, we can better undestand the performance results of (a) curves. For example, in Figure 7, for case study #2, in the curve (a), CP, respectively MILP, provides 80 %, respectively 100 %, of the best solutions for the 70 minutes horizon instance set. However, for this instance set, (b) boxplots indicate that the objective values of both algorithms are very close as the median of the differences is close to zero. The y axis reports the difference between the best objective value given by MILP minus the one given by CP. This means that the points above the origin are those for which CP provides better solutions than MILP,

(c) in the third column, the curves quantify the frequency at which an algorithm is able to prove the optimality of the returned solution during the allowed CPU time.

6.1 Horizon size variation

Figure 7 allows a comparison of the solution quality given by CP and MILP on each case study for the horizon size variation. The x-axis represents the horizon size in minutes, which

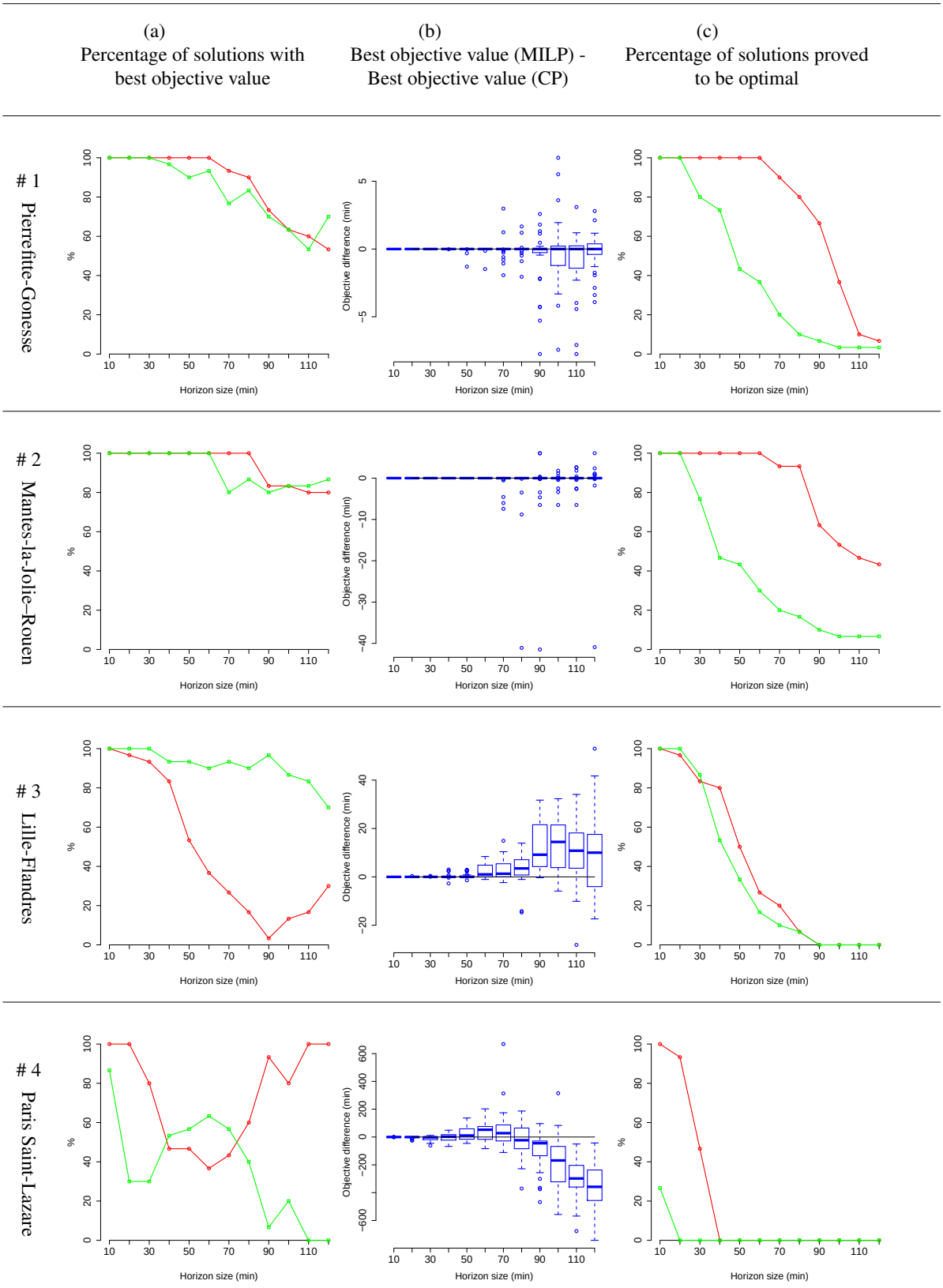


Figure 7: Experimental results for horizon size variation. In column 1 and 3, red circles for MILP, green squares for CP. In column 2, difference MILP minus CP.

is incrementally increased for this set of experiments.
These results show some general tendencies, common for all case studies:

- MILP has a better ability to prove optimality than CP, it outperforms CP for almost all horizon size problems according to this indicator, shown in column 3.. The # 1, # 2 # 3, # 4 case study instances are of increasing difficulty from this point of view: the success frequency of MILP decreases sharply after 90, 40, 30 and 20 minutes horizon size respectively,
- Regarding the best objective values indicator shown in column 1, CP and MILP have generally the same performance for small horizon sizes. As the horizon size increases, CP outperforms MILP starting from the following sizes:
  - around 110 minutes in # 1,
  - around 100 minutes in # 2,
  - before 10 minutes in # 3, note that after 90 minutes, the frequency of MILP increase,
  - around 40 minutes in # 4, note that after 60 minutes, the frequency of MILP increase and another crossing that reverses the performances order of the algorithms occurs around 70 min.

Further analysis shows that, for difficult instances of case studies (#3 and #4), MILP is not able to provide a solution during the all-routes solution phase. Therefore, the solution initially found during the fixed-route search phase is returned. Conversely, in these instances, CP provides poor solutions during the fixed-route search phase and is able to improve them during the all-routes solution phase. However, the improvement is not large enough to overtake the quality of the MILP ones.

6.2 Perturbation rate variation

The results reported in Figure 8 consider another difficulty parameter, namely the rate of perturbations with a fixed one hour size horizon. The x-axis of the plotted curves is the percentage of delayed trains: six configurations of perturbed scenarios are reported with 10 % to 60 % of delayed trains.

The results show different trends according to the case studies:

- For the instances of case studies # 1 and # 2, the best objective curves are close and does not show an important variation as for the case of horizon size variation. Regarding the optimality proof frequencies, MILP have reaches values above 80% , whatever the level of perturbations. This is not the case for CP whose frequencies decrease sharply after the first increase of perturbation level.
- For the instances of case studies # 3 and # 4, the one hour horizon scenarios considered are difficult to solve to the optimum, whatever the level of perturbation. The optimality proof rate of MILP for the instances of case study # 3 decreases under 10 % after 30 % of delayed trains, and the optimality is never proven for instances of case study # 4.

  Regarding the best objective indicator, CP either keeps good frequencies all along the level of perturbation, either improves over MILP above 20 % rate of delayed trains.

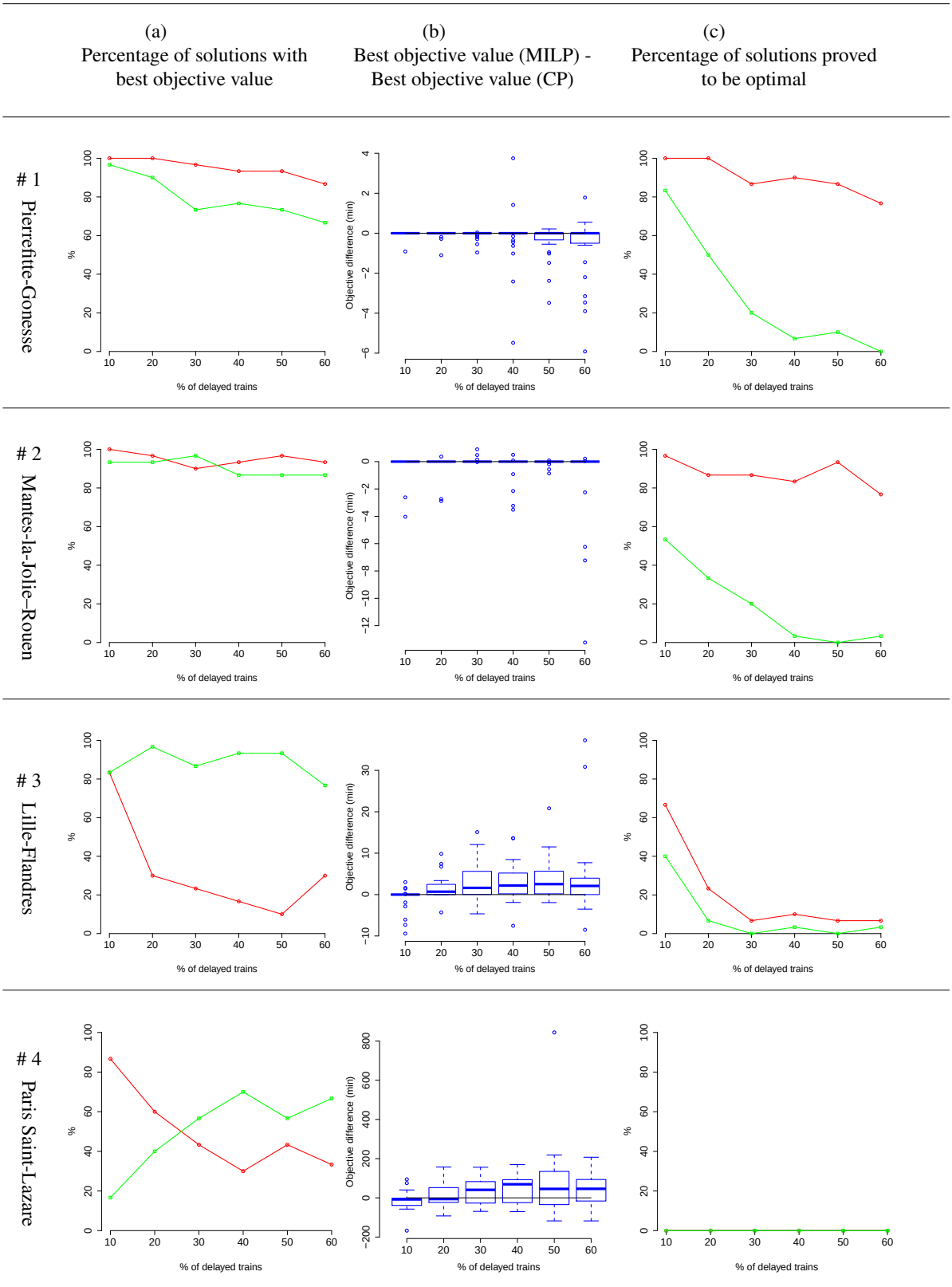


Figure 8: Experimental results for perturbation rate variation. In column 1 and 3, red circles for MILP, green squares for CP. In column 2, difference MILP minus CP.

The boxplots confirm the trend: the more we have perturbations, the more CP gives better objective values than MILP.

7 Conclusion

In this paper, we propose a new formulation of a Constraint Programming for the real-time Railway Traffic Management Problem. It is based on the concept of Time-interval variables which simplifies the formulation of optional activities. The solution method integrates Constraint Programming and Mathematical Programming techniques. Preliminary results show good performance of the proposed approach compared with the state-of-the art RECIFE-MILP algorithm.

The model is fed by two types of information: static information and dynamic information. Static information includes tds characteristics, block limits, number of aspects, platforms positions, train set parameters ... Dynamic information includes scheduled arrival and departure of trains in stations, running and clearing times of tds for trains and train delays. All information was provided by French railways and information on delays are supposed to be provided by a traffic state prediction module a few minutes before the start of the scenario. The prediction module can be based on simulation, statistics or an artificial intelligence learning technique.

This research is an addtional contribution toward the applicability and relevance of the approache of a microscopic model to tackle real-time control of traffic perturbations. Previously, the output of the European project ON-TIME Quaglietta et al. (2016) provided a proof-of-concept of a framework where the RECIFE-MILP algorithm were used in a closed-loop with a simulation environment and tested on different networks of European infrastructure managers.

As perspectives of this research, we will exploit the use of state resources to better manage opposite direction conflicts, hence to improve algorithm performance. In addition, an in-depth analysis of weaknesses and strengths of RECIFE-MILP and RECIFE-CP should allow the proposal of a hybrid solution approach.

References

Cacchiani, V., Huisman, D., Kidd, M., Kroon, L., Toth, P., Veelenturf, L., and Wagenaar, J. (2014). An overview of recovery models and algorithms for real-time railway rescheduling. *Transportation Research Part B: Methodological*, 63:15 – 37.

Fang, W., Yang, S., and Yao, X. (2015). A survey on problem models and solution approaches to rescheduling in railway networks. *IEEE Transactions on Intelligent Transportation Systems*, 16(6):2997–3016.

Großmann, P., Hölldobler, S., Manthey, N., Nachtigall, K., Opitz, J., and Steinke, P. (2012). Solving periodic event scheduling problems with SAT. In *Advanced Research in Applied Artificial Intelligence - 25th International Conference on Industrial Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2012, Dalian, China, June 9-12, 2012. Proceedings*, pages 166–175.

Hansen, I. A.; Pachl, J., editor (2008). *Railway Timetable & Traffic - Analysis, Modelling, Simulation*. Eurailpress.

Kroon, L., Huisman, D., Abbink, E., Fioole, P.-J., Fischetti, M., Maróti, G., Schrijver, A., Steenbeek, A., and Ybema, R. (2009). The new Dutch timetable: The OR revolution. *Interfaces*, 39(1):6–17.

Laborie, P. and Godard, D. (2007). Application to single-mode scheduling problems. In *3rd Multidisciplinary International Conference on Scheduling: Theory and Applications (MISTA)*., page 276–284.

Laborie, P. and Rogerie, J. (2008). Reasoning with conditional time-intervals. In *Twenty-First International FLAIRS Conference*.

Laborie, P. and Rogerie, J. (2016). Temporal linear relaxation in IBM ILOG CP Optimizer. *Journal of Scheduling*, 19(4):391–400.

Lusby, R. M., Larsen, J., Ehrgott, M., and Ryan, D. (2011). Railway track allocation: models and methods. *OR Spectrum*, 33(4):843–883.

Mascis, A. and Pacciarelli, D. (2002). Job-shop with blocking and no-wait constraints. *European Journal of Operational Research*, (143):498–517.

Pellegrini, P., Marlière, G., and Rodriguez, J. (2014). Optimal train routing and scheduling for managing traffic perturbations in complex junctions. *Transportation Research Part B: Mathodological*, 59:58–80.

Quaglietta, E., Pellegrini, P., Goverde, R. M., Albrecht, T., Jaekel, B., Marliere, G., Rodriguez, J., Dollevoet, T., Ambrogio, B., Carcasole, D., Giaroli, M., and Nicholson, G. (2016). The ON-TIME real-time railway traffic management framework: A proof-of-concept using a scalable standardised data communication architecture. *Transportation Research Part C: Emerging Technologies*, 63:23 – 50.

Rodriguez, J. (2007). A constraint programming model for real-time train scheduling at junctions. *Transportation Research Part B: Mathodological*, 41:231–245.

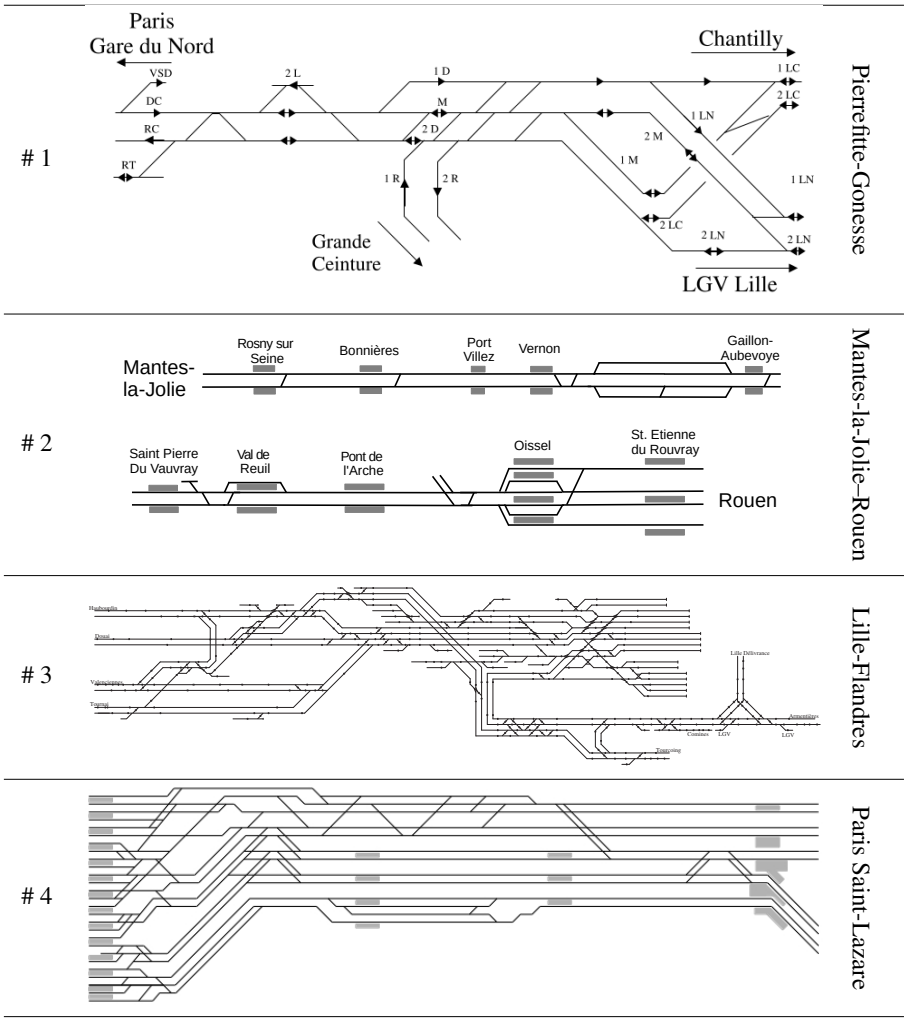
Schrijver, A. and Steenbeek, A. (1994). Timetable construction for railned. Technical report, C. W. I. Center for Mathematics and Computer Science Amsterdam. in Dutch.

Spzigel, B. (1973). Optimal train scheduling on a single track railway. In Ross, M. e., editor, *OR'72*, pages 343–352. North Holland Publishing Co.

Vilím, P., Barták, R., and Čepek, O. (2005). Extension of O(N Log N) Filtering Algorithms for the Unary Resource Constraint to Optional Activities. *Constraints*, 10(4):403–425.

Vilím, P., Laborie, P., and Shaw, P. (2015). Failure-directed search for constraint-based scheduling. In *International Conference on AI and OR Techniques in Constriant Programming for Combinatorial Optimization Problems*, page 437–453. Springer, Cham.

Appendix A - Case-studies layouts



Online microscopic calibration of train motion models: towards the era of customized control solutions

Valerio De Martinis [a, 1], Francesco Corman [a],
[a] Institute for Transport Planning and Systems IVT, ETH Zurich
P.O. 8093 Zurich, Switzerland
[1] E-mail: valerio.demartinis@ivt.baug.ethz.ch,

Abstract
The on-board collection of data related to train operation enables a better calibration of the current train motion models, which are fundamental for the elaboration of optimized train control solutions. Here, the possibility to implement an online calibration of train motion models at microscopic detail, i.e. to set the model's parameters for the single train on the go, is explored. For this purpose, a comparison of different calibration models is proposed. Then, the performances of the models are evaluated according to the requirements for online elaborations. In the end, possible further requirements and limitations on the use are discussed.

Keywords
Calibration, online, train resistances, data driven

1 Introduction

In the last years, the optimization of rail operation for different purposes (e.g., increase of network use, increase of punctuality, increase of energy efficiency) has become a primary goal for railway systems to keep being competitive in the transportation market (Corman and Meng, 2015; Rao et al. 2016). Enabling technologies and advanced modelling contribute to achieving performance goals. Nevertheless, for their effective and correct use, these must be associated with a greater specification of the models behind, to obtain more accurate results and adherence to reality.

Therefore, the calibration of train motion models is a key aspect, since these last ones constitute the backbones for the elaboration of optimized solutions for train control, such as determining precisely running time supplements, determining control actions to ensure energy saving profiles and have succession of trains at minimum separation time, in order to maximize capacity usage in bottlenecks. Some positive experiences have already shown the importance of the offline calibration (Besinovic et al., 2013). However, the variety of possible rolling stock characteristics (e.g. freight trains) and of possible operating conditions (e.g. weather conditions) limit the expected enhancements during implementation.

The sensors invasion and the consequent availability of huge dataset of on-board monitoring systems (e.g. on power used, energy consumed, speed, position, etc.) open to the opportunity to overcome these issues with a proper calibration of the train motion model. Some interesting insights have been already highlighted in recent papers, such as in Hansen et al (2016) and De Martinis and Corman (2018).

In this paper, we focused our attention on online microscopic calibration. When the train is running, we can directly determine the parameters values of the train motion model. This

allows to compute a specific control strategy for that train and to implement it on the go. In other words, there is the possibility to determine the motion model for a specific train in a specific moment, thus enabling customized train control strategies.

Main objective of this paper is therefore to investigate under which conditions an online calibration of train motion models is achievable, identify current gaps, and address possible further research. Within this work, we aim at feeding the discussion with the following contributions to the field:

- Specification of a microscopic calibration model for train motion models based on data collected on board.
- Evaluation of the performance of the proposed calibration model in relation to: track characteristics and motion phases, computation speed, number of observations.
- Discussion on possible requirements and limitations.

Finally, this paper is seen within the stream of research on automation of train control, main aim is to enhance the knowledge for future on-line calibration processes of DASs or ATP/ATO systems

A large data set provided by a Swiss train operator consisting of monitoring data collected in multiple months, related to the operation of different trains on different tracks, for many hundred trains, is used as reference.

2 Train resistances in train motion models and motion phases

Train resistances play a primary role in train operation. These largely affect train performances, travel times and energy consumptions. Travel times relate to possible saving by better exploitation of capacity, for a single train (running time reserves) and multiple trains (buffer time between minimum headway). Regarding energy consumptions, precise estimations of train resistances are needed for generating optimal train control strategies in real time. Nowadays, train resistances are computed through well-known polynomial formulations dependent upon speed (for more details, see Hansel and Pachl, 2014), some of which have been elaborated in the first decades of the last century, and sporadically updated by the scientific community, such as in Rochard and Schmidt (2000) or by train operators. Nevertheless, it usually happens that monitoring data describe a reality that cannot be totally explained by the current models; for example, Bomhauer-Beins et al. (2018) highlighted that the high variability of energy consumption between same passenger trains running on the same track with similar speed profiles can be explained by considering the weather conditions (e.g. wind, snow…).

On the other hand, the expertise acquired with the current formulation of train resistances is a valuable fix point of the train motion modeling and, at authors' knowledge, it is still the most used formulation. Generally, train motion models are used to describe the motion of a train in terms of acceleration, speed time and distance traveled. In particular, this work focuses on dynamic train motion models, which relate the forces applied to a train (tractive efforts, train resistances) with its motion. These are formulated through a dynamic approach based on Newton's second law of motion, and specified as follows:

$$F^{tr}(v) - F^{R}_{veh}(v) - F^{R}_{inf}(s) = f_t * m * {}^{dv}/_{dt} \quad (1)$$

where F^{tr} are the efforts generated by the traction unit, depending on the speed v, F^{R} are the resistances related respectively to vehicle (*veh*, dependent on speed) and to the line (*inf*,

dependent on position *s*), f_t is the mass factor, *m* is the mass of the train and *dv/dt* is the acceleration.

Line resistances (slopes and curves) are modelled as additional resistances that depend on train position. Their values depend on track characteristics (gradient and radius) and the train mass. Specifically:

$$F_{inf}^{R}(s) = F_{slope}^{R}(s) + F_{curve}^{R}(s) = m\ g \sin i(s) + mg\ 700/r(s) \quad (2)$$

In equation (2), the resistance given by slopes depends on the train mass *m*, the gravity acceleration *g* and the gradient *i*. When *i* is negative, slopes contribute positively to the train motion. The resistance given by curves depends on the train mass *m*, the gravity acceleration *g* and the curve radius *r*.

Equation (1) leads to a formulation of train motion that depends on train motion parameters. Specifically, the tractive efforts F^{tr} applied at wheels can be related to the active power measured at the traction unit:

$$P = \frac{1}{\eta} F^{tr} v \quad (3)$$

Where the constant $\eta \in (0,1)$ represents the losses related to power transmission. Vehicle resistances can be computed, according to the consolidated practice, through a polynomial formulation:

$$F_{veh}^{R} = A + B * v + C * v^2 \quad (4)$$

Where parameters *A*, *B* and *C* describe the resistances of the train. For more details please refer to [6].

Air resistances in tunnels are considered as an additional aerodynamic resistance. In particular, coefficient C of (4) is computed in open-air condition and it is increased, when the train is in a tunnel, of a quantity f_{Tun} that depends on tunnel dimensions (e.g. cross-sectional area), train dimensions and train shape (see [4] for more details). The coefficient C is therefore dependent on train position, and equation (4) can be rewritten as:

$$F_{veh}^{R} = A + B * v + \big(1 + f_{Tun} * \gamma(s)\big) C * v^2 \quad (5)$$

Where γ is a dummy variable equal to 1 when the position (s) belongs to a tunnel, otherwise it is equal to 0.

Parameters A, B and C are evaluated according to empirical formulas that consider numerous variables, such as number of axles, number of vehicles, aerodynamic coefficient representing the shape of the vehicles and the cross-sectional area of the vehicles [7].

By defining the train motion model, it is possible to identify, within the train motion data set collected onboard, the train motion phases along the track, i.e. particular regimes which happen more often, like acceleration; cruising at more or less constant speed; coasting; and braking, discussed in what follows.

Acceleration

The acceleration phase is usually intended as a variation of the vehicle speed, i.e. the acceleration rate dv/dt is not null. The positive acceleration of a train is the result of the positive contribution given by the tractive efforts applied (F^{tr} >0). Possibly, specific track characteristics, e.g. a descent slope ($i(s)$ < 0), can also contribute to it. When tractive efforts

are the only forces applied to the vehicle, acceleration is limited by the maximum performances of the train F_{max}^{tr} and the adhesion limits *Ad*:

$$\max {}^{dv}/_{dt} = \min({}^{dv}/_{dt}\,(F_{max}^{tr}), {}^{dv}/_{dt}\,(Ad)\,) \quad (6)$$

A negative acceleration, when tractive efforts are applied, means that the total contribution of the resistances in that part of the track is higher than the tractive efforts. In this condition, the train will decelerate until it will reach a new speed that allows for a new equilibrium state (i.e. resistances dependent on speed will be smaller, and resulting acceleration is null).

Cruising

The cruising phase is characterized by a constant speed.

$$F^{tr}(v) - F_{veh}^{R}(v) - F_{inf}^{R}(s) = 0 \quad (7)$$

In eq.7 the railway line slope is relevant. During descending slopes, it is possible that $F^{tr} < 0$, and this means that electric braking is applied to avoid acceleration. It is not known whether the mechanical braking is also applied or not.

Coasting

This phase is characterized by the train's inertial motion, without additional tractive effort. In this phase, tractive efforts are not applied and the resistances terms drive the train motion.

$$F_{veh}^{R}(v) - F_{inf}^{R}(s) = f_t * m * {}^{dv}/_{dt} \quad (8)$$

Braking

In the braking phase, tractive efforts are not applied and brakes are activated. In modern vehicles there are two main braking systems: the first is the electric braking, i.e. the power flow is inverted and the traction wheels work as fly-wheel, the second is the mechanical braking. . Electric braking is considered in terms of power generated by the traction unit (i.e. $F^{tr} < 0$), while data on mechanical braking efforts given by disc brakes are not collected. This lead to an incomplete knowledge of eq. (1).

3 The train motion model calibration

Typically, every model follows three main steps for its complete identification: specification, calibration and validation. The specification of a model results in its formulation, i.e. relations between the chosen variables by means of some parameters. The calibration phase sets the values of the parameters. The validation tests the goodness of the calibrated model in terms of adherence to reality.

Given a history of *k*-recorded on-board measurements ($k = 1 \ldots N$), i.e. data belonging to a given time window between a departure and a successive arrival at station (speed is 0 at beginning and end of this time window), the scope of the present section is to calibrate the resistance parameters values within such time window. We can vary the time window. Having a very small time window results in a slightly overdetermined problem, size of data and parameters are comparable. This might enable a better fit of the parameters in the time window, but a higher variability across two successive time windows (i.e. the parameters calibrated at time t and those calibrated at time t+1 might have very different values). A

long time window results in a largely overdetermined calibration problem, i.e. much more data than parameters. This results in larger sample deviation from the parameter values, but a more stable computation. A smaller time window is more reactive to changes in environment, a longer time window is less reactive to the same factors.

The calibration model is formulated as an optimization problem for parameters fitting. The setup of this specific problem requires the identification of the following parts:

- GoF (Goodness of Fit). It is the function that evaluates the adherence of the output of the specified model, given a set of coefficients, to a set of data used as reference (e.g. real world data collected).
- MoP (Measure of Performance). It is a variable that is used by the GoF operator for the evaluation of the model. For this work, F^{tr} has been considered as MoP.

The equation (1) is here treated with a difference equation approach. According to the time step Δt_k of the recorded observations (for instance, 1 sec), we formulate (1) as follows.

$$F^{tr}(v_k) - F^R_{veh}(v_k) - F^R_{inf}(s_k) = f_t * m * {}^{(v_{k+1} - v_k)}/_{\Delta t_k} \quad (4)$$

From the set *N* of recorded measurements, we can formulate the problem of calibrating the resistance parameters as follows:

$$\left(\hat{A}, \hat{B}, \hat{C}\right) = \text{argmin } GoF\left(F^{tr}\left(v(k), A, B, C, f_{Tun}\, F^R_{inf}(s_k)\right), \overline{F^{tr}}(k)\right) \forall k = 1 \dots N \quad (5)$$

subject to the following constraints:

$$F^{tr} \leq Ad = \mu_r * g * {}^{m}/_{n};\ \mu_r = 0.161 + {}^{7.5}/_{(3.6 * v + 44)} \quad (6)$$

$$F^{tr}\big(v(k)\big) \leq F^{tr}_{max}(v(k)) \quad (7)$$

Equation (6) is related to adherence conditions, where *m* is the mass of the train, *n* is the number of motorized axes of the vehicle, g is gravity acceleration and $\mu_r$is the adherence coefficient computed with the Curtius & Kniffler formula. Equation (7) ensures that the estimated tractive effort needed to win the resistances does not exceed the one produced by the traction unit in maximum power conditions. Here, the values of tractive efforts from the train motion model are a function of the speeds series (v_k for F^R_{veh}, v_k and v_{k+1} for variation of speed), *A, B, C,* f_{Tun} parameters and line resistances. Here train position is defined as the space traveled from last departure. The mass and the mass factor are considered as constant values.

4 Data set description

The data used in the current work are part of a large set of data collected by a Swiss train operator through onboard monitoring systems. Data available for each train course are:

- Onboard monitoring system. The single record of the onboard dataset is composed by time, speed, latitude and longitude position of the train (via GPS), and measurements at pantograph of power consumed and power generated. Such data is available with sampling frequency of one second.
- Track data. The single record of the track data consider each variation of value

in one or more of the following fields: radius, gradient, speed limitations (per each train category). For each variation, the position along the track is reported.

5 Experimental plan

5.1 Data and setup

Preliminary set up

For the present work, 50 runs between two consecutive stops at stations of passenger trains operating on a Swiss line have been selected. The travel time from timetable is 17 minutes, while average travel time from the monitoring system is 1024 seconds. The track is approximately 37 km long and digitally provided in LV95 Swiss coordinate system.

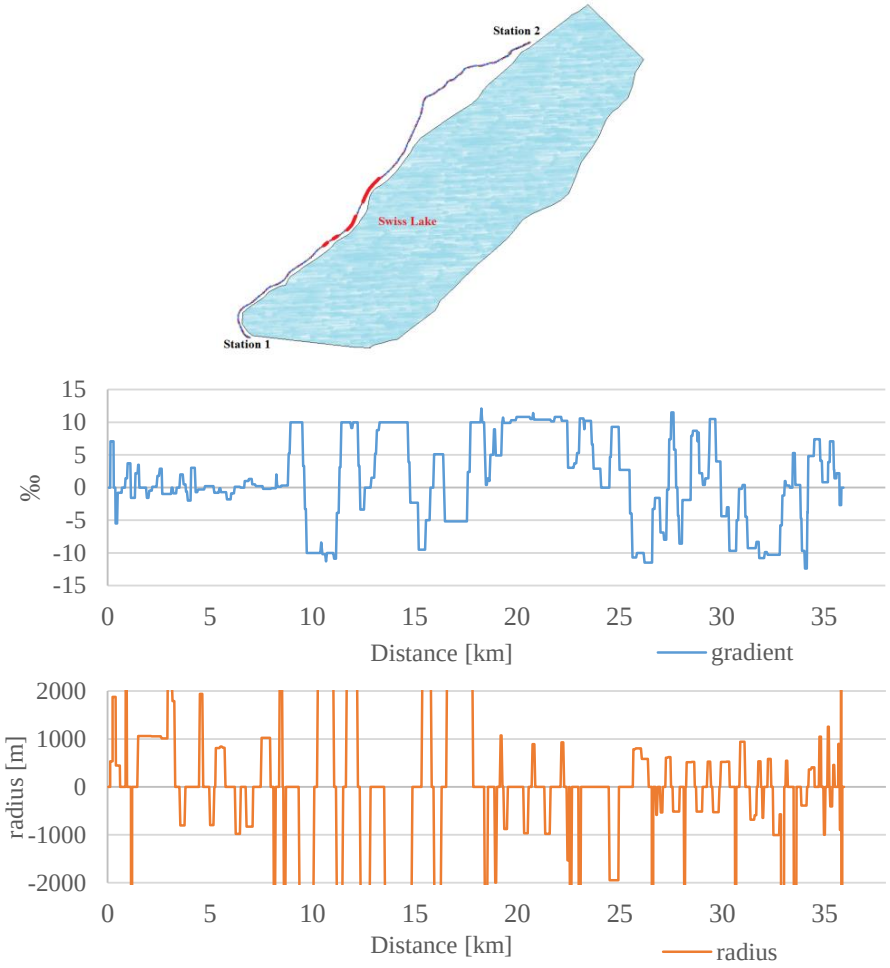


Figure 1. Top: sketch of the case study area. Middle: gradient trend along the track. Bottom: radii trend along the track (infinite radii reported as 0 for visual representation)

Along the track there are four tunnels, respectively 0.35, 0.27, 1.25, and 2.23 km long. In Figure 1, the line development is represented (omitting names of relevant features for confidentiality). The 4 tunnels are highlighted with red thick lines. The maximum allowed velocity on the line is 160 km/h (≅44.4 m/s).

The mass m of the train and the constant η, representing the transmission losses of power from the pantograph to the traction unit, have been given by the train operator. Data from tunnel resistances have been taken from previous empirical studies on Swiss tunnels [4; 8]. For the introduced discrete-time model, we chose the same sampling time of the recorded data $\Delta t = 1\ s$. The measured positions are taken from a reference starting point (0 m).

GoF functions

In this study, two GoF functions are evaluated according with the common knowledge in transport systems. At authors' knowledge, the calibration experiences in this railways are very few, therefore most common GoF used in other fields of transport systems, such as in traffic engineering problem (Ciuffo et al., 2012), are used:

- Sum of Absolute Error (SAE). It is the most widely used GoF for calibration in different fields. Its main characteristics is that it penalizes large errors.
- RMSE (Root Mean Square Error). It may return instabilities if there are low values along the set of data used for calibration. It is sensitive to outliers.

The tests have been made considering different time windows for calibration, namely 10, 30, 60 seconds (time-step of recorded data is 1 second), to understand calibration behavior in terms of speed and stability.

Train motion phase identification

The calibration models have been evaluated according to the different motion regimes, as described in section 2. While coasting and braking have very well defined characteristics in terms of power used (i.e. equal to zero in the first case and negative in the second case), acceleration and cruising phase can be confused; main reason is that even within the cruising phase there are some small speed deviations and related acceleration measures. This can be given by sensor noise or characteristics variability of the track which has not been reported in the input dataset.

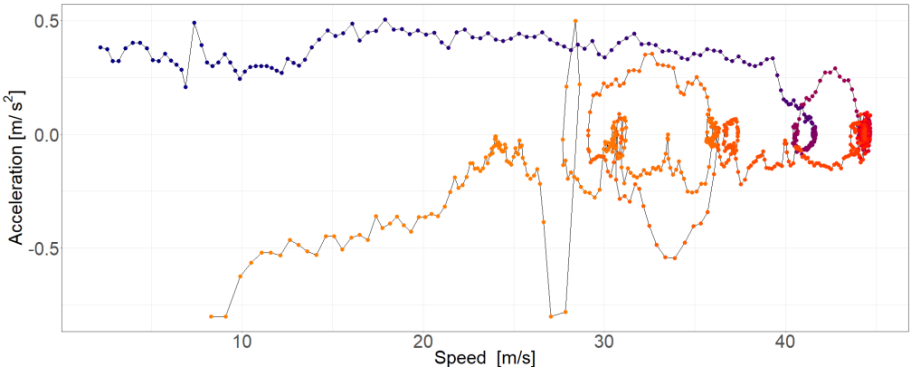


Figure 2 Acceleration vs Speed measurements

Figure 2 reports a speed/acceleration diagram of a typical train trajectory. The color indicates the evolution over space and time from the beginning (blue) to the end (orange). As shown, there are agglomerations of accelerations measurements close to zero value. These correspond to those speeds where cruising occurs; for the trajectory reported, different cruising have been performed at different speeds, like 30, 36, 42 and 44 m/s. The acceleration /braking phases, and the transition to cruising, are well distinguishable

In Figure 3, we plot a histogram of the acceleration recorded for the entire train set of train trajectories. From the acceleration values distribution (Figure 3), two bounds [-0.04, +0.04] have been set up to make the distinction between these two phases. The positive one will separate acceleration to cruising, the negative one will include some small negative accelerations that are not intended as part of the coasting phase.

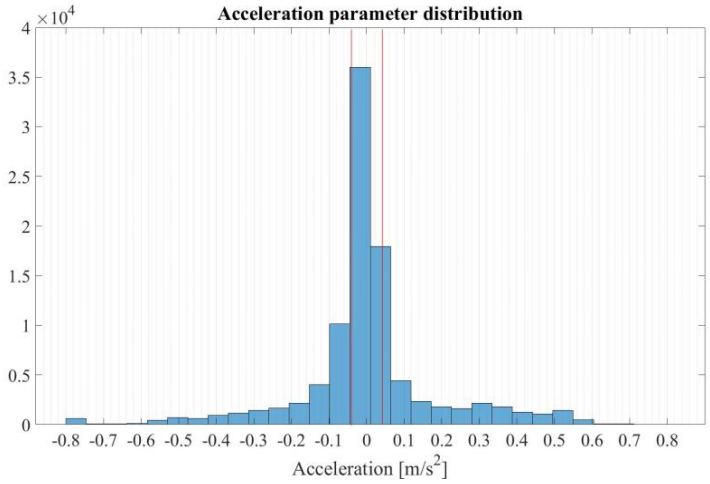


Figure 3. Acceleration measurements distribution

5.2 GoF evaluation

A first evaluation refers to the GoF function that better performs the calibration problem as specified in (5). The calibration problem has been implemented in MatLab, using the Optimization toolbox. The 2 different GoF functions have been tested with a Multi Start Gradient-based solver and 40 random starting points, to avoid being stuck in local minima. The test has been conducted considering the smallest time window of data used for calibration, i.e. 10 seconds, and a single trajectory for reference.

Results are shown in Figure 4. Here the error E_k at each step in terms of kN is reported for both the SAE and the RMSE. Specifically, the error has been evaluated through:

$$E_k = F^{tr}(v_k) - F^R_{veh}\left(\overline{A_k^{10}}, \overline{B_k^{10}}, \overline{C_k^{10}}, v_k\right) - F^R_{inf}(s_k) - f_t * m * {}^{(v_{k+1} - v_k)}/_{\Delta t_k}$$
$$\forall k = 1 \dots N \qquad (8)$$

Where $\overline{A_k^{10}}, \overline{B_k^{10}}, \overline{C_k^{10}}$ are the parameters A, B and C evaluated at time k and calibrated considering the previous 10 measurements. The error has been compared with the A, B, C values computed via Sauthoff formulation (see [6] Ch. 4 for more reference). In this latter case, it results:

A=2048 [1000*$^{kg*m}/_{s^2}$]; B=98,4 [1000*$^{kg}/_{s}$]; C=6,89 [1000*$^{kg}/_{m}$]. (9)

From the Figure 4, the error committed with the calibrated parameters, for both the two GoF functions, is negligible when compared to the error using the classical formulation (which has errors one order of magnitude larger). In this latter case, either the formula underestimates the vehicle resistances (due to different shape of train, weight, etc.) and/or some non-modelled phenomena are affecting the train motion. An online calibration as proposed in this paper would overcome both of these issues.

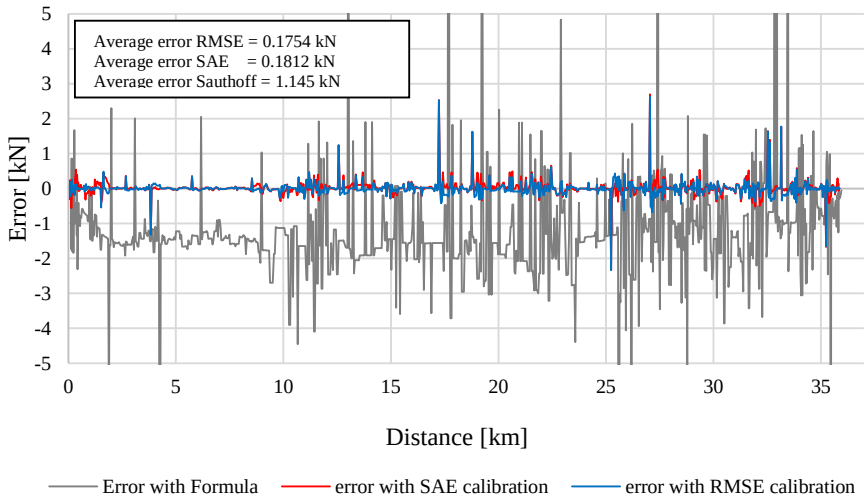


Figure 4. Error for different parameters set for a typical trajectory.

Since the results of the two GoF appear to have no substantial differences, the following elaboration will be conducted with the SAE formula. This mainly because the single iteration of the optimization process is computed faster (6.83 seconds vs 14.55 seconds).

5.3 Calibration results Inter-train characteristics

Given the calibration problem setup, calibrations on 50 different train runs have been performed. Results are shown grouped by the train motion phases as proposed before. For those phases where tractive energy is positive (energy consuming phases) results are reported, in terms of average values per train, in Figure 5 (left: acceleration; right: cruising). Results are reported as histograms of the values of the Parameters A, B and C, respectively in top, middle and bottom row of Figure 5. A kernel estimate of the distribution of values is also provided (red curve).
It is possible to note that there is a quite wide range of values between different trains both during acceleration and during cruising phase. In all cases except for B parameters in cruising phase, there is a well-defined peak value, that reflects a common characteristics among all trains at all samples. This could be for instance the similarity in train characteristics on the track, and/or the same external conditions (e.g. weather, humidity). B parameters in cruising phase have more than one peak, which may be given by a larger

influence of parameter C at higher speeds. Parameter C is well-defined in cruising phase (i.e. a clear peak in the histogram is present). This is because aerodynamics counts more at higher speeds, which are usually those used for cruising regimes.

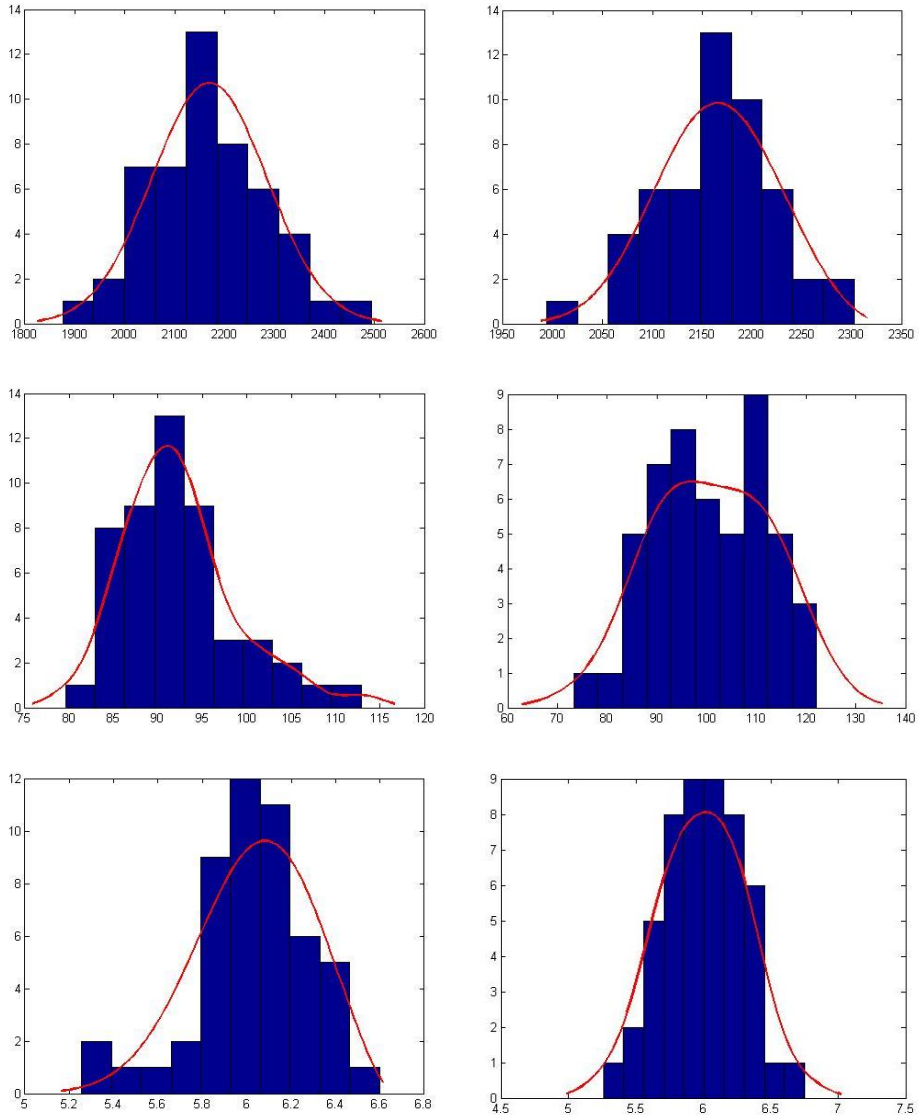


Figure 5. Parameters values distribution for the acceleration (left column) and the cruising phase (right column). Top row: A value distribution; middle row: B value distribution; Bottom row: C value distribution.

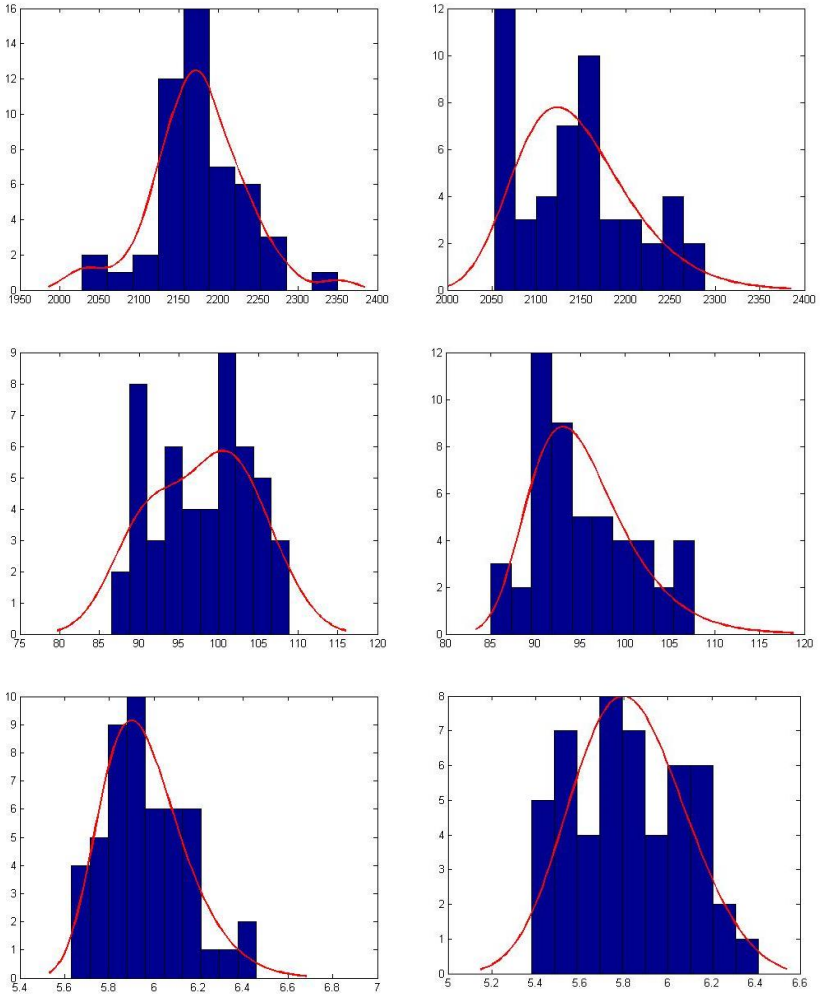


Figure 6. Parameters values distribution for the coasting (left column) and the braking phase (right column). Top row: A value distribution; middle row: B value distribution; Bottom row: C value distribution

Figure 6 shows the calibration results for those motion phases where tractive efforts were not activated, i.e. coasting (left column) and braking phase (right column). The figure shows the histogram of the computed values of A B and C (top, middle, bottom row).

Results show more than one peak of values, especially in the braking phase. This is most probably due to the missing tractive efforts information (i.e. there was no measured tractive effort) that could be considered in the optimization process. In particular B parameters of both coasting and braking phases are very sparse. Probably, the variation in one parameter reflects in the variation in the other parameters, especially B-C. In fact, issues in calibration of Parameter C can be expected, as during the braking phase resistances from the air and

resistances from the braking system work together and it is difficult to separate their working. Moreover a braking action might appear as a less aerodynamic resistance for this same reason. . Additionally, C parameters related square of speed are affected also by the small relevance of aerodynamic resistances at low speeds. Moreover, during braking phase the mechanical braking efforts are not known thus a precise estimation of resulting tractive effort is difficult.

5.4 Calibration results Intra-train characteristics

In this analysis, calibration of all considered trains is evaluated with respect of the train motion phases. Calibrations have been performed considering different time windows, i.e. the dataset for calibration had 10, 30 or 60 seconds (and an equivalent amount of samples).

We report those in Figures 7, 8 and 9, as scatter plots of the parameter estimated, along space. For each motion phase identified, we use a different color. This allows identifying variation in the distribution of the parameters, as dependent on the motion phase, and the position along the track. For each Figure, the three plots report the time window 10, 30 60 seconds, at top, middle, bottom respectively.

We start with parameter A (Figure 7), which shows bigger variances (i.e. larger spread along the vertical axis) at lower speeds (i.e. the beginning and end of the space axis), probably because the other speed-dependent parameters do not explain resistances well at lower speeds. This is associated to acceleration (green) and braking (blue) phases. Increasing the size of data set used for calibration yields, as expected, a lower variation between trains. The cruising phase (yellow) seems to give more stable evaluation along the track. This is expected since A parameter is not speed dependent.

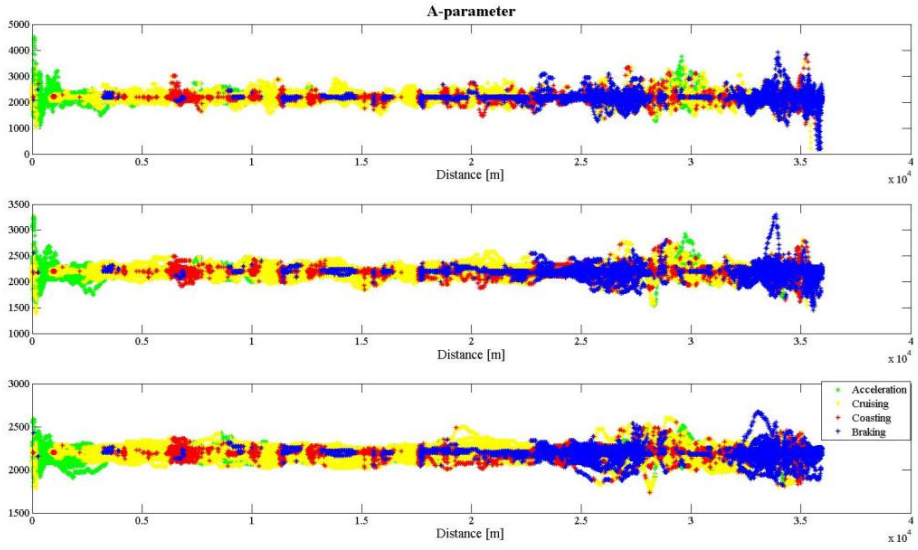


Figure 7. A-parameter values for different trains. Top: data set size is 10 seconds; middle: data set size is 30 seconds; bottom: data set size is 60 seconds

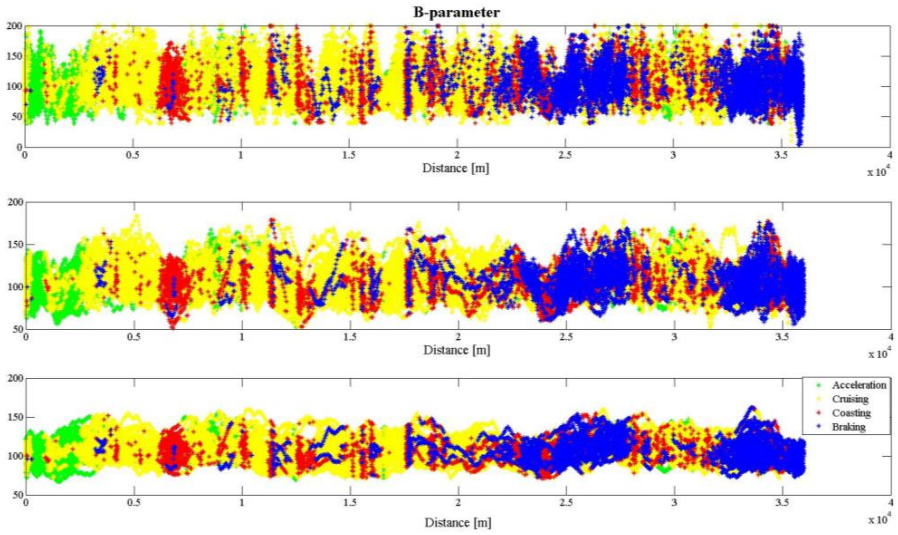


Figure 8. B-parameter values for different trains. Top: data set size is 10 seconds; middle: data set size is 30 seconds; bottom: data set size is 60 seconds

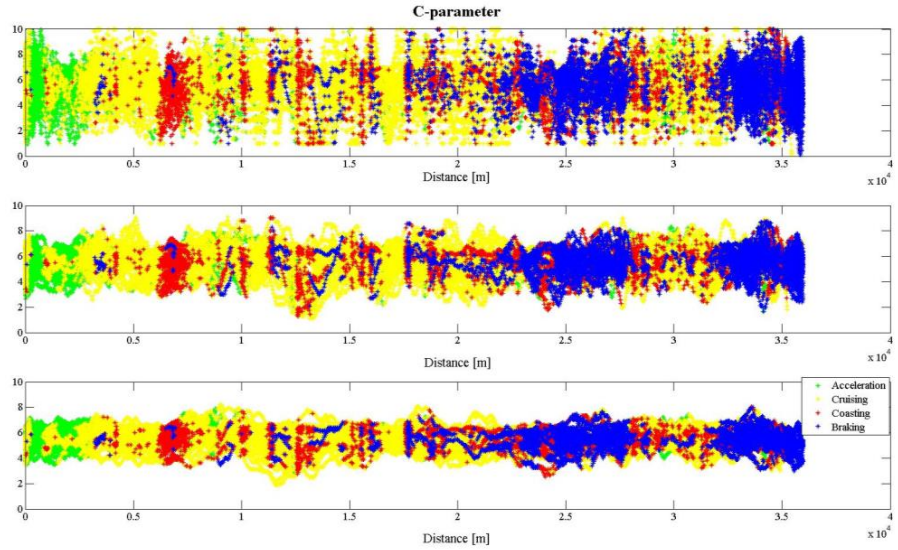


Figure 9. C-parameter values for different trains. Top: data set size is 10 seconds; middle: data set size is 30 seconds; bottom: data set size is 60 seconds

The same analysis for parameters B and C (Figure 8 and 9) shows a larger variability than parameter A, although their variability is reduced by considering a bigger set of data for calibration. Especially when analyzing the variation of parameters values with larger time windows, it is possible to see areas where the variation is lower, especially between the 15th and 20th km. On the one hand, this area corresponds to a large stretch where speed is almost constant, thus calibration is difficult. On the other hand, this suggest a possible influence of track conditions also for parameters B and C that needs further investigations. A possible idea to model variation in resistances at constant speed is to have a more rich description of the generalized resistance as dependent on the position, similar to gradients and curves. The computation times required for the different time windows used, i.e. for the optimization process for the time windows of 10, 30 and 60 seconds, are respectively 6.83, 39.2 seconds and 1 minute 12.54 seconds. In other terms, already with this preliminary implementation the optimization performed with a 10 seconds time window can be run in less than 10 seconds.

6 Recommendations, possible limitations and conclusion

This work aimed at determining the conditions under which a precise estimate of parameters of resistance of train motion can be calibrated at microscopic level (i.e. detail of one second), online (i.e. fast) and specifically per train (i.e. fitting the particular characteristics of each vehicle and train run). Within the set of experiences here presented, some interesting conclusion can be highlighted. Within the data set of trains used for this work, there is a variability between different trains and within the single runs. This can be modelled by defining appropriate set of resistances parameters A, B, C that can fit pretty well the vehicle resistances. Such a probabilistic description of train resistances would require the extension of current train running time estimation models and energy consumption estimation models to distributions rather than crisp numbers.

The calibration procedures work better in case of variation of speed, and even more, in case of positive acceleration. Assuming the current resistance formula dependent on input values of speed gradient and curvature, a variation in resistances when those three input values could not be explained. Thus, calibration of resistance parameters at cruising is particularly challenging, unless some artificial controlled variation of speed are introduced.

Keeping the trinomial formula, the physical meaning of the parameters are thus enlarged to accept non–modelled dynamics. The formula loses a bit of physical interpretation in order to describe more accurately measured resistances. A completely online approach, learning the parameters based on the data and not based on characteristics of trains such as number of axles and weight might result in better fit, but more difficult explanation to domain experts. Nonetheless, the added value of such a black-box approach, able to reduce error by an order of magnitude might well be worth it. More sophisticated calibration and /or regression techniques could be used to identify most relevant dynamics, which could be worthwhile to include in the formulation of resistance. A study of the variation of those parameters and other known unmodelled factors (for instance weather) could also be proposed.

With the proposed algorithm the online calibration for a given dataset is pretty fast. With a time window of 10 seconds of data for calibration, the values of the resistance parameters are returned in less than 7 seconds. By increasing the sample size, the stability of the parameters value over running time is increasing but the time for elaboration increases consequently.

The present work proposed the comparison of 2 main GoF, SAE and RMSE, for the

calibration process. Both alternatives perform well in the calibration procedure. The calibration with RMSE takes a longer time than with SAE, however the time required in this case (14.55 seconds) is also compatible with many online procedures (i.e. speed optimization for energy saving). Both approaches deliver error one order of magnitude lower than the formula from the books (not fitted to the specific conditions of track and train). The infrastructure itself seems to affect the calibration of the vehicle resistances, which are, in the model considered, independent from it. This suggest that possible influences are not modeled and further investigations to deepen this assumption are needed. Extension of the resistance formula might be an interesting development.

An open question is the level of accuracy for the use of online calibration in downstream systems, i.e. from traffic control to train control. This aspect cannot be treated independently, but it must be related to the technology applied and the specific type of train control strategy. Basically, continuous calibration and the variation in optimal driving strategies must consider the current driving system and the effectiveness of this variation. Considering for example energy saving train control strategies, the variation of driver's instructions, which can be enhanced by the proposed online calibration system, should be adapted to e.g. the human attitude to follow the instructions.

The main limitations are seen in the field of application, where the total time from measurement to application is relevant. We must consider: the time for data collection phase, the time for calibration, and the time for elaboration of specific solutions (e.g. for energy saving), presentation to user and/or acceptance, and the duration for implementation. This can determine lead times of multiple tens of seconds, which of course would pose requirements on the process and affect also the expected benefits. A fixed configuration of passenger trains may lead to prefer an offline procedure based on statistical evaluation of resistances parameters. Nevertheless, this means to neglect the influence of specific conditions such as weather and occupation rate of coaches. Freight trains may instead largely benefit from this online specification.

In the end the proposed model can enhance the definition of the current train motion modeling and its application to train control problems, towards more specific and accurate elaborations, such as on energy consumption estimation, braking curves, and train location, which are important key aspects for the next generation rail operation.

Acknowledgement

This research project is financially supported by the Swiss Innovation Agency Innosuisse and is part of the Swiss Competence Center for Energy Research SCCER Mobility.

References

[1]. Corman, F., Meng, L., 2015. "A review of online dynamic models and algorithms for railway traffic management". *IEEE Transactions on ITS*, 16 (3), pp. 1274-1284

[2]. Rao, X., Montigel, M. and Weidmann, U., 2016. "A new rail optimisation model by integration of traffic management and train automation". *Transportation Research Part C: Emerging Technologies*, vol. 71, pp. 382-405.

[3]. Bešinović, N., Quaglietta, E., Goverde, R.M.P., 2013. "A simulation-based optimization approach for the calibration of dynamic train speed profiles". *Journal of Rail Transport Planning and Management*, vol. 3 (4), pp. 126-136.
[4]. Hansen, H.S., Nawaz, M.U. and Olsson, N., 2016. Using operational data to estimate the running resistance of trains. Estimation of the resistance in a set of Norwegian tunnels. *Journal of Rail Transport Planning and Management*, vol. 7, pp. 62-76.
[5]. De Martinis, V., Corman, F., 2018. "Data-driven perspectives for energy efficient operations in railway systems: Current practices and future opportunities". *Transportation Research Part C: Emerging Technologies*, vol. 95, pp. 679-697,
[6]. Hansen, I. A., Pachl, J., 2014. *Railway Timetabling & Operations*. Hamburg: Eurailpress.
[7]. Rochard, B. P., Schmid, F. 2000. "A review of methods to measure and calculate train resistances". *Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*, vol. 214 (4), pp. 185-199.
[8]. Bomhauer-Beins, A., Weidmann, U., Schranil, S., 2018. Estimation of the energy saving potential of automation in railway operations [Abschätzung des Energiesparpotentials der Automatisierung im Bahnbetrieb]. *EB - Elektrische Bahnen*, vol. 116 (4-5), pp. 150-156.
[9]. Ciuffo, B., Punzo, V., Montanino, M., 2012 The Calibration of Traffic Simulation Models : Report on the assessment of different Goodness of Fit measures and Optimization Algorithms MULTITUDE Project – COST Action TU0903. Available at: http://publications.jrc.ec.europa.eu/repository/handle/JRC68403

Predictive Model of Train Delays in a Railway System

Weiwei Mou [a], Zhaolan Cheng [a] , Chao Wen [a,b,1]

[a] National Engineering Laboratory of Integrated Transportation Big Data Application Technology, Chengdu Sichuan 610031, China;

[b] National United Engineering Laboratory of Integrated and Intelligent Transportation, Southwest Jiaotong University, Chengdu Sichuan 610031

[1] E-mail: wenchao@swjtu.cn, Phone: +86 13882255078

Abstract

Delay prediction is an important issue associated with train timetabling and dispatching. Based on real-world operation records, accurate estimation of delays is of immense significance in train operation and decisions of dispatchers. In the study, we establish a model that illustrates the interaction between and the factors affecting the same via train operation records from a Dutch railway system. Based on the main factors that affect train delay and the time series trend, we identify the independent and dependent variables. A long short-term memory (LSTM) prediction model in which the actual delay time corresponded to the dependent variable is established via Python3.6. Finally, the prediction accuracy of the random forest model and artificial neural network model is compared. The results indicate that the LSTM model outperforms other models.

Keywords

Railway, Real-world data, Delay prediction, LSTM model

1 Introduction

Delay prediction is a process of estimating delay probability based on known data at a given checkpoint and is typically measured via arrival (departure) delay. The key to making delay prediction based on actual operational data involves establishing the relationship between train delays and various characteristics of a railway system. This provides a basis for the operator's scheduling decision.

From a strategic and tactical viewpoint, the accurate prediction of train delays is of immense significance. At a strategic level, accurate train delay prediction is conducive to analyzing capacity of railway and effectiveness of its route planning. It is well known that operators tend to reduce train delays by investing in infrastructure. Accurate delay

prediction can also detect habitual delays in railway routes and potential conflicts in train operation in a timely manner. This enables operators to improve infrastructure for specific routes, and thereby promotes the overall transport efficiency of the railway system. With respect to tactical level, accurate delay prediction is tremendously significant in the establishment of a flexible and stable train diagram and aids in improving the stability of train operation plan. Timetables are tested for robustness via probability distributions of process durations that are derived from historical traffic realization data. Conclusions from the tests are subsequently used to improve timetable robustness (Medeossi et al. (2011))

2 Literature Review

Machine learning methods have been widely used in train delay prediction, which are roughly divided into two categories, namely traditional statistical machine methods (including correlation analysis, linear regression, Markov chain, Bayesian network, and random forest) and neural network machine learning (mainly including support vector, neural network, and deep learning).

Traditional statistical machine learning methods consider train operation performance as model-driven data to update algorithm structure and parameters in time such as delay probability updating in Bayesian network and pruning of a decision tree. Berger A. (2011) proposed a stochastic model of delay propagation to predict train arrival and departure delay events. The model is suitable for all public transportation systems and requires online prediction. The actual delay data of the train should be updated in real time. Based on the train operation data of the Netherlands railway network, extant studies established several models via traditional statistical machine learning methods including a train delay prediction model based on network graph (Huisman et al. (2002),Yuan and Hansen (2007)) and a train stop time and train operation performance prediction model based on distribution statistics(Meer et al. (2009),Goverde et al. (2013)). The results obtained by Olsson and Haugland (2004) indicate that passenger management is an important factor that affects train punctuality in congested areas while the management of train crossings is the key factor that affects train punctuality in non-congested areas. Flier H et al. (2009) combined linear regression and combination model to predict delay based on the on-line train delay monitoring data of the Swiss railway network. The model tested the regional corridor of Lucerne and achieved good prediction results without considering station capacity constraints. Gorman (2009) used statistical methods to forecast the average monthly train running time, and the average absolute percentage error corresponded to

4.6%. The train running process is typically considered as a Markov process. Train running delay is predicted (Barta et al. (2012), Şahin (2017),Kecman et al. (2015)) based on the deduction of train running state. The delay probability updating mechanism of the Bayesian network simulates the process of dispatcher updating the delay probability based on experience and train operation data. It is also used to establish a delay prediction model(Lessan et al. (2018), Francesco and Pavle (2018), Kecman and Goverde (2015b)) that utilizes robust linear regression, regression tree, and random forest models to predict the train running time and dwell time. Furthermore, robust linear regression was improved, and a local model was proposed for local routes and sections. The results indicated that the local model exhibited higher prediction accuracy.

It is not necessary for neural network machine learning methods to be based on prior scheduling knowledge. They realize train delay prediction by learning useful features from massive data. Marković et al. (2015) determines the effect of the infrastructure on train delays by experts and then uses the support vector machine model to predict the arrival time of a train at a station. When compared with the ordinary artificial neural network model, this indicates that the support vector machine model exhibited better prediction effect. Based on the actual data of Wuhan-Guangzhou high-speed railway, Chen et al. (2015) proposed three models, namely least squares method, support vector machine and least squares support vector machine models, to determine train location and predict train delay. Specifically, ANN was used to establish the delay prediction model, and a data-driven model was constructed based on the train operation data in Iran and Germany. The model validation results indicated that the prediction accuracy of the model is high (Yaghini et al. (2013)，Peters et al. (2005)).

Most recently, a shallow and deep extreme learning machine (DELM) was proposed in conjunction with the rapid development of big data technologies. Oneto et al. (Oneto et al. (2017b), Oneto et al. (2016)) presented a data-driven TDPS for a large-scale railway network to provide useful information on RTC processes by using state-of-the-art tools and techniques. The system extracted information from a large amount of historical train movement data using big data technologies, learning algorithms, and statistical tools. The described approach and prediction system were validated based on real historical data in six months. The results revealed that the DELM outperformed the current technique, and this was mainly based on the event graph proposed by Kecman and Goverde (2015a). Oneto et al. (2017a) developed a data-driven dynamic train delay prediction system based on the findings of Oneto et al. (2017b).This integrated heterogeneous data sources to deal with varying dynamic systems via DELM. The system exploited state-of-the-art tools and techniques, was completely data-driven, and did not require any prior information on the

railway network.

When compared with the traditional statistical machine learning model, deep learning uses deep neural network models for learning. The steps it learns corresponds to signal-feature-value. The first step involves not determining via learning the structure of the input data and not via random initialization. Therefore, the initial value is closer to the global optimum, and the model achieves better results. Overall, it corresponds to a layer-wise training mechanism. If the traditional neural network reaches more than seven layers, then the residual propagation to the foremost layer is extremely low, and gradient diffusion occurs, and this affects the accuracy of the model. When compared to traditional neural networks, deep learning reduces the number of neural network parameters and adds new structures (for e.g., LSTM and ResNet), a new activation function (ReLU), new weight initialization methods (for e.g., layer-by-layer initialization and XAVIEER), new loss functions, and new over-fitting methods (for e.g., Dropout and BN). It is characterized by a deep neural network selection that overcomes artificial choices. Currently, the prediction model of the train arrival delay is not refined. The research means and prediction accuracy are limited. Generally, from the time series perspective, it is common to consider multi-attributes to obtain the delay prediction. However, a few studies focus on the application of deep learning technology to predict train delays. In the study, the LSTM neural network model in deep learning is applied to prediction of train delays, and this is mainly because the propagation mechanism of train delays is complex and exhibits a non-linear relationship in time and space. The LSTM neural network exhibits a complex structure, and this can be used for non-linear fitting of data related to train delays to realize coding and decoding of time series data. The essential relationship between train delays and impact factors is better revealed via deep learning of large data samples and self-selection of features, and this improves the prediction accuracy of train delays.

Based on the actual running data of the Dutch railway Rotterdam Central to Dordrecht section, the study uses the LSTM model to predict the train arrival delay, and this lays a theoretical foundation for a dispatcher's decisions. The main structure of the study is as follows: Section 3 mainly describes the data of train delays. Section 4 introduces LSTM model for arrival delay prediction. Section 5 presents model forecast accuracy analysis and model evaluation. Section 6 discusses the main conclusions and applications.

3 Data Description

The actual data of train operation in the study ranges from Rotterdam Central to Dordrecht

section of the Dutch railway system, and this contains seven stations, namely Rotterdam Central (Rtd), Rotterdam Blaak (Rtb), Rotterdam Lombardijen (Rlb), Barendrecht (Brd), Zwijndrecht (Zwd), and Dordrecht (Ddr). The data includes delays of all trains in seven stations and six sections. The time span corresponds to 66 working days ranging from September 4, 2017 to December 8, 2017. The data records include the date, train number, train characteristic, location, train activity, planned time, realization, delay jump, and delay cause. A few examples of the data are shown in Table 1.

Table 1:Part of the original data table

Traffic Date	Trainnumber	Train Characteristic	Location	Activity	Planned Time	Realisation
2017/9/4	5195	SPR	Zwd	K_A	1:14:18	1:14:43
2017/9/21	2274	IC	Brd	K_A	22:59:00	22:59:10
2017/9/29	5131	SPR	Ddr	A	9:20:00	9:20:21
2017/11/13	5025	SPR	Rtd	A	7:39:00	7:39:15

4 Train Arrival Delay Prediction Model

4.1 Selection of characteristic variables

Delay prediction is a process of estimating the probability of train delays at subsequent recording points based on train operation history data, and this is typically determined by arrival delays. It is assumed that a train is currently located at station s_n, the former station and the subsequent station to arrive are donated by s_{n-1} and s_{n+1} respectively. s_{n+1}. Based on the train delays at s_n and s_{n-1} stations and scheduled running time of trains at sections (s_{n-1}, s_n), (s_n, s_{n+1}), the study predicts the arrival delays of trains at s_{n+1} stations. As shown in Fig. 1, the train arrives at the station s_n at time t_n^A on schedule and starts at the same station at time t_n^D. However, in the actual operation process, given various interference factors, the train can deviate from the timetable to generate the actual arrival time $\hat{t}_n^A$ and actual departure time $\hat{t}_n^D$. Figure 1 shows successive stations (s_{n-1}, s_n , and s_{n+1}) with the parameters in parentheses indicating the scheduled time and actual time of the event. The train delay can be typically divided into arrival delay and departure delay. The difference between the actual and scheduled times ($\hat{t}_n^A - t_n^A$) and ($\hat{t}_n^D - t_n^D$) indicate the arrival and departure delays, respectively, of the train at station s_n.

The train can be delayed due to various disturbances in the operation process. Six

parameters are selected after the analysis of the train arrival delays at the station to constitute the feature space (F). The study assumes that the parameters affect the future delay of the train, and thus the future arrival of the train is predicted based on the selected parameters.

The feature variables included in the feature space (F) are as follows:

1.Train Characteristic(X_1)

There are three main characteristics of trains running in Rotterdam Central to Dordrecht section of the Netherlands railway system, namely regional train stopping at station (SPR), intercity train stopping at large station (IC), and empty train (LM).

2. Departure delay time of the train at the current station(X_2)

The actual departure delay time of the train at the current station s_n indicates the difference between the actual departure time of the train at station s_n and the planned departure time. The equation corresponds to $\hat{t}_n^D - t_n^D$, which is accurate to seconds.

3. Arrival delay time of the train at the current station(X_3)

The actual arrival delay time of the train at the current station s_n indicates the difference between the actual arrival time of the train at s_n station and planned arrival time. The equation corresponds to $\hat{t}_n^A - t_n^A$, which is accurate to seconds.

4. Departure delay time of the train at the last station(X_4)

The actual departure delay time of the train at the last station indicates the difference between the actual departure time of the train at s_{n-1} station and planned departure time. The equation corresponds to $\hat{t}_{n-1}^D - t_{n-1}^D$,which is accurate to seconds.

5. Planned running time of the train in the last section(X_5)

The calculation equation for the planned running time between the last station s_{n-1} and current station s_n corresponds to $t_n^A - t_{n-1}^D$, whichis accurate to seconds.

6. Planned running time of the train in the next section(X_6)

The calculation equation for the planned running time between the current station s_n and next station s_{n+1} corresponds to $t_{n+1}^A - t_n^D$, which is accurate to seconds.

The output variable of the delay prediction in the study denotes the arrival delay time (Y) of the train at the next station. The delay prediction data based on the aforementioned characteristic variables are shown in Table 2. The expression is detailed as follows:

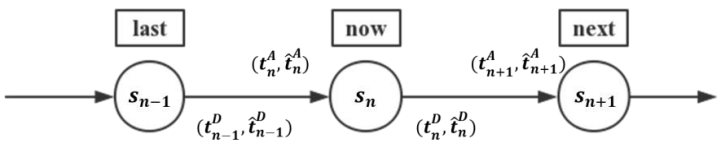


Figure 1: General scheme of train movements at three successive stations

$$Y = \varphi(X_1, X_2, X_3, X_4, X_5, X_6). \quad (1)$$

Where Y denotes the train arrival delay (output variable), X_1, X_2, X_3, X_4, X_5, and X_6 denote the train delay influence factors (input variables), and φ denotes the machine learning algorithm model.

Table 2:Modeling data table

Date	Train number	The Last Station	The Current Station	The Next Station	X1	X2	X3	X4	X5	X6	Y
2017/9/4	5139	Brd	Zwd	Ddr	SPR	113	17	-7	180	300	140
2017/10/2	2216	Sdm	Rtd	Rtb	IC	106	124	138	132	300	124
2017/10/26	2214	Rtd	Rtb	Rtz	IC	819	809	781	120	132	812
2017/11/7	2218	Rlb	Brd	Zwd	IC	215	215	190	60	240	179
2017/12/8	5027	Dtz	Sdm	Rtd	SPR	128	94	116	378	300	80

4.2 LSTM model

The LSTM model was proposed by Hochreiter et al. to improve the model based on RNN. In a conventional RNN, the hidden layer generally corresponds to an extremely simple node such as Tanhwhile the LSTM improves the simple node of the hidden layer into a storage unit. The basic structure of the storage unit is shown in Figure 2. The storage unit is composed of an input gate *i*, an output gate *o*, a forgetting gate *f*, and a memory cell *c*. In forward propagation, the input gate determines when to activate the incoming storage unit while the output gate determines when to activate the outgoing storage unit. In reverse propagation, the output gate determines when to allow errors to flow into the storage unit, and the input gate determines when to let it flow out of the storage unit. The input gate, output gate, and forgetting gate constitute keys to control information flow. The operation principle of the storage unit is expressed in terms of equations (2)–(6) (Bengio et al. (2002),Greff et al. (2016) ,Gers et al. (2002)) as follows:

$$i_t = \delta(\boldsymbol{W_i} x_t + \boldsymbol{U}_i h_{t-1} + \boldsymbol{V}_i C_{t-1} + b_i). \quad (2)$$

$$f_t = \delta\left(\boldsymbol{W_f} x_t + \boldsymbol{U}_f h_{t-1} + \boldsymbol{V}_f C_{t-1} + b_f\right). \quad (3)$$

$$c_t = f_t \cdot c_{t-1} + i_t \cdot tan\, h(\boldsymbol{W_c} x_t + \boldsymbol{U}_c h_{t-1} + b_c). \quad (4)$$

$$o_t = \delta(\boldsymbol{W_o} x_t + \boldsymbol{U}_o h_{t-1} + \boldsymbol{V}_o C_t + b_o). \quad (5)$$

$$h_t = o_t \cdot tan\, h(c_t). \quad (6)$$

Where ct denotes the calculation method of memory cells at time t; ht denotes all outputs of LSTM units at time t; W, U, V, and b denote the matrix of coefficients and vector of offset; δ denotes the activation function sigmoid; · denotes a point multiplication operation; and it, ft, and ot denote the calculation methods of the input gate, forgetting gate, and output gate at time t, respectively. As shown in Figure 2, the outputs of the three gates of the input gate, forgetting gate, and output gate are connected to a multiplier element to control the input and output of information flow and the status of cell units respectively.

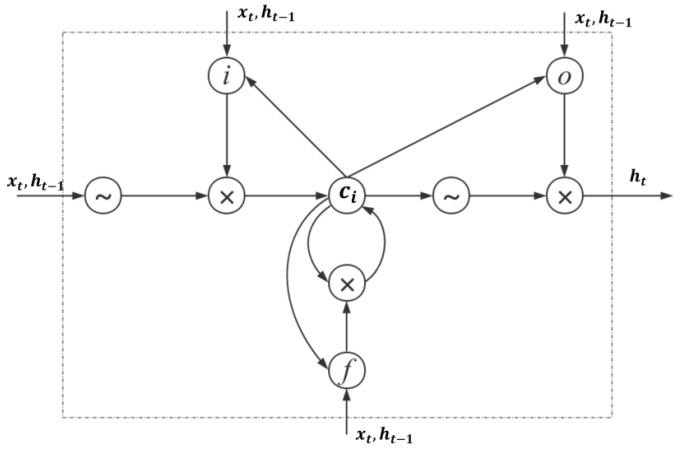


Figure 2: Basic structure of the LSTM storage unit

In the actual operation of trains, given the mutual restriction between trains, the delay of the forward train can affect the backward train and result in the lateral propagation of the delay. The LSTM model assumes time series format data as input, and its results at any *t*-time are based on the results at the previous time and input data at the current time. This mechanism enables the preservation and reuse of time series information in the model for a long period such that it learns the knowledge of time series correlation in time series data.

The LSTM model for delay prediction is constructed as follows:

(1) Seven stations in Rotterdam Central to Dordrecht are selected to extract the arrival delay time Y and corresponding feature space (F). All train delays and their extraction attributes are sorted based on the actual train operation sequence, and training data sets and test data sets are divided. As shown in Fig. 3, the first row in the figure indicates the

train arrival delay time (Y), and the second row indicates the characteristic space (F) of the influence factors of the delay time. Specifically, i denotes the train number; s_n denotes the station number; and the sliding window length l denotes the number of trains that are predicted to be entered each time. Hence, the effect of the previous l trains is considered on the current train delay.

(2) Determination of parameter l: The delay time and influencing factors of each train are treated as time series. The model considers the interaction relationship between different train numbers by inputting multiple trains each time. After repeated verification, when l=1 the best predictions can be obtained. Thus, only the effect of the previous train delay on the arrival of the current train is considered. This is mainly due to the long arrival time interval between different trains in Rotterdam Central to Dordrecht section of the Netherlands and tweak interaction between trains.

(3) After determining the optimal number of input trains, the model structure and parameters (for e.g., hidden layer number, neuron number, learning rate, optimizer, and dropout rate) are optimized to obtain the optimal parameters and structure of the model and predict the arrival delay of the train at the station. Finally, the LSTM model with time series input form is shown in Fig. 4. The arrival delay time($Y_i^{s_n}$)of the current train is predicted based on the feature space ($F_i^{s_n}$) of the current train and the effect of only the previous train ($F_{i-1}^{s_n}$). The aforementioned step is repeated to finally realize the prediction for all stations from Rotterdam Central to Dordrecht section.

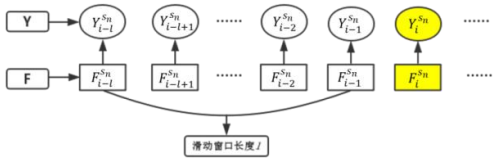


Figure 3: LSTM input data format

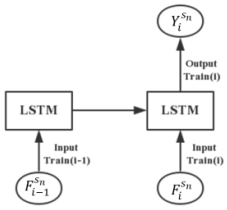


Figure 4: LSTM prediction model

5 Precision and Evaluation of the Model Prediction

5.1 Model prediction accuracy analysis

In order to evaluate the prediction effect of the model, the following analysis is initially performed. As shown in Fig. 5, the actual and predicted arrival delays of trains at stations are compared. Second, as shown in Figure 6, the scatter plots of the observed and predicted arrival delays of trains are illustrated. The results indicate that the predicted values of train arrival delays exhibit a good match with the observed values. Specifically, in the interquartile range, the whiskers and right tail closely match in the figures for each station. Furthermore, as shown in in Fig. 6, the majority of predictions are close to the depicted diagonal lines for arrival events ,which implies that the predicted value is extremely close to the observed value.

Figure 7 shows the distribution of predicted residuals for train arrival delays at different stations. The figure assumes the train actual arrival delay time as abscissa and the residuals as ordinate for visualization purposes. As shown in the figure, in the seven stations of Rotterdam Central to Dordrecht section of the Netherlands railway system, all stations (with the exception of the Rtd station) exhibit good prediction results. Increases in the prediction error of the Rtd station can be due to the increasingly significant influence of the outliers. Figure 8 shows the prediction accuracy histogram of LSTM model for the seven stations. As shown in the figure, the model accuracy corresponds to 87.6% with an allowable error within 30 s, and thus the model exhibits a good prediction effect.

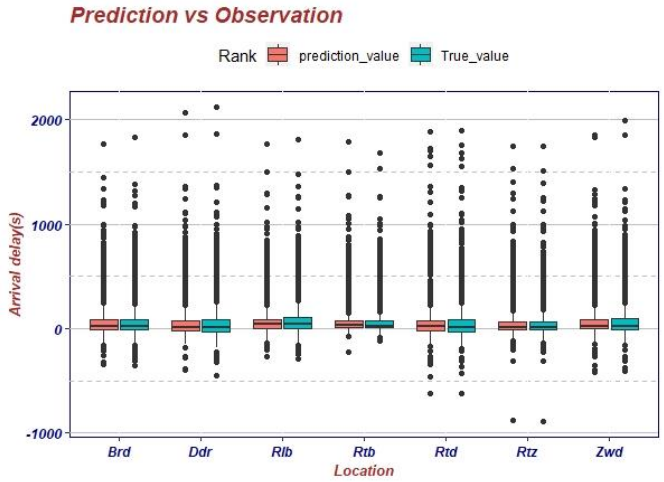


Figure 5: Comparison of predicted and observed arrival delay distribution for different stations

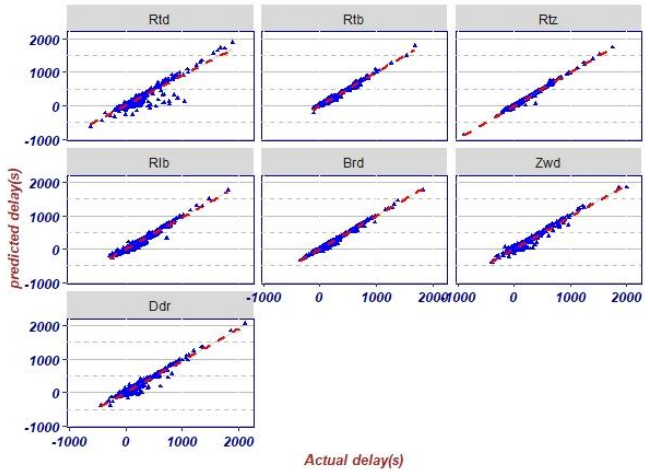


Figure 6: Scatter plots of actual vis-á-vis predicted arrival delays.

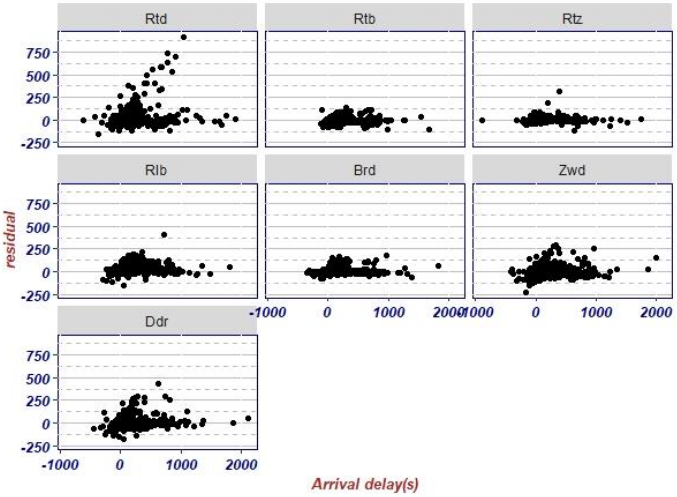


Figure 7: Distribution of the residuals of train arrival delays at different stations

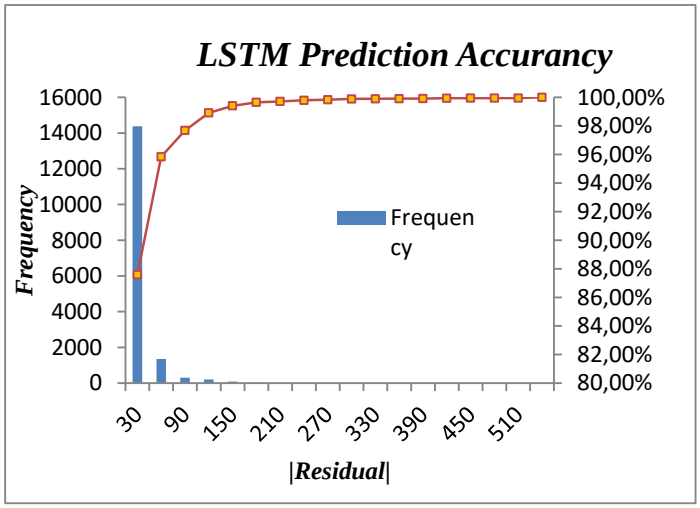


Figure 8: LSTM model prediction accuracy

5.2 Model evaluation

(1) Benchmark model

In order to better evaluate the prediction effect of the model, two benchmark models are selected and compared with the LSTM model, namely the random forest model and artificial neural network model. They are detailed as follows:

Random forest: The random forest is a joint prediction model that is composed of multiple decision trees (Cutler et al. (2004),Loh (2011)), and this can be used as a fast and effective classification and prediction model. Each decision tree in RF consists of several forks and nodes. Each decision tree is regressed and predicted. Finally, the predictive effect of random forest is determined via the predictive effect of multiple decision trees. The random forest corresponds to an ensemble learning algorithm, which belongs to the Bagging type. The final result is voted or averaged by combining multiple weak classifiers, and thus the overall model results exhibit higher accuracy and generalization performance. Thus, the model obtains good results, and this is mainly due to the "random" and "forest" elements, which make it resistant to overfitting and increase the precision.

Artificial neural networks: An artificial neural network is one of the most commonly used train delay prediction model (Peters et al. (2005),Yaghini et al. (2013) ,Malavasi (2001)). It mainly models the relationship between a set of input signals and set of output signals. The model is derived from the reaction of the human brain to stimuli from a sensory input. In a manner similar to how the brain uses a network of interconnected cells

of a neuron to create a large parallel processor, artificial neural networks use artificial neurons or a network of nodes to solve learning problems. There are three main characteristics of artificial neural networks as follows: ① Activation function that converts the net input signal of a neuron into a single output signal for further propagation in the network; ②network topology that describes the number of neurons in the model, number of layers, and the manner in which the layers are connected; and ③training algorithm that specifies the setting of the connection weight to suppress or increase the proportion of neurons in the input signal. This model is suitable for situations involving simple input and output data albeit an extremely complex input-to-output process.

(2) Model evaluation index

With respect to model evaluation, the study mainly selects MAE and RMSE as evaluation indexes. The equation to calculate the index is as follows:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|p_i - y_i|. \tag{7}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(p_i - y_i)^2}. \tag{8}$$

where p_i and $y_i$denote the predicted and observed delay values for *ith* arrival events, respectively, and *n* denotes the total number of observations. The measures quantify the average deviation of the predictions from the observed values. The model's performance level improves when the measures are closer to zero.

Figure 9 and Figure 10 show a comparison of MAE and RMSE values for LSTM, RF, and ANN models of different stations. As shown in Fig. 9, from the MAE perspective, the prediction effects of the LSTM model and the random forest model do not significantly differ and both are superior to the artificial neural network model. As shown in Fig. 10, from the RMSE perspective, the prediction effect of LSTM significantly exceeds that of random forest and artificial neural network models. In summary, the LSTM model exhibits a good predictive effect.

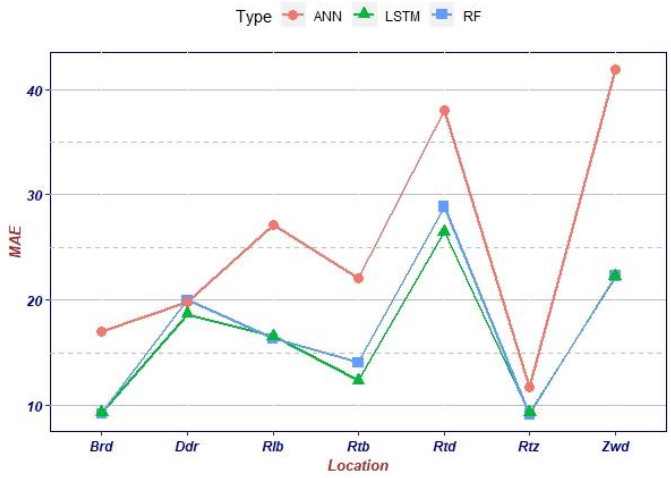


Figure 9: Comparison of MAE values at different stations

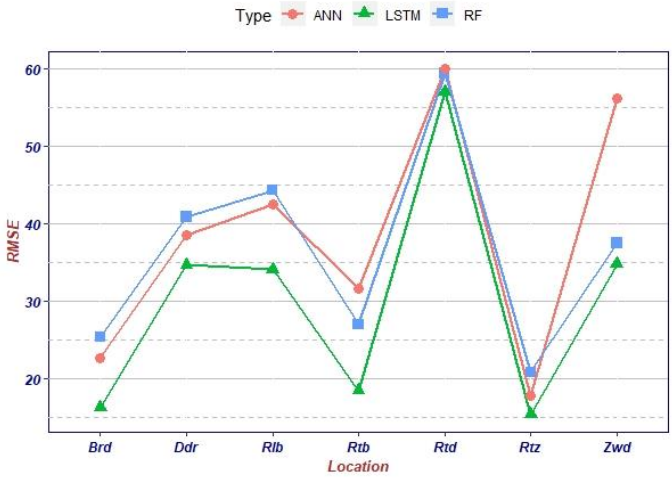


Figure 10: Comparison of RMSE values for different stations

6 Conclusions

The study presents a machine learning model to analyze the relationship between train

arrival delays and various characteristics of a railway system, which is important for planning changes and investments to reduce delays. In the study, the LSTM model is used to construct a prediction model of train arrival delay, and the model is trained and tested based on the historical data of train operation. The results show that the LSTM model exhibits a better predictive effect than random forest and artificial neural network models. The performance of the LSTM model is superior as indicated by the data validation results. Specifically, the LSTM model exhibits better MAE and MSE values, and its prediction accuracy reaches 87% within 30 s.

The LSTM model is a good measure of the lateral and vertical propagation of train delays. This feature ensures that the model exhibits good generality and can be extended to other high-speed railway routes. Additionally, the model exhibits two main advantages as follows: (a) The simplicity of the model makes it more explanatory and efficient. (b) It includes interrelationships between various delay factors and superposition of arrival delays.

The model in the study can be applied to other stations although similar data must be collected. With respect to the expansion direction of the model, the current model does not consider an excessive number of infrastructure factors. With respect to further model expansion, it is possible to consider additional train delay influence factors and extract increasingly accurate feature variables to obtain better prediction results.

Acknowledgment

This work was supported by the National Nature Science Foundation of China [grant number 71871188] and the Science & Technology Department of Sichuan Province [grant number 2018JY0567]. We are grateful for the contributions made by our project partners.

References

Barta, J., Rizzoli, A.E., Salani, M., Gambardella, L.M.,2012.Statistical modelling of delays in a rail freight transportation network. *Wint Simul C Proc.* 2012.

Bengio, Y., Simard, P., Frasconi, P., 2002.Learning long-term dependencies with gradient descent is difficult. *IEEE Trans Neural Netw* 5(2):157-166.

Berger A., Gebhardt, A., Müller-Hannemann, M., Ostrowski, M.,2011.Stochastic delay prediction in large train networks. In: OASIcs-OpenAccess Series in Informatics (Vol. 20). Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Chen, D., Wang, L.,Li, L.,2015.Position computation models for high-speed train based on support vector machine approach. *Applied Soft Computing* 30(C):758-766.

Cutler, A., Cutler, D.R.,Stevens, J.R.,2004.Random Forests. *Machine Learning* 45(1):157-176.

Francesco, C.,Pavle, K.,2018.Stochastic prediction of train delays in real-time using Bayesian networks. *Transportation Research Part C: Emerging Technologies.*

Flier, H., Graffagnino, T., Nunkesser, M.,2009. Finding Robust Train Paths in Dense Corridors. IAROR RailZurich 2009.

Gers, F.A., Schraudolph, N.N.,Schmidhuber, J.,2002.Learning Precise Timing with LSTM Recurrent Networks. *Journal of Machine Learning Research* 3(1):115-143.

Gorman, M.F.,2009.Statistical estimation of railroad congestion delay. *Transportation Research Part E Logistics & Transportation Review* 45(3):446-456.

Goverde, R.M.P., Corman, F.,D ' Ariano, A.,2013.Railway line capacity consumption of different railway signalling systems under scheduled and disturbed conditions. *Journal of Rail Transport Planning & Management* 3(3):78-94.

Greff, K., et al.,2016.LSTM: A Search Space Odyssey. *IEEE Transactions on Neural Networks & Learning Systems* 28(10):2222-2232.

Huisman, T., Boucherieb, R.J.,Dijka, N.M.V.,2002.A solvable queueing network model for railway networks and its validation and applications for the Netherlands. *European Journal of Operational Research* 142(1):30-51.

Kecman, P., Corman, F.,Meng, L.,2015.Train delay evolution as a stochastic process. International Conference on Railway Operations Modelling & Analysis-railtokyo, 2015.

Kecman, P.,Goverde, R.M.P.,2015a.Online data-driven adaptive prediction of train event times. *IEEE Transactions on Intelligent Transportation Systems* 16(1):465-474.

Kecman, P.,Goverde, R.M.P.,2015b.Predictive modelling of running and dwell times in railway traffic. *Public Transport* 7(3):1-25.

Lessan, J., Fu, L.,Wen, C.,2018.A Hybrid Bayesian Network Model for Predicting Delays in Train Operations. *Computers & Industrial Engineering*.

Loh, W.Y.,2011.Classification and regression trees. *Wiley Interdisciplinary Reviews Data Mining & Knowledge Discovery* 1(1):14-23.

Malavasi, G.,2001.Simulation of stochastic elements in railway systems using self-learning processes. *European Journal of Operational Research* 131(2):262-272.

Marković, N., Milinković, S., Tikhonov, K.S., Schonfeld, P.,2015.Analyzing passenger train arrival delays with support vector regression. *Transportation Research Part C: Emerging Technologies* 56:251-262.

Medeossi, G., Longo, G.,Fabris, S.D.,2011.A method for using stochastic blocking times to improve timetable planning. *Journal of Rail Transport Planning & Management* 1(1):1-13.

Meer Van der, D.J., Goverde, R.M.P.,Hansen, I.A.,2009.Prediction of train running times using historical track occupation data. *Delft University of Technology*, 2009..

Meester, L.E.,Muns, S.,2007.Stochastic delay propagation in railway networks and phase-type distributions. *Transportation Research Part B Methodological* 41(2):218-230.

Olsson, N.O.E.,Haugland, H.,2004.Influencing factors on train punctuality—results from some Norwegian studies. *Transport Policy* 11(4):387-397.

Oneto, L., Fumeo, E., Clerico, G., Canepa, R., Papa, F., Dambra, C., Mazzino, N., Anguita, D.,2016.Advanced Analytics for Train Delay Prediction Systems by Including Exogenous Weather Data. *IEEE International Conference on Data Science & Advanced Analytics*, 2016. *IEEE, Montreal, Canada*, pp. 458-467.

Oneto, L., Fumeo, E., Clerico, G., Canepa, R., Papa, F., Dambra, C., Mazzino, N., Anguita, D.,2017a.Dynamic Delay Predictions for Large-Scale Railway Networks: Deep and Shallow Extreme Learning Machines Tuned via Thresholdout. *IEEE Transactions on Systems Man & Cybernetics Systems* PP(99):1-14.

Oneto, L., Fumeo, E., Clerico, G., Canepa, R., Papa, F., Dambra, C., Mazzino, N., Anguita, D.,2017b.Train Delay Prediction Systems: A Big Data Analytics Perspective. *Big Data Research*11, 54-64

Peters, J., Emig, B., Jung, M., Schmidt, S.,2005.Prediction of Delays in Public Transportation using Neural Networks. *Computational Intelligence for Modelling, Control and Automation*, 2005 and *International Conference on Intelligent Agents, Web Technologies and Internet Commerce, International Conference on*, 2005, pp. 92-97.

Şahin, İ.,2017.Markov chain model for delay distribution in train schedules: Assessing the effectiveness of time allowances. *Journal of Rail Transport Planning & Management* 7(3).

Yaghini, M., Khoshraftar, M.M.,Seyedabadi, M.,2013.Railway passenger train delay prediction via neural network model. *Journal of Advanced Transportation* 47(3):355-368.

Yuan, J.,Hansen, I.A.,2007.Optimizing capacity utilization of stations by estimating knock-on train delays. *Transportation Research Part B Methodological* 41(2):202-217.

Real-time Train Platforming and Routing at Busy Complex High-speed Railway Stations

Jia Ning [a,1], Qiyuan Peng [a,2], Gongyuan Lu [a,3]
[a] School of Transportation and Logistics, Southwest Jiaotong University, Chengdu, China
[1] E-mail: ningjia@my.swjtu.edu.cn
[2] E-mail: qiyuan-peng@home.swjtu.edu.cn
[3] E-mail: lugongyuan@home.swjtu.edu.cn, Phone: +0086-13880609100

Abstract
This paper focuses on the real-time train platforming and routing problem at a busy complex high-speed railway station in a disrupted situation caused by malfunctioning railway infrastructure or train primary delay. When the disruption occurs, dispatchers need to reassign trains to the conflict-free platform and route, even reschedule the arrival and departure times. To this end, we develop a mixed-integer linear programming formulation that determines platforming and routing decision simultaneously, while allowing trains to be rescheduled when the initial schedule imposes irreconcilable conflicts. The objective of our model is to minimize the overall deviation from the planned schedule and the original platforming plans. To improve the solving efficiency, an iterative algorithm is proposed to compute near-optimal solutions in a short computation time, which is based on the decomposition of the overall problem into two sub-problems: (i) platform and route assignment with fixed arrival and departure times, (ii) partial conflict trains rescheduling. The connecting information between two sub-problems concerns the index of conflict trains and the new train timetable. To solve sub-problem (i) efficiently, we develop a branch and bound algorithm which includes implicational rules enabling to speed up the computation and still can acquire optimal solutions. Since the model of sub-problem (ii) is the same as the model of original problem with a relative small scale, it can be solved by CPLEX solver efficiently. A real-world instance with operation data of Zhengzhou East high-speed railway station, is implemented to demonstrate the performance and effectiveness of the proposed algorithm.

Keywords
Train platforming and routing problem, Real-time conflict resolution, Linear programming, Branch and bound algorithm

1 Introduction

In this paper, we focus on an important operational problem in the railway industry, namely real-time train platforming and routing at a single large busy complex high-speed railway station (rtTPR). This problem is one variant of the more general problem of routing trains through railway stations considered by Zwaneveld et al. (2001).

Within a high-speed railway system, the platform, the route, as well as the arrival and departure time of each train are set in the dispatching system in advance, then trains run as scheduled. When the disruption occurs which can be caused by malfunctioning railway infrastructure or train primary delay, dispatchers need to reassign trains to the non-conflicting platforms and routes, even reschedule the arrival and departure times when the

initial schedule imposes irreconcilable conflicts.

Large stations, such as hub stations, are usually located at the intersection of multiple railway lines. Such stations typically have multiple entering/existing points, dozens of platforms and hundreds of inbound/outbound routes, with several hundred or over a thousand trains per day (Carey and Carville, 2003). Dispatchers should coordinate the station operations of trains from different entering/existing directions every day. When a disruption occurs, a high-quality platform reallocation plan can absorb train delay to some extent. This process requires effective solution within minutes, which is difficult for the dispatchers.

Most research on train dispatching is concerned with rescheduling trains on lines, usually single lines, taking into account the microscopic layout of each station on the railway line. But they are not concerned with large stations with multiple entering/existing points, or do not consider detailed route occupancy and release (for example, D'Ariano et al. 2008, Lamorgese and Mannino, 2015). However, in Europe and China, large busy complex high-speed railway stations are key components of high-speed railway networks, and are usually the locations of most train conflicts (Carey and Carville, 2003). Since the rtTPR is of great significance to reduce the impact of the disruptions on train operation and station working order, we focus on this problem in the present paper. And the optimization results can also provide suggestions for line dispatching.

The train platforming and routing at a single railway station (TPR) is an important planning problem in the railway industry and arises at each of the strategic, tactical and operational planning levels (Sels et al., 2014). While the strategic and tactical level variants, which address future station capacity and the generation of feasible timetable, respectively, have been well studied, the operational variant has received relatively limited attention in the literature. Lusby et al. (2011) survey a large number of papers in the field of routing trains through railway junctions at the strategic, tactical and operational levels. Cacchiani et al. (2014) present an overview of literatures on real-time timetable rescheduling in case of disturbances or disruptions, considering a microscopic or a macroscopic view on the railway system. We refer the interested reader to Lusby et al. (2011) and Cacchiani et al. (2014). Here we review contributions in the area of real-time train platforming and routing at a single station. Without providing an exhaustive review, we attempt to provide an overview of the models and methods that have been adopted in the literature.

Zwaneveld et al. (1996) address the problem of routing trains through railway stations at the strategic level. In this paper the authors propose a node-packing formulation which allows time deviations on the arrival and departure times of train paths. Furthermore, they present a branch-and-cut approach combined with reduction techniques to solve the problem. However, in a follow-up paper, Zwaneveld et al. (2001) state that this approach was unable to solve the routing problem for two of the larger stations in Netherlands. Zwaneveld et al. (2001) is an extension of Zwaneveld et al. (1996). In this paper the problem is formulated as a weighted node-packing model. More sophisticated preprocessing and reduction techniques are included in the branch-and-cut solution approach, and the authors conclude that this approach is sufficient for solving the routing problem arising at any Dutch railway station, but the efficiency of this approach cannot be applied to the operational level.

Carey and Carville (2003) address the TPR at the tactical level. They develop scheduling heuristics analogous to those successfully adopted by dispatchers using "manual'' methods. The algorithm considers one train at a time and finds and resolves all conflicts for that train before considering the next train. When considering a train, the

algorithm considers assigning it to each platform in turn, to find the best platform. With successive refinements, the algorithm eventually takes only a few seconds to run.

Constraint programming has also been used to model the TPR. Rodriguez and Kermad (1998) and Rodriguez (2007) are an attempt to model the operational variant of the problem. This approach attempts to find the minimum total delay necessary in keeping the trains on their assigned paths. Instances with between 6 and 24 trains are considered. However, the problem considered in this paper permits trains to wait on track sections if the subsequent track section is unavailable, which is not common in China and is not allowed in our paper.

D'Ariano et al. (2007, 2008) propose an alternative graph formulation for the operational variant of the TPR problem. A pair of alternative arcs is used to enforce a train sequencing order at any block section where two trains are in conflict. An alternative graph is built using one path per train. The model may include hundreds of machined (block sections) and jobs (trains) for real-life instances, and is therefore very hard to be solved in real-time. To overcome this issue, D'Ariano et al. (2007) propose a branch-and-bound algorithm which includes dynamic and static implication rules enabling to speed up the computation. This algorithm is extended by D'Ariano et al. (2008) to include a local re-routing strategy. The iterative procedure first computes an optimal train sequencing for given train routes and then improves this solution by locally rerouting some trains.

Caimi et al. (2012) propose a closed-loop discrete-time control framework for the TPR at the operational level. This framework resolve conflicts by re-timing and re-routing of trains as well as partial speed profile coordination. In this approach the time horizon is discretized, and a binary variable is associated with every operation and every period in the time horizon. Computational experiments indicate clearly the great potential of this approach.

Lusby et al. (2011) address the strategic-level variant of the TPR and present a set-packing model. A resource based constraint system is used, in which resources correspond to track sections. Then the authors prove that this formulation is tighter than the conventional node-packing model and develop a branch-and-price algorithm that utilizes the dual representation of any basic feasible solution. Lusby et al. (2013) extend this method and apply it to the operational level. In this paper, the authors develop a branch-and-price approach that exploits the flexibility of the model to be dynamically updated. Numerical results indicate the efficiency of this approach by confirming that, with a given time limit of 270 seconds, practical problems can be solved within 3.5 percent of optimality.

Caprara et al. (2011) deal with a general version of the TPR problem. Each train to be assigned a platform is assumed to have a number of possible patterns consisting of an inbound path, outbound path, and platform, as well as arrival and departure times at the platform. The authors present an integer linear programming formulation that is based on a node-packing problem. This model has a continuous relaxation that leads to strong bounds on the optimal value. A branch-and-cut-and-price solution approach based on the linear programming relaxation is proposed in this paper.

Boccia et al. (2013) present a new mixed integer programming model to tackle the real-time railway traffic management problem. A set of routes for each train is considered and tracks are subdivided into sections. The model uses binary variables indicating whether a route in the set is assigned to a train, and continuous variables representing the time at which a train reaches a block section. Two heuristic algorithms are proposed.

The same problem is considered by Meng and Zhou (2014). They propose a cumulative flow variables-based model based on a time-space network modelling

framework. A Lagrangian relaxation solution framework is developed. Then the original problem is decomposed into a sequence of single train optimization sub-problems. For each sub-problem, a dynamic programming algorithm is proposed to find the time dependent least cost path on a time-space network.

The similar problem is considered by Pellegrini et al. (2015, 2019). Pellegrini et al. (2015) present a mixed-integer linear programming formulation, which models the infrastructure in terms of track-circuits and the route-lock sectional release interlocking system. The model calls for assigning a route to each train, as well as a possible delay for the train on each track circuit. Since the difficulty in solving the MILP model is mostly due to the multiplicity of both the alternative routes and the potential conflicts. Pellegrini et al. (2015, 2019) propose valid inequalities to boost the solution algorithm. In addition, Pellegrini et al. (2019) reformulate this model based on a reduced number of scheduling binary variables.

Sama et al. (2016) deal with the real-time train rerouting and rescheduling in the railway network. This paper studies the problem of selecting the best subset of routing alternatives for each train among all possible alternatives, which is formulated as an integer linear programming formulation and solved via an algorithm inspired by the ant colonies' behaviour. Then, the real-time train rerouting and rescheduling problem takes as input the best subset of routing alternatives and is solved as a mixed-integer linear program.

In this paper, we focus on the rtTPR. The *Degree of Conflict* is defined to describe the conflict between any inbound/outbound routes. A bi-objective mixed integer linear programming model is formulated to determine platform, inbound and outbound route, and arrival and departure times simultaneously, which is also a universal model for the route-lock sectional/integral release interlocking system. Similar to the normal practice of railway dispatchers, this model is decomposed into two sub-models and an iterative algorithm which combines a branch-and-bound algorithm and CPLEX solver is developed. Finally, a real-world instance of Zhengzhou East high-speed railway station in China is tested.

2 Problem Description

A railway station consists of platforms and of a large number of track sections. Trains enter a railway station from *entering points* and leave it through *leaving points*. An *inbound route* is a sequence of track sections linking an entering point to a platform; while an *outbound route* is a sequence of track sections linking a platform to a leaving point. Notably, there may be more than one inbound/outbound route linking the same entering/leaving point and the same platform. A *path* is composed of an inbound and an outbound route linking the same platform and the linked platform, and it is a sequence of track sections connecting an entering point to a leaving point. There are generally multiple different paths between a given pair of entering and leaving points.

As soon as a train arrives at its entering point of the station, it claims an inbound route to a platform. At the same time, the linked platform is also claimed. Similarly, before a train leaves its platform, it claims an outbound route to its leaving point. Moreover, as a train traverses its inbound/outbound route, it sequentially releases each of the track sections comprising the route. Since any track section can only be claimed by at most one train at any time, a *conflict* will occur if two chosen routes simultaneously attempt to claim the same track sections. Thus, the exact calculation of the time at which the common sections are released by the previous route is the key to rule out the conflict

between any two routes. To this end, Degree of Conflict (DOC) is defined to describe the conflict relations between routes.

Definition1. (Degree of Conflict) Given a route pair $r - r'$, if route r is claimed ahead of route r', and route r and r' have common track sections, DOC between the route pair $r - r'$, denoted as $\gamma_{r,r'}$, indicates the time elapsed from the time when route r starts to be claimed until the common sections are all released by route r (i.e., the time at which route r' can be claimed).

As shown in Figure 1, the sequenced sections list of route r are recorded as {s_9, s_8, s_7, s_4, s_3, s_2, s_5, s_6}, and the sequenced sections list of route r' are recorded as {s_1, s_2, s_3, s_4, s_{10}}. The common sections of route r and r' are recorded as {s_2, s_3, s_4}. If route r is claimed ahead of route r', then $\gamma_{r,r'}$ is equal to the duration of traversing sections {s_9, s_8, s_7, s_4, s_3, s_2} sequentially. Whereas, if route r' is claimed ahead of route r, then $\gamma_{r',r}$ is equal to the duration of traversing sections {s_1, s_2, s_3, s_4} sequentially. Therefore, the DOC between routes is asymmetry (i.e., $\gamma_{r,r'} \neq \gamma_{r',r}$).

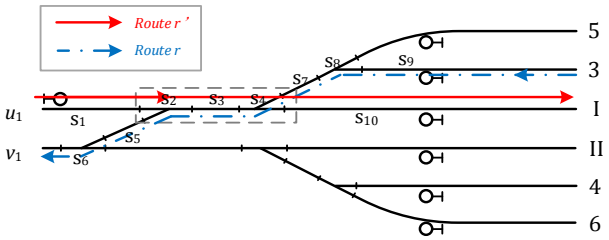


Figure 1: Illustration of Degree of Conflict.

Assume that all trains travel at the same constant speed regardless of route choice. DOC between each route pair can be calculated according to the sequenced sections list of each route, the common sections list, the length of each section and the speed of trains.

In addition to conflicts between routes, a conflict also arises whenever two or more trains require the same platform at the same time. To rule out platform conflicts, safety time interval is imposed between two adjacent trains occupying the same platform.

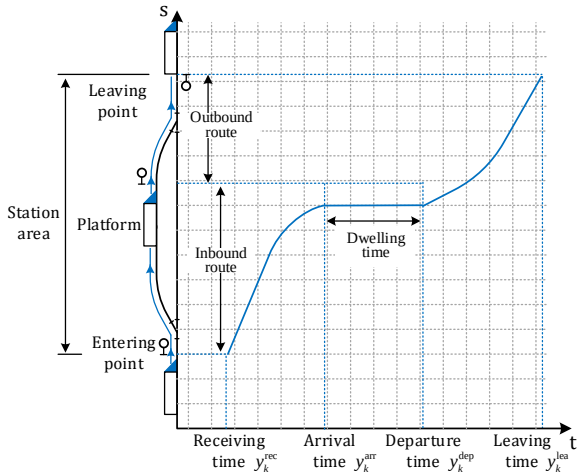


Figure 2: Time definition in train operation process.

For recording traveling process in railway station of each train, following variables are defined:

1. **Receiving time $y_k^{\mathbf{rec}}$:** The receiving time of train k is the time at which the train arrives at its entering point of the station and its inbound route and platform have just been claimed.
2. **Arrival time $y_k^{\mathbf{arr}}$:** The arrival time of train k is the time at which the train stops at a platform, after traveling along an inbound route.
3. **Departure time $y_k^{\mathbf{dep}}$:** The departure time of train k is the time at which the train begins to leave the platform along an outbound route.
4. **Leaving time $y_k^{\mathbf{lea}}$:** The leaving time of train k is the time at which the tail of the train releases the last section of its outbound route and leaves the station from its leaving point.

Notably, if a train does not stop at a railway station, which is called through train for the station, when it arrives at its entering point, its outbound route should be claimed with its inbound route and its platform. Thus, as for the through train, its receiving time indicates the time when the train arrives at its entering point of the station and its inbound and outbound route and platform have all been claimed. The arrival time of a through train indicates the time when the train has released the last section of its inbound route. And the departure time of a through train is equal to its arrival time. In addition, the definition of the leaving time of a through train is same as that of other trains.

Note that above 4 types of time definition correspond to the operation process of a train in a railway station, as illustrated in Figure 2, so relation can be modelled between these time variables. The arrival time is equal to the receiving time plus the time required by a train to release its inbound route completely (including the time for creating the inbound route, driver response time, the time for the train to approach the protection signal, the travel time of the train via the route, and the time for the train to clear and release the last section of the inbound route). Similarly, the leaving time is equal to the departure time plus the time required by a train to release its outbound route completely.

3 Mathematical Formulation

3.1 Assumptions

Firstly, the following assumptions are made throughout this paper.

Assumption 1. All trains travel at the same constant speed regardless of route choice. The DOC between each route pair and the traversal time of each route are pre-given.

Assumption 2. The ideal arrival and departure times of the trains derive from the planned scheduled at the tactical level. The original platform allocation plan and route assignment plan are per-given. The initial train delay information and track maintenance information are also pre-given.

3.2 Definitions

The sets, parameters and decision variables used in this paper are described in the Tables 1 and 2, respectively.

Table 1: Definition of sets and parameters.

Symbol	Definition
K	Set of trains, index by k and h, i.e., $k, h \in K$.
T	Set of platforms.
R	Set of routes, index by r and r', i.e., $r, r' \in R$.
R_k	Set of train paths that train k potentially travel through, index by i and j, i.e., $i, j \in R_k$. Each path i is composed of three parts: inbound route $r_{k,i}^{\text{in}}$, platform $r_{k,i}$ and outbound route $r_{k,i}^{\text{out}}$, and $r_{k,i} \in T$, $r_{k,i}^{\text{in}}, r_{k,i}^{\text{out}} \in R$.
$\gamma_{r,r'}$	The Degree of Conflict between route r and route r'.
w_k	0-1 train parameter, equal to 1 if train k is a through train, 0 otherwise.
τ_k^{arr}	Ideal arrival time of train k.
τ_k^{dep}	Ideal departure time of train k.
Δ_k	Minimum dwelling time of train k.
$c_{k,i}$	Weight of train k's path i.
$tt_{k,i}^{\text{in}}$	Traversal time of inbound route $r_{k,i}^{\text{in}}$.
$tt_{k,i}^{\text{out}}$	Traversal time of inbound route $r_{k,i}^{\text{out}}$.
$\theta_{k,i}$	0-1 availability parameter, equal to 1 if platform $r_{k,i}$ is available during the whole considered time horizon, 0 otherwise.
$t_{k,i}^{\text{start}}$	Start time of the maintenance on platform $r_{k,i}$.
$t_{k,i}^{\text{end}}$	End time of the maintenance on platform $r_{k,i}$.
H_{start}	Start time of considered time horizon.
H_{end}	End time of considered time horizon.
ε	The shortest time interval permitted between two adjacent trains occupying the same platform.
M	A sufficiently large positive number.

Table 2: Definition of variables.

Symbol	Definition
$x_{k,i}$	0-1 path selection variable, equal to 1 if train k chooses path i, 0 otherwise.
y_k^{arr}	Actual arrival time of train k.
y_k^{dep}	Actual departure time of train k.
y_k^{rec}	Actual receiving time of train k.
y_k^{lea}	Actual leaving time of train k.
$B_{k,i}$	0-1 occupancy-maintenance sequence variable, equal to 1 if train k claiming platform $r_{k,i}$ precedes the maintenance on platform $r_{k,i}$, 0 otherwise.
$B_{k,h}^{\text{p}}$	0-1 platform occupancy sequence variable, equal to 0 if train k claiming its platform precedes train h, 1 otherwise.
$B_{k,h}^{\text{in-in}}$	0-1 inbound-inbound route occupancy sequence variable, equal to 0 if train k claiming its inbound route precedes train h claiming its inbound route, 1 otherwise.
$B_{k,h}^{\text{in-out}}$	0-1 inbound-outbound route occupancy sequence variable, equal to 0 if train k claiming its inbound route precedes train h claiming its outbound route, 1 otherwise.
$B_{k,h}^{\text{out-in}}$	0-1 outbound-inbound route occupancy sequence variable, equal to 0 if train k claiming its outbound route precedes train h claiming its inbound route, 1 otherwise.
$B_{k,h}^{\text{out-out}}$	0-1 outbound-outbound route occupancy sequence variable, equal to 0 if train k claiming its outbound route precedes train h claiming its outbound route, 1 otherwise.

3.3 Mathematical Model

The mathematical model for real-time Train Platforming and Routing problem is given in

the following.

Objective function

On one hand, the reallocation of platform and the reassignment of route will influence regular station working order. Hence, the first objective maximizes the preferences of the adjusted platform allocation plan in order to minimize the impact on regular station working order. This objective function can be formulated as follows:

$$\max Z_1 = \sum_{k\in K}\sum_{i\in R_k} c_{k,i}x_{k,i} \tag{1}$$

Here $c_{k,i}$ reflects the impact of choosing path i for train k. A higher value of $c_{k,i}$ indicated a smaller impact on regular station working order.

On the other hand, in order to prevent train delay from propagating through the network, the second objective minimizes the overall deviation from the ideal arrival and departure times for each train. This objective function can be functioned as follows:

$$\min Z_2 = \sum_{k\in K}(y_k^{\mathrm{arr}} - \tau_k^{\mathrm{arr}}) + \sum_{k\in K}(y_k^{\mathrm{dep}} - \tau_k^{\mathrm{dep}}) \tag{2}$$

Constraints

(1) Time relation constraints.

The time variables of the model include actual receiving time, actual arrival time, actual departure time and actual leaving time. Constraints (3) specify that the actual arrival time of train k is equal to its actual receiving time plus the time required to release its inbound route completely. Similarly, constraints (4) specify that the actual leaving time of train k is equal to its actual departure time plus the time required to release its outbound route completely.

$$y_k^{\mathrm{arr}} = y_k^{\mathrm{rec}} + \sum_{i\in R_k} tt_{k,i}^{\mathrm{in}} x_{k,i} \qquad \forall k \in K \tag{3}$$

$$y_k^{\mathrm{lea}} = y_k^{\mathrm{dep}} + \sum_{i\in R_k} tt_{k,i}^{\mathrm{out}} x_{k,i} \qquad \forall k \in K \tag{4}$$

(2) Platform conflict free constraints

Each platform can be occupied by at most one train at any time. Whether it is a through train or a non-through train, its platform is claimed at its receiving time and is released at its departure time. Constraints (5) and (6) impose the safety time interval between two adjacent trains k and h on the same platform. One of the constraints (5) and (6) will be activated if both the variables $x_{k,i}$ and $x_{h,j}$ are equal to 1 and platform $r_{k,i}$ and $r_{h,j}$ refer to the same platform. The activation of a constraint (5) means train k precedes train h and thus $B_{k,h}^{\mathrm{p}}$ is equal to 0. In this case, the time when train h claims its allocated platform minus the time at which train k releases its allocated platform must be larger than or equal to the shortest time interval permitted between two adjacent trains occupying the same platform. Constraints (6) have a similar function of constraints (5) except that train h precedes train k.

$$\mathrm{M}\left(2 + B_{k,h}^{\mathrm{p}} - x_{k,i} - x_{h,j}\right) + y_h^{\mathrm{rec}} - y_k^{\mathrm{dep}} \geq \varepsilon \qquad \forall k \in K, \forall h \in K, \forall i \in R_k, \forall j \in R_h : h > k, r_{k,i} = r_{h,j} \tag{5}$$

$$\mathrm{M}\left(3 - B_{k,h}^{\mathrm{p}} - x_{k,i} - x_{h,j}\right) + y_k^{\mathrm{rec}} - y_h^{\mathrm{dep}} \geq \varepsilon \qquad \forall k \in K, \forall h \in K, \forall i \in R_k, \forall j \in R_h : h > k, r_{k,i} = r_{h,j} \tag{6}$$

(3) Route conflict free constraints

Whether it is a through train or a non-through train, its inbound route is claimed at its receiving time and is released at its arrival time. However, as for the through train, its outbound route is claimed at its receiving time and is released at its leaving time; while as for the non-through train, its outbound route is claimed at its departure time and is

released at its leaving time.

The DOC of a given route pair indicates the duration time from the time when the previous route is claimed to the time when the latter route can be claimed. Constraints (7)-(14) impose minimum required headway times (i.e., the Degree of Conflict) between any two trains on two conflicting routes. It is worth noting that if a train is a through train, its outbound route conflicts with a route chosen by the other train, and the train claiming its outbound route precedes its conflict route, then the minimum required headway time is equal to the DOC between the two conflict routes plus the time required to release the through train's inbound route. Constraints (7) and (8) deal with conflicts between inbound routes. Constraints (9)-(12) deal with conflicts between inbound routes and outbound routes. Constraints (13) and (14) deal with conflicts between outbound routes.

$$\mathrm{M}\left(2 + B_{k,h}^{\mathrm{in-in}} - x_{k,i} - x_{h,j}\right) + y_h^{\mathrm{rec}} - y_k^{\mathrm{rec}} \geq \gamma_{r_{k,i}^{\mathrm{in}}, r_{h,j}^{\mathrm{in}}} \quad \forall k \in K, \forall h \in K, \forall i \in R_k, \forall j \in R_h : h > k, \gamma_{r_{k,i}^{\mathrm{in}}, r_{h,j}^{\mathrm{in}}} \neq 0 \quad (7)$$

$$\mathrm{M}\left(3 - B_{k,h}^{\mathrm{in-in}} - x_{k,i} - x_{h,j}\right) + y_k^{\mathrm{rec}} - y_h^{\mathrm{rec}} \geq \gamma_{r_{h,j}^{\mathrm{in}}, r_{k,i}^{\mathrm{in}}} \quad \forall k \in K, \forall h \in K, \forall i \in R_k, \forall j \in R_h : h > k, \gamma_{r_{h,j}^{\mathrm{in}}, r_{k,i}^{\mathrm{in}}} \neq 0 \quad (8)$$

$$\mathrm{M}\left(2 + B_{k,h}^{\mathrm{in-out}} - x_{k,i} - x_{h,j}\right) + w_h y_h^{\mathrm{rec}} + (1 - w_h) y_h^{\mathrm{dep}} - y_k^{\mathrm{rec}} \geq \gamma_{r_{k,i}^{\mathrm{in}}, r_{h,j}^{\mathrm{out}}} \quad \forall k \in K, \forall h \in K, \forall i \in R_k, \forall j \in R_h : h > k, \gamma_{r_{k,i}^{\mathrm{in}}, r_{h,j}^{\mathrm{out}}} \neq 0 \quad (9)$$

$$\mathrm{M}\left(3 - B_{k,h}^{\mathrm{in-out}} - x_{k,i} - x_{h,j}\right) + y_k^{\mathrm{rec}} - w_h y_h^{\mathrm{rec}} - (1 - w_h) y_h^{\mathrm{dep}} \geq \gamma_{r_{h,j}^{\mathrm{out}}, r_{k,i}^{\mathrm{in}}} + w_h t t_{h,j}^{\mathrm{in}} \quad \forall k \in K, \forall h \in K, \forall i \in R_k, \forall j \in R_h : h > k, \gamma_{r_{h,j}^{\mathrm{out}}, r_{k,i}^{\mathrm{in}}} \neq 0 \quad (10)$$

$$\mathrm{M}\left(2 + B_{k,h}^{\mathrm{out-in}} - x_{k,i} - x_{h,j}\right) + y_h^{\mathrm{rec}} - w_k y_k^{\mathrm{rec}} - (1 - w_k) y_k^{\mathrm{dep}} \geq \gamma_{r_{k,i}^{\mathrm{out}}, r_{h,j}^{\mathrm{in}}} + w_k t t_{k,i}^{\mathrm{in}} \quad \forall k \in K, \forall h \in K, \forall i \in R_k, \forall j \in R_h : h > k, \gamma_{r_{k,i}^{\mathrm{out}}, r_{h,j}^{\mathrm{in}}} \neq 0 \quad (11)$$

$$\mathrm{M}\left(3 - B_{k,h}^{\mathrm{out-in}} - x_{k,i} - x_{h,j}\right) + w_k y_k^{\mathrm{rec}} + (1 - w_k) y_k^{\mathrm{dep}} - y_h^{\mathrm{rec}} \geq \gamma_{r_{h,j}^{\mathrm{in}}, r_{k,i}^{\mathrm{out}}} \quad \forall k \in K, \forall h \in K, \forall i \in R_k, \forall j \in R_h : h > k, \gamma_{r_{h,j}^{\mathrm{in}}, r_{k,i}^{\mathrm{out}}} \neq 0 \quad (12)$$

$$\mathrm{M}\left(2 + B_{k,h}^{\mathrm{out-out}} - x_{k,i} - x_{h,j}\right) + w_h y_h^{\mathrm{rec}} + (1 - w_h) y_h^{\mathrm{dep}} - w_k y_k^{\mathrm{rec}} - (1 - w_k) y_k^{\mathrm{dep}} \geq \gamma_{r_{k,i}^{\mathrm{out}}, r_{h,j}^{\mathrm{out}}} + w_k t t_{k,i}^{\mathrm{in}} \quad \forall k \in K, \forall h \in K, \forall i \in R_k, \forall j \in R_h : h > k, \gamma_{r_{k,i}^{\mathrm{out}}, r_{h,j}^{\mathrm{out}}} \neq 0 \quad (13)$$

$$\mathrm{M}\left(3 - B_{k,h}^{\mathrm{out-out}} - x_{k,i} - x_{h,j}\right) + w_k y_k^{\mathrm{rec}} + (1 - w_k) y_k^{\mathrm{dep}} - w_h y_h^{\mathrm{rec}} - (1 - w_h) y_h^{\mathrm{dep}} \geq \gamma_{r_{h,j}^{\mathrm{out}}, r_{k,i}^{\mathrm{out}}} + w_h t t_{h,j}^{\mathrm{in}} \quad \forall k \in K, \forall h \in K, \forall i \in R_k, \forall j \in R_h : h > k, \gamma_{r_{h,j}^{\mathrm{out}}, r_{k,i}^{\mathrm{out}}} \neq 0 \quad (14)$$

(4) Path availability constraints

Some train paths may be unavailable due to track maintenance. Depending on the location of the track maintenance, it may incur the unavailability of some inbound routes, some outbound routes or some platforms during a predetermined time period, and then leads to the unavailability of some train paths. In this paper, we take platform maintenance as an example.

If a platform needs to be maintained, constraints (15) and (16) ensure that the time period of any train occupying the platform cannot overlap with the time period of the maintenance on the platform. When train k selects path i, if $B_{k,i}$ is equal to 0, constraint (15) is used to ensure that train k can only claim the platform $r_{k,i}$ after the maintenance on this platform has been executed; otherwise, constraint (16) enforces that train k must exit

from the platform $r_{k,i}$ before the start of maintenance on this platform.

$$\mathrm{M}\left(1+B_{k,i}-x_{k,i}\right)+y_k^{\mathrm{rec}}-t_{k,i}^{\mathrm{end}}\geq 0 \quad \forall k\in K,\forall i\in R_k\colon \theta_{k,i}=0 \tag{15}$$

$$\mathrm{M}\left(2-B_{k,i}-x_{k,i}\right)+t_{k,i}^{\mathrm{start}}-y_k^{\mathrm{dep}}\geq 0 \quad \forall k\in K,\forall i\in R_k\colon \theta_{k,i}=0 \tag{16}$$

(5) Earliest arrival/departure time constraints

To guarantee passengers' boarding activity, each train is not permitted to depart earlier than its ideal departure time. Moreover, since in this paper, we reschedule trains without considering the train movement in line sections and other adjacent stations, to guarantee that the adjusted train schedule is also feasible on the whole network, each train is not permitted to arrive earlier than its ideal arrival time.

$$y_k^{\mathrm{arr}}\geq \tau_k^{\mathrm{arr}} \quad \forall k\in K \tag{17}$$

$$y_k^{\mathrm{dep}}\geq \tau_k^{\mathrm{dep}} \quad \forall k\in K \tag{18}$$

(6) Minimum dwelling time constraints

The time required by passengers to board and alight dictates the minimum amount of dwelling time required. For each train k, constraint (19) ensures that its actual dwelling time is larger than or equal to its minimum dwelling time Δ_k. Obviously, the minimum dwelling time is set to 0 if train k is a through train.

$$y_k^{\mathrm{dep}}-y_k^{\mathrm{arr}}\geq \Delta_k \quad \forall k\in K \tag{19}$$

(7) Path selection constraints

Each train k can select exactly one path.

$$\textstyle\sum_{i\in R_k} x_{k,i}=1 \quad \forall k\in K \tag{20}$$

(8) Domain of variables

The domain of variables in the model is defined by expressions (21)-(23) and is next summarized. The actual receiving time, the actual arrival time, the actual departure time and the actual leaving time of each train are defined as integer variables. The rest of the variables are defined as binary variables.

$$y_k^{\mathrm{rec}},y_k^{\mathrm{arr}},y_k^{\mathrm{dep}},y_k^{\mathrm{lea}}\in N \quad \forall k\in K \tag{21}$$

$$x_{k,i},B_{k,i}\in\{0,1\} \quad \forall k\in K,\forall i\in R_k \tag{22}$$

$$B_{k,h}^{\mathrm{p}},B_{k,h}^{\mathrm{in-in}},B_{k,h}^{\mathrm{in-out}},B_{k,h}^{\mathrm{out-in}},B_{k,h}^{\mathrm{out-out}}\in\{0,1\} \quad \forall k\in K,\forall h\in K \tag{23}$$

The proposed model is a mixed-integer linear programming formulation that can be solved by commercial solvers. However, solver efficiency of the model is still a matter in large scale problem solving due to 3 aspects of issues: (1) two types of conflict (platform and route) need to be resolved separately which expand the scale of the model; (2) sequence of trains are set as decision variable; and (3) arrival and departure times may need to be rescheduled once conflict occurs.

Thus, the following section aims to develop a heuristic algorithm that can efficiently obtain a high-quality solution in a much short time for the model. Next, we firstly decompose the overall problem into two sub-problems and then detailed techniques of the algorithm are introduced.

4 Solution Approaches

4.1 Decomposition of MILP Model

When platform/route conflict occurs, dispatchers generally first consider reassigning trains to the conflict-free paths. If the conflict still cannot be resolved, then dispatchers will modify the arrival and departure times of the relevant trains. Based on the normal practice of dispatchers, the real-time Train Platforming and Routing problem can be

decomposed into two sub-problems: (i) train path selection sub-problem with fixed arrival and departure times (TPSWFT sub-problem), (ii) partial conflict trains rescheduling sub-problem (PCTR sub-problem).

Carey and Carville (2003) also develop a heuristic algorithm which is analogous to the "manual'' methods adopted by dispatchers. The algorithm considers one train at a time and finds and resolves all conflicts for that train before considering the next train. Although this algorithm takes only a few seconds to run, it may facilitate the propagation of train delay on the railway network.

Lamorgese and Mannino (2015) decompose the real-time train dispatching problem into two sub-problems (i.e., line dispatching sub-problem and station dispatching sub-problem). The line dispatching sub-problem attempts to reschedule trains in order to minimize the deviations from the original timetable; the station dispatching sub-problem is the train platforming feasibility problem based on a given timetable. Compared with the above paper, the PCTR sub-problem in this paper is used to reschedule trains in real time and the TPSWFT sub-problem is used to assign non-conflicting platform and routes to each train. The decomposition approach is similar to the decomposition approach proposed by Lamorgese and Mannino (2015). However, the TPSWFT sub-problem in this paper is the train platforming optimization problem, while the station dispatching sub-problem is the train platforming feasibility problem. And a high-quality platform reallocation plan can absorb train delay to some extent. In addition, when solving the station dispatching sub-problem, Lamorgese and Mannino (2015) only consider the trains from or to two specific entering/existing points, while the station considered in this paper usually have multiple entering/existing points, we can collaboratively optimize the allocated platform (and inbound and outbound route) and arrival and departure times of all trains from or to different entering/existing points.

Dollevoet et al. (2014) consider the problem of delay management. They propose an iterative heuristic which first solves the delay management model with a fixed platform track assignment and then improves this platform track assignment in each step. However, for the rtTPR, the impact of rescheduling trains on train operations is more severe than the impact of reassigning trains to new platforms and routes, thus the strategy of reassigning trains to new platforms and routes should be given priority.

TPSWFT sub-problem

This sub-problem attempts to reallocate conflict-free paths (composed of inbound routes, platforms, and outbound routes) for as many trains as possible without modifying the arrival and departure time of each train. The paths are selected to minimize the impact on regular station working order. The arrival and departure time of each train are taken from the initial train schedule and train delay information or updated by PCTR sub-problem. The TPSWFT formulation is as follows:

TPSWFT: $\max Z = \sum_{k\in K} \sum_{i\in R_k} c_{k,i} x_{k,i}$

Subject to:

Constraints (3)-(16), (21)-(23)

$$y_k^{\mathrm{arr}} = \tau_k^{\mathrm{arr}} \quad \forall k \in K \tag{24}$$

$$y_k^{\mathrm{dep}} = \tau_k^{\mathrm{dep}} \quad \forall k \in K \tag{25}$$

$$\sum_{i\in R_k} x_{k,i} \leq 1 \quad \forall k \in K \tag{26}$$

PCTR sub-problem

This sub-problem aims to further resolute platform and route conflicts through

rescheduling the arrival and departure times of trains. The set of trains involved in solving this sub-problem, denoted by K', includes all the trains that cannot be allocated a conflict-free path in the TPSWFT sub-problem. The model of the PCTR sub-problem is similar to the model of the original problem, except that the train sets considered by these two problems are different. The PCTR formulation is as follows:

PCTR: $\min Z = \sum_{k\in K'}(y_k^{\mathrm{arr}} - \tau_k^{\mathrm{arr}}) + \sum_{k\in K'}(y_k^{\mathrm{dep}} - \tau_k^{\mathrm{dep}})$

Subject to:

Constraints (3)-(23)

4.2 Solution Approach of TPSWFT Sub-problem

When the arrival and departure times of each train are fixed, the conflict relationship between any two paths of different trains can be determined. Let $\delta_{k,i,h,j}$ denote whether the path i of train k conflicts with the path j of train h, which is equal to 0 when the two paths conflict with each other (because of platform conflict or route conflict), and 1 otherwise. In addition, the binary parameter $\delta_{k,i,h,j}$ has the following characteristics:

(1) Symmetry, i.e., $\delta_{k,i,h,j} = \delta_{h,j,k,i}, \forall k \in K, \forall h \in K, \forall i \in R_k, \forall j \in R_h$.
(2) If the path i of train k is unavailable, then $\delta_{k,i,h,j} = 0$ and $\delta_{h,j,k,i} = 0$, $\forall h \in K, \forall j \in R_h$.
(3) Since each train can be assigned to at most one path, $\delta_{k,i,k,j} = 0(\forall k \in K, \forall i,j \in R_k: i \neq j)$ and $\delta_{k,i,k,j} = 1(\forall k \in K, \forall i,j \in R_k: i = j)$.

An undirected conflict graph $G = (V, A)$ is built based on the conflict relationships between train paths, where each vertex $v_{k,i} \in V$ corresponds to a possible path i for train k and is assigned a weight $c_{k,i}$, and each arc $a_{k,i,h,j} \epsilon A$ connecting the two vertexes (vertex $v_{k,i}$ and vertex $v_{h,j}$) indicates that the corresponding train paths (path i of train k and path j of train h) are compatible with each other (i.e., when $\delta_{k,i,h,j}$ is equal to 1, the arc $a_{k,i,h,j}$ exists). It is worth noting that there is no connection between any vertexes corresponding to paths for the same train.

The TPSWFT sub-problem attempts to reallocate conflict-free paths for as many trains as possible. Based on the undirected conflict graph, this sub-problem can be formulated as the Maximum Vertex Weight Clique Problem (MVWCP). The MVWCP can be described as follows:

> Given an undirected graph $G = (V, A)$, a clique is a set $C \subseteq V$ such that there is exactly one arc connecting any two vertexes of C. And for a clique C of G, define its weight as $W(C) = \sum_{v_{k,i}\in C} c_{k,i}$. The MVWCP is to determine a clique C^* of maximum weight, i.e., $\forall C \in \Omega, W(C^*) \geq W(C)$ where Ω is the set of all possible cliques of the graph.

Furthermore, for each vertex $v_{k,i}$ of G, the *vertex weight degree* $wd_{k,i}$ is defined to reflect the maximum weight that the clique may reach if vertex $v_{k,i}$ is selected. The formula for calculating $wd_{k,i}$ is as follows:

$$wd_{k,i} = \sum_{h\in K} \max\{\delta_{k,i,h,j} c_{h,j} | j \in R_h\} \quad (27)$$

Many algorithms and methods have been proposed to solve MVWCP, see Wu and Hao (2015). In this paper, we develop a branch and bound algorithm to solve it which includes implicational rules enabling to speed up the computation and still can acquire optimal solutions.

The branch and bound algorithm is developed in the form of a tree structure. Each level of the tree, denoted by l, represents assigning path for train l ($l \neq 0$). Each node on any particular level, denoted by $n_{l,p}$, represents assigning path p for train l ($p \leq |R_l|$) or indicates that no path can be assigned for the train ($p = |R_l| + 1$). Leaf nodes define feasible train path selection plans or partial maximum clique.

For each node $n_{l,p}$ of the branch-and-bound tree, the following variables are defined:

(1) the current clique $C_{l,p}$ on node $n_{l,p}$ is used to record all chosen vertexes currently.
(2) the current weight $cw_{l,p}$ on node $n_{l,p}$ is defined to reflect the accumulated weight of the current clique.
(3) the upper bound $ub_{l,p}$ on node $n_{l,p}$ is defined to reflect the maximum weight that the partial maximum clique may reach if the branch is continued based on node $n_{l,p}$ until one leaf node is obtained. The formula for calculating $ub_{l,p}$ is as follows:

$$ub_{l,p} = \sum_{h \in K} \max\{c_{h,j} \times \min\{\delta_{k,i,h,j} | v_{k,i} \in C_{l,p}\} | j \in R_h\} \quad (28)$$

(4) the conflict relationship matrix $E_{l,p}$ on node $n_{l,p}$ is defined to reflect whether any path of any train conflicts with any chosen path of the current clique. Each element $e_{k,i}^{l,p}$ of matrix $E_{l,p}$ can be computed by formula (29).

$$e_{k,i}^{l,p} = \min\{\delta_{k,i,h,j} | v_{h,j} \in C_{l,p}\} \quad (29)$$

Branch and Bound algorithm procedure

Step 0. Initialization. Set $l = 0$ and $p = 1$, and generate the root node $n_{0,1}$. Set $C_{0,1} = \emptyset$ and $cw_{0,1} = 0$. Each element $e_{k,i}^{o,1}$ of matrix $E_{0,1}$ is set to 1. The upper bound $ub_{0,1}$ on the root node $n_{0,1}$ is also the upper bound of the overall TPSWFT sub-problem, denoted by UB, which can be computed by formula (30). Turn to Step 1.

$$UB = ub_{0,1} = \min\{\max\{wd_{k,i} | i \in R_k\} | k \in K\} \quad (30)$$

Step 1. Node selection. If all nodes of the branch-and-bound tree are leaf nodes, the branch and bound algorithm terminates; otherwise, pick the node of the last level with the maximum current weight and the maximum upper bound, turn to Step2.

Step 2. Branching and Bounding. Based on the selected node in Step 1, assign a path for the next train, and the vertex in the conflict graph corresponding to the specified path is added into the current clique of the selected node accordingly. Hence, generate a series of nodes, and the number of newly generated nodes is equal to the number of possible paths of the next train plus one. For each newly generated node $n_{l,p}$, calculate $C_{l,p}$, $cw_{l,p}$, $ub_{l,p}$ and $E_{l,p}$, and turn to Step 3. When the last newly generated node has been checked, turn to Step 1.

Step 3. Pruning. For each newly generated node $n_{l,p}$, (1) if the newly added vertex is not connected to each vertex in the current clique of the selected node, node $n_{l,p}$ will be removed; (2) if $l = |K|$, i.e., node $n_{l,p}$ is a leaf node, and if the current weight $cw_{l,p}$ is greater than the current optimal solution, then update the current optimal solution. And if the current optimal solution is equal to UB, the branch and bound algorithm terminates; otherwise, for each node $n_{m,q}$ of the branch-and-bound tree, if its upper bound $ub_{m,q}$ is less than or equal to the current optimal solution, then node $n_{m,q}$ is removed; (3) if $l < |K|$ and the upper bound $ub_{l,p}$ is less than or equal to the current optimal solution, then node $n_{l,p}$ is removed; (4) if $l < |K|$

and the upper bound $ub_{l,p}$ is greater than the current optimal solution, turn to Step 4 to check whether node $n_{l,p}$ meets *Equivalence Rule*.

Step 4. Equivalence Rule. For the newly generated node $n_{l,p}$ and any node $n_{l,q}$ which is on the same level and still exists on the branch-and-bound tree after Step 3, if the matrix $E_{l,p}$ and the matrix $E_{l,q}$ are equivalent, which shows that for any leaf node derived by continuous branching based on node $n_{l,p}$, there must be a leaf node derived by continuous branching based on node $n_{l,q}$, and these two leaf nodes have the same weight, then node $n_{l,p}$ is removed. The conditions for the equivalence of matrix $E_{l,p}$ and matrix $E_{l,q}$.are described as follows:

$\forall h \in K, h > l$ and $\forall j \in R_h$,

(1) if $r_{h,j} = r_{l,p}$, then

$$\begin{cases} e_{h,j}^{l.p} = 0, e_{h,j}^{l.q} = 0, & \text{if } \forall jj \in R_h, r_{h,jj} \neq r_{l,q} \\ e_{h,j}^{l.p} = e_{h,jj}^{l.q}, & \text{if } \exists jj \in R_h, r_{h,jj} = r_{l,q} \end{cases} \tag{31}$$

(2) if $r_{h,j} = r_{l,q}$, then

$$\begin{cases} e_{h,j}^{l.p} = 0, e_{h,j}^{l.q} = 0, & \text{if } \forall jj \in R_h, r_{h,jj} \neq r_{l,p} \\ e_{h,j}^{l.p} = e_{h,jj}^{l.q}, & \text{if } \exists jj \in R_h, r_{h,jj} = r_{l,p} \end{cases} \tag{32}$$

(3) if $r_{h,j} \neq r_{l,p}$ and $r_{h,j} \neq r_{l,q}$, then

$$e_{h,j}^{l.p} = e_{h,j}^{l.q} \tag{33}$$

In conclusion, a key strategy in the reduction of the computational effort of branch and bound algorithm procedures for the TPSWFT sub-problem is that based on the concept of the *vertex weight degree*, high quality upper bound can be obtained which serves both as an efficient pruning strategy, as well as an efficient stopping criterion. In addition, the *Equivalence Rule* is also used to tremendously reduce the size of the branch-and-bound tree as a more efficient pruning strategy. Based on these above implicational rules, the branch and bound algorithm enable to acquire optimal solutions within short time limits.

4.3 Solution Approach of PCTR Sub-problem

The trains which cannot be allocated a conflict-free path in the TPSWFT sub-problem, should be rescheduled in the PCTR sub-problem. At the same time, paths will be reassigned for the conflict trains in order to minimize the overall deviation from the ideal planed schedule. Since the model of the PCTR sub-problem is similar to the model of the original problem with a relative small scale, it can be solved by CPLEX solver efficiently. The algorithm for solving the PCTR sub-problem, which is called synchronous adjustment algorithm, is described as follows:

Synchronous adjustment algorithm procedure

Step 0. Generate conflict trains set J. The unassigned trains set K' is the set of trains which cannot be allocated a conflict-free path in the TPSWFT sub-problem. For each train k of set K', calculate its conflict trains set J_k. If the station occupancy time of train k (i.e., from its receiving time to its leaving time) overlaps with the station occupancy time of train h ($\forall h \in K$), which implies that the changes of the arrival and departure times of train k may cause the infeasible path of train h or rescheduling the arrival and departure times of train h may rule out the conflict between train k and other trains, then train h is added into the set J_k. And $J =$

$\{J_k | \forall k \in K'\}$. Turn to Step 1.

Step 1. Set operations. $\forall k \in K'$, $\forall h \in K'$, and $k \neq h$, if $J_k \cap J_h \neq \emptyset$, then make $J_k = J_k \cup J_h$, remove J_h from J. Turn to Step 2.

Step 2. Synchronous adjustment. For each conflict trains set J_k, call PCTR model to further resolve the conflicts.

4.4 Iterative Algorithm

Since not all trains will input to the PCTR sub-problem, the feasibility is not guaranteed for that rescheduled trains may lead to new conflicts. Therefore, these two sub-problems need to be solved iteratively until all conflict are resolved.

In addition, to ensure that the iterative algorithm stops within the time limit, we use an additional criterion to terminate the algorithm, i.e., if the current iteration index is greater than a pre-given threshold, the iterative algorithm terminates. Specifically, χ denotes the iteration index, and N denoted the maximum number of iterations.

When the algorithm terminates, if there are still some conflicts between the trains, for each unassigned train, assign it to its original allocated platform and inbound and outbound routes, and delay it until its arrival and departure times are greater than each of the assigned trains on its original allocated platform and make sure it is compatible with all other trains. Thus a feasible solution is obtained.

The iterative algorithm framework is as follows:

Iterative algorithm procedure

Input: The ideal planed train schedule, original platform allocation plan, original route assignment plan, detailed station yard topology, track maintenance information, initial train delay information, and so on.

Step 0. Initialization. Generate all possible paths for each train and the Degree of Conflict between any two routes. Set iteration index $\chi = 0$, and turn to Step 1.

Step 1. TPSWFT sub-problem. Calculate the binary parameter $\delta_{k,i,h,j}$, construct the undirected conflict graph, and call the branch and bound algorithm to reallocate conflict-free paths for as many trains as possible. If the number of unassigned trains is equal to 0, the iterative algorithm terminates. Set $\chi = \chi + 1$, if $\chi \leq N$, then turn to Step 2; otherwise, generate a feasible solution, and the iterative algorithm terminates.

Step 2. PCTR sub-problem. Collect all unassigned trains in set K', and call the synchronous adjustment algorithm to further resolute conflicts through rescheduling the arrival and departure times of trains. Turn to Step 3.

Step 3. Update the arrival and departure times of each train and the conflict relationship between any two paths of different trains. Turn to Step 1.

Output: The adjusted arrival and departure times of each train, the adjusted path selection plan (including platform reallocation plan and the route reassignment plan).

5 Case Study

We performed the numerical experiment using operational data from the Zhengzhou East high-speed railway station to test how well the proposed algorithm may be applied in the real-world instance. The following experiment is performed on a computer Intel® Core™ i7-4790 CPU @ 3.6GHz processor and 16GB RAM.

As shown in Figure 3, this station includes 6 entering points, 5 leaving points, 12 platforms and 67 inbound and outbound routes. The traversal time of each route is set according to the length of the route and the average train speed in the yard. The shortest safety time interval permitted between two adjacent trains occupying the same platform ε is set to 180 seconds.

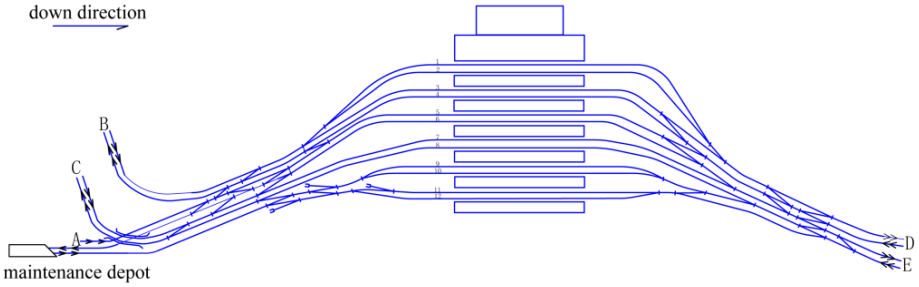


Figure 3: Layout of Zhengzhou East high-speed railway station.

Table 3 shows the detailed information for 84 trains. For each train, we record its entering point (EnP), its leaving point (LeP), its ideal arrival time (ArrTime), its departure time (DepTime) and its original allocated platform (Platform). The earliest train arrives at time 10:07:30 and the last train leaves station at 11:58:00. The entire considered time horizon is 2 hours.

Table 3: Train information.

ID	EnP	LeP	ArrTime	DepTime	Platform	ID	EnP	LeP	ArrTime	DepTime	Platform
1	5	11	10:07:30	10:10:00	3	43	6	3	10:40:50	10:50:00	3
2	8	11	10:09:20	10:15:00	4	44	6	7	11:04:40	11:11:40	2
3	4	1	10:08:20	10:16:00	5	45	8	1	10:54:50	11:00:50	3
4	10	9	10:07:30	10:15:00	9	46	8	3	11:00:40	11:05:00	4
5	6	9	10:07:30	10:09:50	6	47	8	7	11:05:40	11:11:40	1
6	2	7	10:04:30	10:06:30	4	48	10	11	10:55:00	11:01:40	10
7	5	1	10:02:30	10:11:00	1	49	2	11	11:04:10	11:11:40	11
8	8	1	10:04:00	10:06:00	2	50	2	9	10:59:10	11:06:10	8
9	10	11	10:02:30	10:20:00	10	51	2	9	11:09:10	11:16:10	10
10	2	11	10:09:20	10:25:00	7	52	10	11	11:00:50	11:06:40	9
11	2	9	10:14:20	10:20:00	8	53	10	9	11:05:50	11:11:10	12
12	2	9	10:19:20	10:25:00	9	54	5	11	11:09:50	11:16:40	3
13	10	11	10:12:30	10:30:00	11	55	8	11	11:14:40	11:21:40	6
14	10	9	10:17:30	10:30:00	12	56	4	1	11:06:20	11:11:40	5
15	4	3	10:01:40	10:03:40	5	57	10	9	11:15:00	11:22:00	9
16	4	7	10:20:40	10:22:40	5	58	6	9	11:21:00	11:27:00	5
17	6	3	10:20:40	10:29:40	6	59	2	7	11:14:40	11:16:40	4
18	6	7	10:15:40	10:17:40	1	60	5	1	11:14:50	11:19:20	2
19	8	3	10:10:50	10:19:40	2	61	8	1	11:19:50	11:24:20	1
20	5	3	10:14:50	10:24:40	3	62	10	11	11:20:00	11:26:50	11
21	5	7	10:22:30	10:27:40	1	63	2	11	11:19:40	11:31:50	12
22	4	9	10:27:30	10:35:00	5	64	2	9	11:25:00	11:32:00	7
23	8	7	10:30:40	10:32:40	2	65	2	9	11:30:00	11:37:00	8
24	8	9	10:35:40	10:40:00	6	66	10	11	11:25:50	11:36:50	10
25	6	1	10:25:40	10:27:40	4	67	10	9	11:30:50	11:42:00	9
26	6	11	10:32:20	10:35:00	4	68	4	3	11:26:20	11:33:20	6
27	2	1	10:34:10	10:36:10	3	69	4	7	11:31:40	11:37:40	5
28	4	11	10:39:40	10:41:40	5	70	6	3	11:31:10	11:38:30	3
29	2	3	10:39:40	10:45:00	4	71	6	7	11:26:50	11:32:40	4
30	10	11	10:29:10	10:46:40	10	72	8	3	11:31:20	11:43:50	1

31	2	11	10:44:40	10:56:40	11	73	5	3	11:37:30	11:48:50	2
32	2	9	10:49:40	10:51:10	8	74	5	7	11:44:10	11:47:40	3
33	2	9	10:54:10	10:56:10	7	75	10	11	11:35:50	11:41:50	11
34	10	11	10:37:30	10:51:40	9	76	2	11	11:37:10	11:46:50	12
35	10	9	10:50:00	11:01:10	12	77	2	9	11:42:10	11:47:00	10
36	4	1	10:46:20	10:50:50	5	78	2	9	11:47:10	11:52:00	9
37	4	3	10:51:20	10:57:10	6	79	10	11	11:49:10	11:51:50	11
38	4	7	10:56:20	11:01:40	5	80	2	9	11:52:10	11:57:30	10
39	5	1	10:40:00	10:45:50	2	81	10	11	11:54:10	11:56:50	12
40	5	3	10:32:30	10:40:00	1	82	4	1	11:42:30	11:48:20	5
41	5	7	10:49:50	10:56:40	1	83	5	1	11:53:40	11:58:00	1
42	6	1	10:49:40	10:55:50	4	84	4	3	11:47:30	11:52:10	6

The following disruption situation is considered: (1) train 6 is two minutes behind schedule; (2) platform 5 is unavailable from 10:30:00 to 10:40:00; (3) train 44 is five minutes behind schedule. In this case, the total number of train paths is 379. The model of the overall problem has 1,433,882 constraints and 35,954 variables, which takes 3600s to obtain a solution with 58.94% duality gap by using the CPLEX solver. By using the iterative algorithm, the computational time of this instance is less than two seconds. The number of iterations is two. The minimum overall deviation from the ideal planed schedule of each train is 320 seconds.

The instance size is similar to the medium sized railway station Arnhem presented by Zwaneveld et al. (2001). Arnhem has 16 platform and is visited by about 40 trains per hour. But since the problem considered by Zwaneveld et al. (2001) is at the strategic level and their computer computing power is different from ours, we cannot compare the computational efficiency of our approach with their solving methods. However, the results still clearly demonstrate the great potential of our iterative algorithm.

Figure 4 shows the original platform allocation plan. Figure 5 shows the adjusted platform allocation. In these two figures, each rectangle represents a train occupying a platform, the length of each rectangle represents the duration of occupying the platform and the number to the right of each rectangle indicates the train ID. The rectangle with light colour implies that the corresponding train's arrival and departure time, its platform, inbound and outbound routes are all not changed; while the rectangle with dark colour implies that the corresponding train is rescheduled or reassigned a different platform, inbound route or outbound route. Table 4 shows the information of trains which is rescheduled or reassigned a different path.

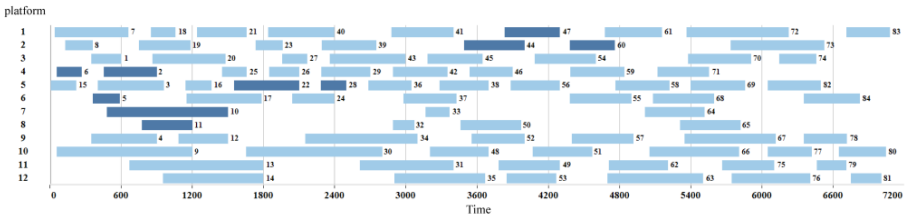


Figure 4: The original platform allocation plan.

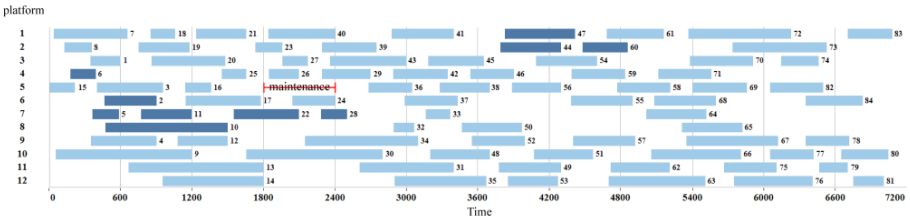


Figure 5: The adjusted platform allocation plan.

Table 4: Information of trains which is rescheduled or reassigned a different path.

ID	Original plan			Adjusted plan		
	ArrTime	DepTime	Platform	ArrTime	DepTime	Platform
2	10:09:20	10:15:00	4	10:09:20	10:15:00	6
5	10:07:30	10:09:50	6	10:07:30	10:09:50	7
6	10:02:30	10:04:30	4	10:04:30	10:06:30	4
10	10:09:20	10:25:00	7	10:09:20	10:25:00	8
11	10:14:20	10:20:00	8	10:14:20	10:20:00	7
22	10:27:30	10:35:00	5	10:27:30	10:35:00	7
28	10:39:40	10:41:40	5	10:39:40	10:41:40	7
44	10:59:40	11:06:40	2	11:04:40	11:11:40	2
47	11:05:40	11:11:40	1	11:05:40	11:13:40	1
60	11:14:50	11:19:20	2	11:16:30	11:21:00	2

6 Conclusions

In this paper we studied the real-time train platforming and routing problem at a busy complex high-speed railway station in a disrupted situation. A bi-objective mixed integer linear programming model was formulated to solve the problem, which is also a universal model for the route-lock sectional/integral release interlocking system. An iterative algorithm is designed to solve the MILP model efficiently. The results of real-world experiment based on the Zhengzhou East high-speed railway station show the proposed model and solution approach have great potential to be applied in the real-world operations.

Our future research will focus on the following significant aspects:

1. This paper assumes that all trains travel at the same constant speed regardless of route choice. The speed profile for each train can be taken into consideration in order to meet the operational requirements.
2. To the feasibility of the solution, this paper enforces that all trains are not permitted to arrive earlier than its ideal arrival time. In fact, trains are permitted to arrive a few minutes earlier without violating the condition of sections. Thus, the proposed model and algorithm can be extended to the railway network.
3. At the tactical level, based on the fixed arrival and departure times, reasonable adjustment to the platform allocation plan and the route assignment plan can increase buffer time. Buffer time can be used to absorb train delay to some extent. Thus, the robustness of platform allocation and route assignment plan can be another research direction.

Acknowledgements

The research was supported by National Key Research and Development Program of China (No. 2017YFB1200700-1) and National Natural Science Foundation of China (No. U1834209).

References

Boccia, M., Mannino, C., Vasilyev, I., 2013. "The dispatching problem on multitrack territories: Heuristic approaches based on mixed integer linear programming", *Networks*, vol. 62, pp. 315-326.

Cacchiani, V., Huisman, D., Kidd, M., et al., 2014. "An overview of recovery models and algorithms for real-time railway rescheduling", *Transportation Research Part B: Methodological*, vol. 63, pp. 15-37.

Caimi, G., Fuchsberger, M., Laumanns, M., et al., 2012. "A model predictive control approach for discrete-time rescheduling in complex central railway station areas", *Computers & Operations Research*, vol. 39, pp. 2578–2593.

Caprara, A., Galli, L., Toth, P., 2011. "Solution of the train platforming problem", *Transportation Science*, vol. 45, pp. 246–257.

Carey, M., Carville, S., 2003. "Scheduling and platforming trains at busy complex stations", *Transportation Research Part A: Policy and Practice*, vol. 37, pp. 195–224.

D'ariano, A., Pacciarelli, D., Pranzo, M., 2007. "A branch and bound algorithm for scheduling trains in a railway network", *European Journal of Operational Research*, vol. 183, pp. 643-657.

D'Ariano, A., Corman, F., Pacciarelli, D., et al., 2008. "Reordering and local rerouting strategies to manage train traffic in real time", *Transportation science*, vol. 42, pp. 405-419.

Dollevoet, T., Huisman, D., Kroon, L., et al., 2014. "Delay management including capacities of stations", *Transportation Science*, vol. 49, pp. 185-203.

Lamorgese, L., Mannino, C., 2015. "An exact decomposition approach for the real-time train dispatching problem", *Operations Research*, vol. 63, pp. 48-64.

Lusby, R. M., Larsen, J., Ehrgott, M., et al., 2011. "Railway track allocation: models and methods", *OR spectrum*, vol. 33, pp. 843-883.

Lusby, R. M., Larsen, J., Ryan, D., et al., 2011. "Routing trains through railway junctions: a new set-packing approach", *Transportation Science*, vol. 45, pp. 228-245.

Lusby, R. M., Larsen, J., Ehrgott, M., et al., 2013. "A set packing inspired method for real-time junction train routing", *Computers & Operations Research*, vol. 40, pp. 713-724.

Meng, L., Zhou, X., 2014. "Simultaneous train rerouting and rescheduling on an N-track network: A model reformulation with network-based cumulative flow variables", *Transportation Research Part B: Methodological*, vol. 67, pp. 208-234.

Pellegrini, P., Marlière, G., Pesenti, R., et al., 2015. "RECIFE-MILP: An effective MILP-based heuristic for the real-time railway traffic management problem", *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, pp. 2609-2619.

Pellegrini, P., Pesenti, R., Rodriguez, J., 2019. "Efficient train re-routing and rescheduling: Valid inequalities and reformulation of RECIFE-MILP", *Transportation Research Part B: Methodological*, vol. 120, pp. 33-48.

Rodriguez, J., 2007. “A constraint programming model for real-time train scheduling at junctions”, *Transportation Research Part B: Methodological*, vol. 41, pp. 231-245.

Rodriguez, J., Kermad, L., 1998. “Constraint programming for real-time train circulation management problems in railway nodes”, *WIT Transactions on The Built Environment*, vol. 37.

Sama, M., Pellegrini, P., D’Ariano, A., et al., 2016. “Ant colony optimization for the real-time train routing selection problem”, *Transportation Research Part B: Methodological*, vol. 85, pp. 89-108.

Sels, P., Vansteenwegen, P., Dewilde, T., et al., 2014. “The train platforming problem: The infrastructure management company perspective”, *Transportation Research Part B: Methodological*, vol. 61, pp. 55-72.

Zwaneveld, P.J., Kroon, L.G., Van Hoesel, S.P.M., 2001. “Routing trains through a railway station based on a node packing model”, *European Journal of Operational Research*, vol. 128, pp. 14-33.

Zwaneveld, P.J., Kroon, L.G., Romeijn, H.E., et al., 1996. “Routing trains through railway stations: Model formulation and algorithms”, *Transportation science*, vol. 30, pp. 181-194.

Wu, Q., Hao, J., 2015. “A review on algorithms for maximum clique problems”, *European Journal of operational research*, vol. 242, pp. 693-709.

Analysis of Timetable Rescheduling Policy for Large-scale Train Service Disruptions

Rieko Otsuka [a,1], Masao Yamashiro [a], Itaru Otsuchibashi [b], Sei Sakairi [c]
[a] Research and Development Group, Hitachi Ltd.
1-280 Higashi-koigakubo, Kokubunji-shi, Tokyo 185-8601, Japan
[1]E-mail: rieko.otsuka.gd@hitachi.com; Phone: +81-70-3874-6607
[b] Information and Communication Technology Business Division, Hitachi, Ltd.
6-27-18 Minami-ohi, Shinagawa-ku, Tokyo, 140-8572, Japan
[c]Research and Development Center of JR East Group, East Japan Railway Company
2-479 Nisshin-cho, Kita-ku, Saitama 331-8513, Japan

Abstract
When a train disruption happens, dispatchers should make a train rescheduling plan quickly and manage both trains and crews to reduce serious congestion at disrupted area. Timetable rescheduling is quite domain specific work that requires a lot of experience, knowledge and decisive skills for dispatchers to make an adequate operation plan by taking account of various constraints. Furthermore, in recent years, dispatchers are expected to consider less impact on passengers to improve transport service quality. Thus, workload of train dispatchers become larger year by year. Then we have developed a dispatcher's decision-making support tool for train rescheduling. Our proposed prototype enables to recommend when and where turn around operations should be performed based on analysis results using historical operated train timetable data.

Keywords
Timetable Rescheduling, Simulation, Big Data Analysis

1 Introduction

In the Tokyo metropolitan area, approximately 15 million people use railway service every day. Train disruption in Tokyo area affects a great number of passengers for a long time. During a train disruption, dispatchers must modify train timetable to restart train service operation as soon as possible. Modification work of train timetable is called train rescheduling. In the operation control centre of JR East company, hundreds of train dispatchers need to quickly make a rescheduling timetable plan by taking account of delays of trains, train resource, human resource (i.e. crews), congestion, weather condition and so on.

Timetable rescheduling work gives strong stress to dispatchers because it requires a lot of experience and knowledge. Especially, in recent years, dispatchers are expected to reduce impact for passengers in addition to quick recovery to a planned timetable by performing turn around operations to provide enough transport capacity during a train disruption. Though work of a dispatcher will become more harder and more stressful, dispatching operation is still manually performed and depends on individual skill and experience. Furthermore, workload reduction and efficiency are desired because train rescheduling know-how is not shared among dispatchers effectively.

Therefore, we propose a new decision-making support system for dispatchers to make

a better train rescheduling plan. We introduce data-driven approach to recommend train rescheduling operations using historical train operation data derived from train management system. By utilizing many dispatcher's operation history, we believe that train rescheduling skills of an individual dispatcher improve efficiently.

Some technologies for supporting dispatchers work have been already proposed. To support dispatcher's decision-making, two process are essential. One is generating process of train rescheduling plan. Many research based on mathmatical optimization approach have been presented as shown in Shakibayifar et al.(2018) and Huang et al.(2018). The other is evaluation process for effectiveness of train timetable plans. A macroscopic indicator named SCORE for evaluating a train disruption impact quantitatively was proposed by Tsunoda et al. (2015). With SCORE, dispatchers can review their rescheduling work from the viewpoint of passengers after a train disruption convergence. As a succeeding research, a SCORE-based simulator for making a train rescheduling plan was developed by Yamashiro et al. (2017). Dispatchers can make an optimal plan by comparing predictive SCOREs calculated based on normal travel demand of passengers. In addition, by their works, the effects of turn around operations during a train disruption have been revealed from the viewpoint of passenger's extra travel time. Other simulation researches also have been proposed. Kunimatsu et al. (2015) have evaluated turn around operations from a passenger's perspective with their simulator. In addition, monitoring system for train delay and congestion for dispatchers have been developed by Sakairi et al. (2016). Integration framework of primary work process for disruption management have been proposed by Besinovic et al. (2015). However, there have not been established method to recommend a rough design policy such as whether turn around operation should be performed or not for a dispatcher. Our motivations of this research are as follows.

(1) Support a dispatcher to make an optimal train rescheduling plan at the initial stage
(2) Indicate possibility of utilizing historical operated data in railway operation field

This paper is structured as follows. In Section 2, we describe train rescheduling operation problem and propose our research methodology. In Section 3, train rescheduling operation extraction method are presented. In Section 4, evaluation results are shown. Finally, Section 6 concludes the paper.

2 Research Methodology

2.1 Train rescheduling operation

Train rescheduling is one of the most important tasks for train dispatchers to recover train operation during a train disruption happens. Train dispatching workflow is as follows.

(i) Emergency notification

Dispatchers receive emergency notification from a station staff or a train attendant. Once an occurrence of train disruption is confirmed, they immediately stop all related trains.

(ii) Monitor the situation

Dispatchers receive follow-up reports of the incidents from stations and trains involved in the accident.

(iii) Rough design of train rescheduling plan

Table 1 shows a list of major train rescheduling operations. Required level of train rescheduling operations depends on a train disruption scale. When a large-scale disruption

happens, dispatchers estimate operation restarting time and make a train rescheduling plan roughly. In a long time, service suspension case, dispatchers are often expected to provide extra trains and turn around operations to supply transport capacity for passengers as much as possible. On the other hand, in case of small-scale disruptions, combination of local rescheduling operations such as changing train departure orders are focused.

(iv) Confirm assets and human resource

After deciding an initial design for rescheduling plan, dispatchers make a detailed plan by considering train assets and human resource. Especially crew assignment is a complicated issue. For local train service operation in Tokyo metropolitan area, both train driver and attendant, i.e. at least two crews are essential to operate. Once asset and human resource are secured, dispatchers input rescheduling contents to train management system one by one.

(v) Catch up on planned timetable

When dispatchers restarts operation of disrupted trains, they concentrate recovering to original planned timetable. In other words, they gradually reduce train delays by combination train cancellation, changing train departure order and departure time. As mentioned in the above, train operation in the Tokyo area requires many human resources. Then it is desired that they operate according to the planned timetable from a perspective of train crews.

Table 1: Primal train rescheduling operation in Tokyo metropolitan area

No.	Rescheduling Operation
1	Extra train
2	Extend operational section
3	Change train type (local, rapid, express)
4	Change departure time of a train
5	Cancellation (fully and partially)
6	Turn around operation
7	Change train track
8	Change train id in train diagram
9	Change departure order at a station

During large-scale train disruptions, a dispatcher repeats monitoring process and re-planning process by taking account of various resource constraints. Figure 1 shows difference of two rescheduled timetables. We choose two disruption cases happened in the similar situation, i.e. both disruptions happened in same service line and almost same time. Main difference between case1 and case2 in Figure 1 is whether turn around operation was operated or not.

The left rescheduled timetable of Figure 1 has the following features.

- Most of trains have been stopped for nearly an hour.
- Significant impact on passengers because trains didn't move entirely.
- Train operation were restarted at once.
- Rapidly recover to the planned timetable after a disrupted train restarted

On the contrary, the features of the right rescheduled timetable of Figure 1 is as follows.

- Some trains have been operated even immediately after a disruption happened.
- Less impact on passengers than the right case of Figure 1 because they could take a

train even in during a disruption.

- It is considered that workload of dispatchers was heavy because it has taken a long time to recover.

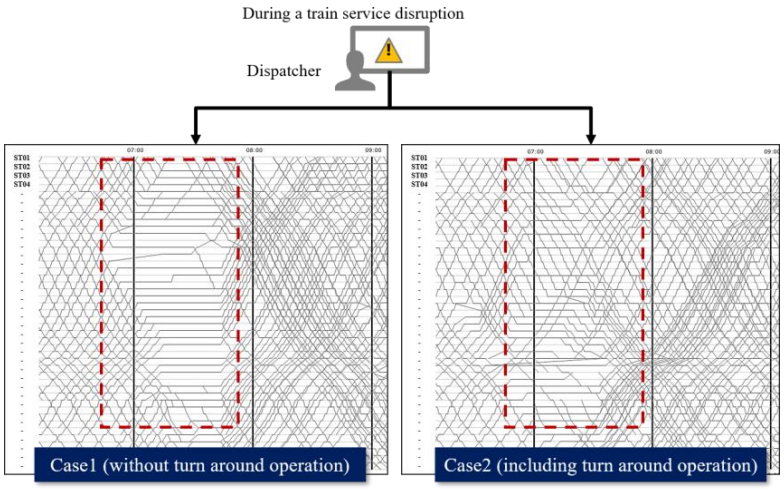


Figure 1: Comparison of two train rescheduling operations

Table 2 shows analysis results regarding the effects of turn around operations. It shows a ratio of transport capacity and an extra travel time per passenger for one actual disruption case. In this disruption case, accidental train service line can be divided in 9 operation sections and turn around operations were performed from section 7 to section 9. Alternative train service line exists around section 4, 5 and 6. From table 2, it is found that turn around operations increased transport capacity remarkably. Hereby, extra travel time per passenger decreased according to turn around operations.

Table 2: Comparison of influence on passengers by operation section

Operation Section	Ratio of Transport Capacity (disrupted day / normal day)	Extra Travel Time per Passenger (min)
All section	66%	12.4
Section 1	7%	27.0
Section 2	14%	27.8
Section 3	27%	24.5
Section 4	60%	18.7
Section 5	72%	15.0
Section 6	77%	14.9
Section 7	119%	9.8
Section 8	119%	7.9
Section 9	120%	7.2

Recently, railway operators have been strongly encouraged to manage train timetable to

minimize the effect on passengers. As one method to achieve that, turn around operations during a disruption are focused. However there haven't been enough analyzed that when and where turn around operations should be performed. Currently, train dispatchers decide it depending their experience and individual intuitive. Therefore, we propose a dispatchers' decision-making support system which can recommends turn around operation plan based on historical data analysis. Nevertheless, thinking process history of rescheduling planning weren't stored in any existing systems. Then we focus utilizing operated train timetable data history to reproduce train rescheduling arrangement process.

2.2 Train operated data

The key dataset in this research is historical train operated data obtained from the train management system. A train management system oversees the location of each train and it is possible to know how long the train is delayed comparing to the actual location with the original transportation plan.Table 3 shows primal information included train operated data. Table 3(a) is an example of normal operation. Table 3(b) shows am example of turn around operation. These train operated data are stored daily as a text-based data. It includes spatial and temporal information of all trains, i.e. stopping station, planned departure and arrival time, operated departure and arrival time as shown in Table 3. In the case of turn around operation, operated time records were disappeared partially. In addition to this, it includes previous and next train Ids. With these datasets, we can trace change of train Ids thorough one day.

Table 3: Example of operated train timetable data

(a) Normal operation

Direction	**Train Id**	**Train Type**	**Order**	**Station**	**Arr. Time (planned)**	**Dep. Time (planned)**	**Arr. Time (operated)**	**Dep. Time (operated)**
Inbound	1xxA	Local	1	ST01		8:52:30		8:52:30
Inbound	1xxA	Local	2	ST02	8:54:40	8:54:50	8:54:45	8:54:55
Inbound	1xxA	Local	3	ST03	8:56:20	8:56:40	8:56:30	8:56:45
Inbound	1xxA	Local	4	ST04	8:58:30	8:58:50	8:58:35	8:58:55
Inbound	1xxA	Local	5	ST05	8:59:45	9:00:15	8:59:50	9:00:15
Inbound	1xxA	Local	6	ST06	9:01:20	9:01:40	9:01:20	9:01:45
Inbound	1xxA	Local	7	ST07	9:03:40	9:04:00	9:03:50	9:04:30
Inbound	1xxA	Local	8	ST08	9:05:40	9:06:00	9:06:00	9:06:20
---	---	---	---	---	---	---	---	---

(b) Turn around operation (operated from ST05 to ST08)

Direction	Train Id	Train Type	Order	Station	Arr. Time (planned)	Dep. Time (planned)	Arr. Time (operated)	Dep. Time (operated)
Inbound	1xxB	Local	1	ST01		8:52:30		
Inbound	1xxB	Local	2	ST02	8:54:40	8:54:50		
Inbound	1xxB	Local	3	ST03	8:56:20	8:56:40		
Inbound	1xxB	Local	4	ST04	8:58:30	8:58:50		
Inbound	1xxB	Local	5	ST05	8:59:45	9:00:15		9:00:15
Inbound	1xxB	Local	6	ST06	9:01:20	9:01:40	9:01:20	9:01:45
Inbound	1xxB	Local	7	ST07	9:03:40	9:04:00	9:03:50	9:04:30
Inbound	1xxB	Local	8	ST08	9:05:40	9:06:00	9:06:00	

If an operation of train is cancelled, operated time data of that train Id will be blank. In case of turn around operations, operated time data will be recorded from an intermediate station. Extension of operation section can be detected by finding only operated time data were recorded, i.e. planned time data are blank for extended section. In addition, extra trains and change of train type can be detected by comparing another day's operated train timetable. From the preliminary analysis, we considered that it can be automatically estimated that how train rescheduling operation were performed by comparing each train Id's normally planned information and operated history. Then we start to generate normal operated train data named "regular train data" to extract train rescheduling operation.

2.3 Our goal

To develop a decision-making support tool that is enabled to recommend train rescheduling operation plan for train dispatchers, we propose two-step analysis method.

Step1: Generate a regular train timetable based on historical train operation data.

Step2: Extract differences between a disrupted day's timetable and a regular timetable.

Our objectives of this paper are as follows:

(1) To extract train rescheduling operations automatically from historical data
(2) To evaluate accuracy of the proposed method by comparing with dispatcher's manual report
(3) To demonstrate a prototype of decision-making support tool

3 Train Rescheduling Operation Extraction Method

In the subsection 3.1, we describe a data process for generating a regular train timetable. Then, in the subsection 3.2, we explain how train rescheduling operation can be extracted automatically.

3.1 Regular train timetable

First, we defined regular train timetable data schema shown in Table 4. As essential factors of a regular train timetable, it is considered that each train's property, a sequence of stopping stations, planned arrival time and planned departure time are necessary. Train property includes line name, direction, train Id, previous train Id and next train Id. In addition, based on the preliminary analysis result, operated day factor was included because it is found that some trains are operated in limited season, day of week and so on.

Table 4: Data schema of regular train timetable data

Direction	Train Id	Operated day	Station #1	Dep. Time #1	Station #2	Arr. Time #2	Dep. Time #2	---
Inbound	1xxA	weekday	ST01	8:52:30	ST02	8:54:00	8:54:30	---
Inbound	1xxA	holiday	ST01	8:50:00	ST02	8:51:30	8:51:50	---
Inbound	1xxC	weekday, only Monday	ST05	9:00:20	ST06	9:02:00	9:02:30	---
---	---	---	---	---	---	---	---	---

Data process for generating a regular train timetable is as follows.

(1) Count operated numbers for each train Id by yearly and by day of the week.
(2) Extract train Ids that meet the following conditions.
-Number of operated days is over than 2.
-Year-based operated ratio is over than 80%
(3) Set "Operated day" flag for train Ids that seems be operated in only limited day of the week.

It is necessary to generate regular train timetables at least by yearly for train operations in Tokyo metropolitan area because a large-scale train timetable revision is held in every March. We have analysed 7-years train operation data with the proposed method. Table 5 shows the result of regular timetable generation for Line A that is operated in Tokyo central area.

Approximately 800 trains were extracted as regularly operated trains for weekday. Holiday regular train numbers is less than weekday. In FY2011, the reason why number of trains operated with limited-time were large is found that impact of Great East Japan earthquake (March 11th, 2011).

Table 5: Number of regular trains on Line A

Fiscal Year	Weekday			Holiday	
	Total	Only Friday	Limited time	Total	Limited time
2011	835	2	83	649	2
2012	778	3	13	644	1
2013	777	4	3	649	0
2014	784	3	9	651	0
2015	779	1	16	659	0
2016	776	0	3	661	0

Though generating a regular train timetable requires only calendar information, i.e. it is necessary to calculate operated ratio by year, other data processing flow can be implemented automatically. Then it is expanded for new dataset easily.

3.2 Extract differences

Train rescheduling operations can be detected by comparing a disrupted day's train operation data with a regular train timetable. The rules or data processing flows to detect each train rescheduling operation are shown in Table 6. Comparison between planned time data and operated time data brings long-time train stop detection besides. Extra operated time of a train can be easily calculated by subtract planned travel time from actual travel

time. Each travel time between adjacent stations is calculated by using departure time at one station and arrival time at a next station.

Table 6: Rules for detecting train rescheduling operation

No.	Rescheduling Operation	Detection Rule or Process
1	Extra train	Not included in a regular train timetable
2	Extend operational section	Compare operational section (a pair of origin station and destination station of each train)
3	Change train type (local, rapid, express)	Compare train type
4	Change departure time of a train	Compare planned departure time
5	Cancellation (fully and partially)	Operated time are not recorded
6	Turn around operation	Until an intermediate station, operated time are blank. From the intermediate station, operated time are recorded.
7	Change train track	Operated time are not recorded from an intermediate station to another intermediate station
8	Change train id in train diagram	Compare previous train Id/next train Id
9	Change departure order at a station	Make a list of train Ids based on a departure time by a station for both a regular train timetable and a disrupted day's data. Then compare departure order.

Data extraction results of train rescheduling operations are stored with data schema as shown in Table 7. In addition to No.1~No.9 listed in Table 7, detection results of long-time train stop can be stored in the same data table. As we described in the above, the proposed method enables to automatically extract all of train rescheduling operations in Table 7. In other words, amount of extracted data increases permanently day by day.

Table 7: An example of train rescheduling operation extraction result

Date	Train Id	Property	Rescheduling No.	Rescheduling Time	Detail	---
1st Feb 2015	1xxA	---	6 (turn around)	13:22:38	Operated from ST05 to ST08	---
1st Feb 2015	1xxB	---	5 (cancellation)	13:26:10		---
1st Feb 2015	1xxC	---	1 (extra train)	13:50:23	Operated from ST03 to ST10	---
1st Feb 2015	9xA	---		13:10:50	Long-time stop at ST06 for 15mins	---
---	---	---	---	---	---	---

4 Evaluation

4.1 Extraction results

With 7 years dataset from July 2011 to April 2016, we analysed all of train rescheduling operations for Line A. Approximately 400 disruptions happens in Line A thorough 7 years. Figure 2 and Figure 3 analysis results of the two disruptions used in Figure 1.

Trains that stopped for a long-time are highlighted in Figure 2. Threshold for detecting "long-time" stopping trains used in this paper is 10 mins. Comparing two images in Figure 2, it is found that turn around operation case took longer time to recovery. With these extraction results of long-time stopping trains, we can estimate disrupted sections and timeslot easily. There is possibility of accumulating disruption information automatically instead of dispatcher's manual input.

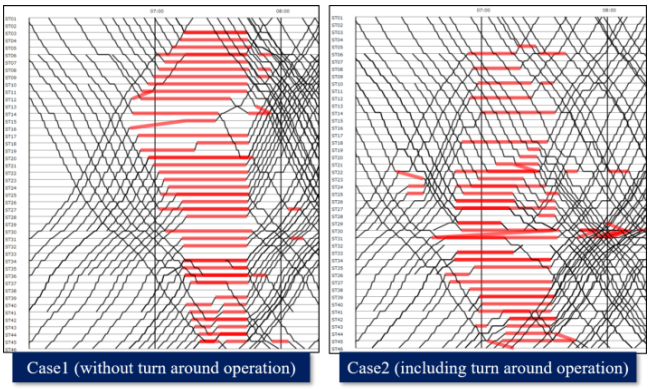


Figure 2: Detection results of long-time stop trains (same cases are used in Fig. 1)

Figure 3 shows turn around trains with blue lines. As described in Section 2, turn around operations had been performed only in Case2. Our proposed method extracted three turn operated trains as shown in Figure 3. It was confirmed that those extracted result was correct by referring daily operating report written by dispatchers.

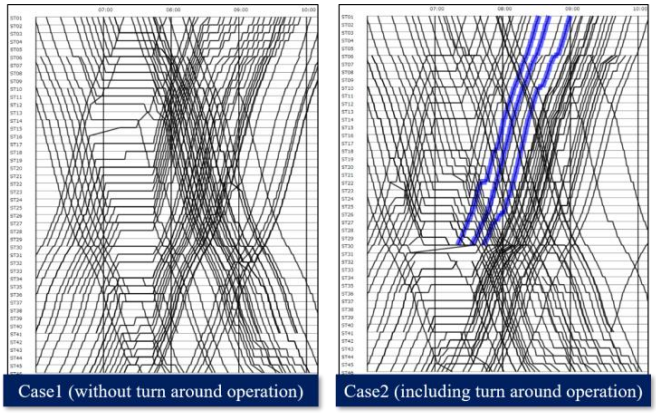


Figure 3: Detection results of turn around operated trains (same cases are used in Fig. 1)

4.2 Accuracy evaluation for turn around operation detection

To evaluate accuracy of detecting method for turn around operation, we compared extraction results with historical operating report manually input by dispatchers. Operating report includes disruption information and some of train rescheduling operations such as where turn around operations were performed as shown in Table 8. We selected 40 disruption case of Line A in which turn around operations were performed and compared operating reports with extraction results regarding turn around operation. Total number of turn around operations is 252 as for 40 disruption case. Then our proposed methods extracted 249 turn around operations. It was confirmed that operated sections, i.e. where turn around train run, were completely correct. Consequently, it is found that accuracy of the proposed method is quite high concerning turn around operations.

Table 8: Dataset of manual report for train rescheduling operation

Date	Line Name	Disruption Information	Number of turn around trains	Detail	---
1st Feb 2015	Line A	Station: ST06 Occurring Time: 10:15 Restart Time: 10:57	10	4 (at ST03, Inbound) 6 (at ST15, Outbound)	---
5th Feb 2015	Line A	Station: ST015 Occurring Time: 15:42 Restart Time: 16:38	8	2(at ST03, Inbound) 2(at ST05, Inbound) 4 (at ST20, Outbound)	
5th Feb 2015	Line B	Station: ST24 Occurring Time: 08:09 Restart Time: 09:30	5	2 (at ST20, Inbound) 3 (at ST28, Outbound)	
---	---	---		---	---

5 Application and Case Study

We have developed a dispatcher's decision-making support tool for train rescheduling using historical train operation data. Our prototype enables to recommend when and

where turn around operations should be performed based on the analysis results described in the above. The processing flow of the prototype is as shown in Figure 4.

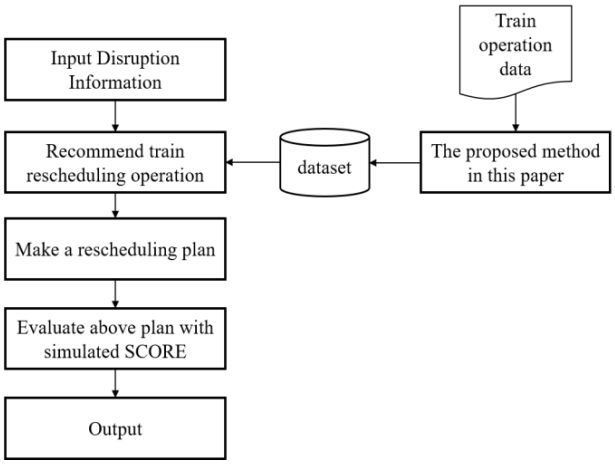


Figure 4: Framework of the proposed system

(1) Input the disruption information
An operator manually input three data, i.e. disrupted line, time of occurrence and predicted time to restart.
(2) Search a similar disruption case based on input information and recommend turn around operation performed in the past.
(3) Make a rescheduling plan (train timetable) with GUI by referring recommendation results.
(4) Evaluate a planned rescheduling timetable with SCORE simulation algorithm.
(5) An operator can confirm an evaluation result for a planned train timetable. Then he/she can continue to make a better rescheduling plan by back to step (3), if necessary.

According the framework in the above, we have implemented the prototype as shown in Figure 5. Basic functionality for making a new train timetable and evaluating it with a simulated SCORE have been already developed (Yamashiro, et al. 2017). We have improved the simulator to recommend a rough design policy for a dispatcher when he/she starts a train rescheduling work. Firstly, the prototype receives a disruption information from a dispatcher. Then the prototype searches a similar disruption among analysed results of historical train operation data and recommends number of turn around operations and at which station to be performed. As parameters of similarity, we used a timeslot of disruption occurrence (i.e. morning, daytime, evening), a service suspended time and a disrupted section. A train timetable making process itself needs some manual modification currently because the prototype doesn't take account of all constraints such as facility and human resources.

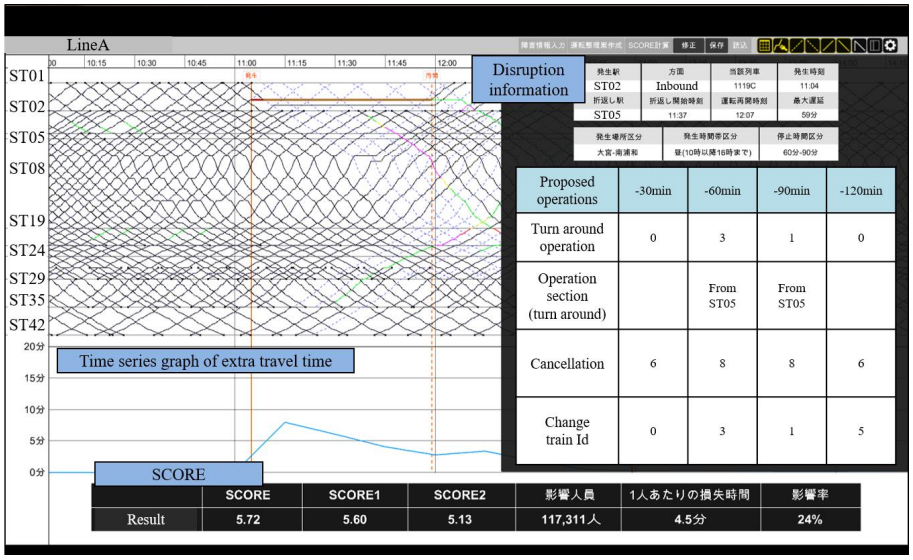


Figure 5: Overview of the prototype tool

To verify the prototype tool, we have compared the proposed timetable with operated timetable regarding the one disruption case in which turn around operations were performed. Case description is as follows.

- Disruption happened at ST05, outbound direction, at 11:05am
- Operation restarted at 11:59am
- Four turn around trains were operated at ST06 until restarting

We have input disruption information (place of a disruption happens, direction, time of occurrence, time of restart) to the prototype tool and confirmed the proposed timetable as shown in the left image in Figure 6. The prototype recommends three turn around operations at ST06. It is found that the operated station of turn around trains were correct though the number of turn around trains was slightly different. Although with a limited example, we have confirmed effectiveness and remaining issues of the prototype. Train headway and movement of each train should be considered to improve validity of the prototype.

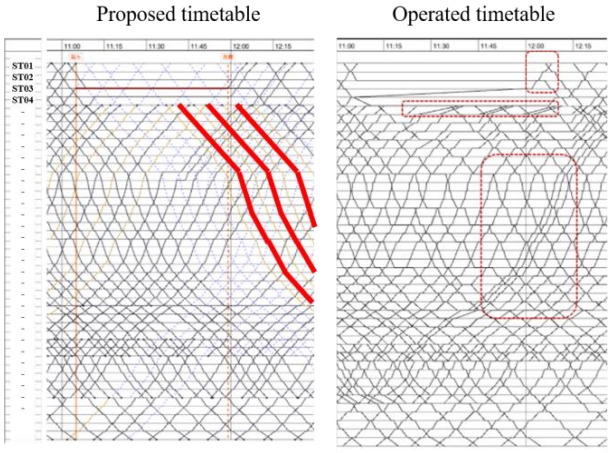


Figure 6: Comparison of proposed and operated timetables

6 Conclusions

We have proposed a new data-driven approach to support train rescheduling operation by analysing historical train operation data derived from train management system. It was confirmed that extraction accuracy as for turn around operations was sufficiently high. Furthermore, we have developed the prototype tool that recommends train rescheduling plan for dispatchers.

Future studies will focus on functional improvements of recommendation system. Prototype system should be improved to help dispatchers make more practical rescheduling plan. In addition, it is necessary to extent our extraction method for more complicated train service line including changes of train departure order.

References

Tsunoda, F., Yamashiro, M., Otsuka, R., Kato, M., Sukeda, H., Ozeki, K., 2015. "Customer-Oriented Evaluation Method of Railway Performance", In: *Proceedings of The 6th International Conference on Railway Operations Modelling and Analysis (RailTokyo2015)*, Tokyo, Japan.

Besinovic, N., Cacchiani, V., Dollevoet, T., Goverde, Rob M.P., Huisman, D., Kidd, M.P., Kroon, L.G., Quaglietta, E., Rodriguez, J., Toth, P., Veelenturf, L., Wagenaar. J.,2015. "Integrated Decision Support Tools for Disruption Management", In: *Proceedings of The 6th International Conference on Railway Operations Modelling and Analysis (RailTokyo2015)*, Tokyo, Japan.

Kunimatsu, T., Sakaguchi, T., 2016. "Evaluation of Facility Improvement for Turn-back Operations from a Passenger Service Viewpoint", *Quarterly Report of RTRI*, Vol. 57, No. 1, pp. 22-18.

Sakairi, S., Shikoda, Y., Otsuchibashi, I., Otsuka, R., 2016. "New Transport Arrangements using ICT", In: *Proceedings of the 11th World Congress of Railway Research*, Italy.

Yamashiro, M., Otsuka, R., Otsuchibashi, I., Ito, K., Sakairi, S., 2017. “A Simulator for Evaluating Customer-oriented Rescheduling for Large-scale Traini-service Disruptions”, In: *Proceedings of The 7th International Conference on Railway Operations Modelling and Analysis (RailLille2017)*, France

Shakibayifar, M., Sheikholeslami, A., Jamili, Amin., 2018. “A Multi-Objective Decision Support System for Real-Time Train Rescheduling”, *IEEE Intelligent Transportation Systems Magazine*, Vol. 10, pp. 94-109.

Huang, H., Li K, Schonfeld, P., 2018 “Real-time energy-saving metro train rescheduling with primary delay identification”, PLoS ONE 13(2): e0192792. https://doi.org/10.1371/journal.pone.0192792

Applying Geometric Thick Paths to Compute the Maximum Number of Additional Train Paths in a Railway Timetable

Anders Peterson [a], Valentin Polishchuk [a], Christiane Schmidt [a]
[a] Communications and Transport Systems, ITN, Linköping University, Norrköping, Sweden. Email: firstname.lastname@liu.se

Abstract
Railway timetabling is a prominent research area in railway research. The timetable is usually shown as a time-space diagram. However, even algorithms that try to adapt/add to an existing timetable rely mainly on mixed integer programming, but do not use the geometric representation of the timetable. In this paper, we consider the problem of determining residual train paths in an existing timetable. We aim to restrict possible disturbance on existing (passenger) traffic, and, hence, insert train paths of a specified minimum temporal distance to other trains. We show how we can extend algorithms for thick paths in polygonal domains to compute the maximum number of trains with a specified robustness to insert.

Keywords
Railway Timetabling, Residual Capacity, Polygonal Domains, Geometric Thick Paths

1 Introduction

Both passenger traffic and freight traffic volumes in Sweden significantly increased over the last 20 years—from 1996 to 2016 by 82% (from about 7 to 12.8 billion passenger kilometers) and by 23% (from about 55 to about 68 million tonne-kilometers), respectively, see Trafikanalys (2017a,b). Over the last years the freight volume transported via railway within the EU has stagnated, but both road congestion and oil prices make road transport more expensive and less attractive. In contrast, railway transport is safer and more environmental friendly. However, already with the current traffic load, railway infrastructure is often overloaded. This is particular true for marshalling yards: trains that are already completed occupy highly demanded space until their departure. To free this capacity, both the freight operator and the infrastructure manager (IM) often agree in their goal to depart ahead of schedule. Today, such a request is answered manually by looking a few stations ahead, and if the completed freight train will not interrupt operations on this limited considered stretch, an earlier departure will be permitted. This procedure hardly takes into account the already congested rail network, where freight traffic interacts with passenger traffic with high requirements on punctuality. The early departed freight train might be stuck at a sidetrack before its destination for long stretches of time due to the limited spatial and temporal horizon considered for the decision by the IM. Having several early departing freight trains only worsens this situation. Similarly, early arrivals contribute to congestion, when no track capacity is available at the destination yard.

To make sure both that the existing (passenger) traffic is not affected by the train path of the freight train and that the freight train actually obtains a feasible train path to its destination, it is essential to optimize the process. In Ljunggren et al. (2018), we proposed an algorithm that computes a maximum robust train path for inserting a single additional train (at a time). Here, we aim to determine how many additional trains with certain properties can be added to the existing timetable, that is, we aim to determine the residual capacity for additional train paths within given time windows. This could be particularly interesting for adding freight trains, but also adding passenger trains can be of interest.

Timetabling is a problem that has been extensively studied, in the majority a new timetable, or a larger part of it, is constructed from scratch, see, e.g., Hansen and Pachl (2014); Liebchen (2008) for an overview.

Adding a new train to an existing timetable was considered, e.g., by Burdett and Kozan (2009). Flier et al. (2009)(see also Flier (2011)) present a shortest path model using a time-expanded graph, which integrates linear regression models based on extensive historical delay data, that gives Pareto optimal train paths w.r.t. travel time and risk of delay. Ingolotti et al. (2004) consider adding new trains to a heterogeneous, heavily loaded railway network, and aim to minimize the traversal time for each additional train. Cacchiani et al. (2010) also consider the problem of inserting a single freight train into an existing schedule of fixed passenger trains. They assume that the operator specifies an ideal timetable that the IM can modify, which also includes the use of a different path. Cacchiani et al. aim to add the maximum number of new freight trains, such that their timetable is as close as possible to the ideal one. To do so, they use a heuristic algorithm based on a lagrangian relaxation of an Integer Linear Program (ILP).

UIC (2004) has developed a compression technique for computing capacity utilization. This technique is widely used for assessing capacity utilization in the railway network. For example the Swedish infrastructure manager routinely makes an annual report about the network congestion (Trafikverket (2018)). The corresponding analysis for year 2011 has also been presented in English, see Grimm (2012). The UIC 406 compression technique is an easy and effective way of estimating the capacity consumption, but it is possible to expound it in different ways leading to different estimates. Landex et al. (2006), who explain how the method has been implemented in Denmark, show the importance of choosing the right length of the line sections and examine how line sections with multiple tracks are considered. Also Lindner (2011) discusses some aspects of this problem. In particular, the UIC 406 code calculates a capacity consumption, that is, it evaluates how much of the available capacity is consumed by the existing traffic. It first compresses the timetable, that is, the existing train paths on the considered line section are shifted as close together as possible. At this stage, they represent trains running within a certain time interval, but no longer are considered during the actual time they occupy the line. After this shifting certain blocking time elements and indirect occupancies are integrated to obtain the capacity utilization. The extension of the compression technique given in the second edition (UIC (2013)) mainly concerns how the method can be applied at station areas and in complex nodes. For the aggregated type of analysis in an annual report, we believe that the UIC compression technique is well suited. However, for determining the actual number of executable train paths between two nodes, which can be added to the existing timetable, the method is insufficient. In this paper, we address the problem to obtain as many train paths that such a network part still allows. Moreover, we allow a trade-off between adding further trains and influencing the existing trains as little as possible: the required temporal distance

to the existing train is an input parameter to our computation. When we use the minimal required temporal distance we can add more trains than with a larger temporal distance, however, this comes at the price of a higher impact on other trains. Additionally, we may add train paths over topological different routes.

Pellegrini et al. (2017) and Lucchini et al. (2001) considered the saturation problem: an existing (possibly empty) timetable and a set of saturation trains are given, and the goal is to add as many trains to the timetable as possible. Lucchini et al. (2001) use the CAPRES method—in which stations and junctions are modeled as a graph on which a constraint program is solved—to determine how many freight trains can run on the North-South rail corridor in Switzerland. Pellegrini et al. (2017) used a MILP approach, . In the saturation problem, various train types (possibly with number of trains per type) are considered, while we assume a specific type, but aim at disturbing the passenger traffic as little as possible, and obtain a trade-off with the temporal distance to other trains. CAPRES uses heuristics, Pellegrini et al. (2017) output the best feasible solution found until a time limit is reached, while we present an optimal solution.

A timetable is usually shown as a time-space diagram. However, even when we only aim at inserting something into an existing timetable, or make some limited adaptations to it, this geometric representation is not used in algorithms (while it is used in the practical, mainly manual, process). We present a roadmap on how we will make use of this geometric representation in Section 2. There exist various results on thick paths and flows within a polygonal domain, we present basic definitions and the results important for this paper in Section 3. In Section 4 we describe our general approach, and detail in Subsection 4.1 how we construct our polygonal domain, in Subsection 4.2 how to extend the path computation to our needs, and in Subsection 4.3 how we combine these to compute the maximum number of additional trains.

2 Roadmap for Our Strategy

We aim to insert additional trains to a given timetable, where we consider the existing trains as fixed. When we consider the time-space diagram of the given timetable (where we consider time on the x-, and space on the y-axis), inserting new trains means to route paths from their start to their end station. However, these paths cannot be arbitrarily close to each other: we need to keep a certain temporal distance to consecutive trains on any track. Let d_s and d_o denote this temporal distance for trains running in the same and for trains running in the opposite direction as the trains to be inserted, respectively (these values may coincide, but will usually not). So, instead of thinking of the existing trains as line segments in the time-space representation of the timetable, we can think of them as "blown-up" line segments (blown up by the temporal distance), that is polygons. Similarly, the trains that we route are not just curves in $\mathbb{R}^2$, but *thick paths*, where the thickness represents the temporal distance we need to keep to neighboring trains, d. For our algorithm, we can choose d according to the minimal necessary temporal distance between trains, or—with the motivation of disturbing the existing trains as little as possible, e.g., because we insert freight trains shortly before operation, see Ljunggren et al. (2018)—we can choose a larger value.

We will use concepts from Computational Geometry that allow routing such thick paths. Of course, if we would just "blow up" all existing trains, no thick paths could overtake any other train at a station, as the line segments representing the stations and the existing

trains would constitute obstacles. To enable such options, we will need to make certain adaptations to the "blown up" time-space representation, we show how to construct the appropriate polygonal domain for our problem in Subsection 4.1. We are given a time-window for possible departure on the start station and a time-window for possible arrival at the end station, these will coincide with special edges, the *source* and *sink*, of the polygonal domain, between which we need to route the thick paths.

To determine the maximum number of trains that we can insert into a timetable, we then need to determine the maximum number of thick paths that we can route in that polygonal domain. However, we do not want to route arbitrary thick paths, for example, paths that are parallel to the y-axis, would mean that our trains run with infinite speed. We are given a maximum speed, and this limits the slope of the feasible train paths. If we denote time along the x-axis, we of course aim for x-monotone paths (as we should not allow our trains to go back in time, implementing none-x-monotone paths will result in definite problems). Hence, we aim for thick (non-crossing) x-monotone paths of a limited slope.

Polishchuk (2007) presented an algorithm to compute the maximum number of (x-) monotone thick non-crossing paths, we will describe this algorithm in Section 3.

We need to extend the algorithm for x-monotone thick paths to compute thick paths of a limited slope, see Subsection 4.2. We will then combine this general algorithm (Subsection 4.2) and the constructed polygonal domain (Subsection 4.1) to determine the maximum number of additional train paths in Subsection 4.3.

3 Routing a Maximum Number of Thick Paths through a Polygonal Domain

Various authors studied maximum flows in geometric domains Hu (1969); Hu et al. (1992); Strang (1983); Mitchell (1990); Eriksson-Bique et al. (2014), Mitchell (1990) presented efficient algorithms for computing maximum flows in polygonal domains. Here, we are, however, interested in routing thick paths through a domain, and not a flow, see Polishchuk (2007). A *thick path* is the Minkowski sum of a "normal" (zero-thickness, or "thin") path and a disk (for some metric). Hence, we try to wire a maximum number of "threads" (of a specific thickness) between given edges of the domain. The domain is usually given by a polygon. We detail the necessary notation in Subsection 3.1, and present a polynomial-time algorithm by Arkin et al. (2010) (see also Polishchuk (2007)) to compute the maximum number of thick paths in Subsection 3.2.

3.1 Notation

We use the notation given by Polishchuk (2007). A polygon can either be simple, or contain holes. Both are possible inputs: We are given a polygonal domain, Ω, defined by the outer simple polygon, P, and a set $\mathcal{H}$ of h holes $H_1, \ldots, H_h$ within P. (In case of a simple polygon, we have $\mathcal{H} = \emptyset$.) For any set $Q \subset \mathbb{R}^2$ we let $\delta(Q)$ denote its boundary; if $Q \neq \delta(Q)$, that is, if Q has interior points, we will assume that Q is open (i.e., $\forall p \in \delta(Q), p \notin Q$).

Two edges on $\delta(P)$ are specifically marked: they are the *source* and the *sink* in our domain—that is, the edges from and to which we want to route the thick paths. We denote them by Γ_s and Γ_t. $\delta(P) \setminus (\Gamma_s \cup \Gamma_t)$ has two connected components, the "top" T, and the "bottom" B.

A "thin" (normal) path π is a simple curve. For $r > 0$ we let $\mathcal{C}_r$ denote the open disk of radius r centered at the origin; we use $\mathcal{C} = \mathcal{C}_1$. For a set $S \subset \mathbb{R}^2$ we let $(S)^r$ denote the Minkowski sum $S \bigoplus \mathcal{C}_r$, with $S \bigoplus \mathcal{C}_r = \{x + y | x \in S, y \in \mathcal{C}_r\}$. A *thick path* is the Minkowski sum of a "thin" (normal) reference path, that is, a curve in $\mathbb{R}^2$, and a unit disk (or, more general, a disk of radius r): a thick path Π with reference path π is the Minkowski sum of π and $\mathcal{C}$, that is, $\Pi = (\pi)^1$, such that Π does not intersect P's exterior.

3.2 Computing the Maximum Number of Thick Paths

We now aim to find the maximum number of thick paths from Γ_s to Γ_t, where the paths should avoid all the obstacles (holes) and be non-crossing, that is, for any wo paths Π_i, Π_j we have $\Pi_i \cap \Pi_j = \emptyset$. This requires the interiors of paths to be disjoint, thick paths may share boundary.

No path can run outside Ω. We add B and T to the set of holes $\mathcal{H}$, that is, $H_0 = T, H_{h+1} = B$. Arkin et al. (2010) used the concept introduced by Mitchell (1990), and we follow this idea: we assume that Ω has been "perforated" at Γ_s and Γ_t, and that Riemann flaps were glued to Ω at these two perforated edges. This circumnavigates complications from a thick path protruding through Γ_s and Γ_t.

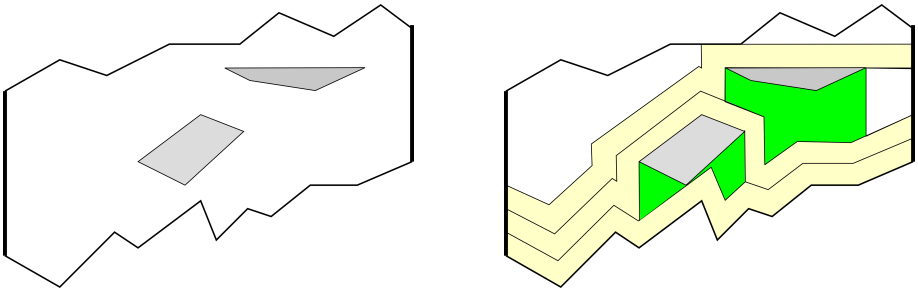
Arkin et al. (2010) showed that the maximum number of x-monotone thick non-crossing paths can be found in $\mathcal{O}(nh + n \log n)$. The routing of these thick non-crossing paths, or wires, is different to just routing the maximum number of self-overlapping thick paths (for example, there exist domains in which only the latter exist at all).

A single self-overlapping thick path from Γ_s to Γ_t that avoids all obstacles can be found by first building the offset of all holes by 1 (that is, building the Minkowski sum $(H_i)^1 \forall i$), and then solving the "usual" shortest path problem in the presence of these new, offset obstacles. The path found by this procedure yields the reference path for an optimal thick path Liu and Arimoto (1995); Chen et al. (2001). This does not translate to the case of wires (non-crossing thick paths).

The idea of the algorithm by Arkin, Mitchell and Polishchuk is to use an adaptation of the so-called "grass fire" analogy from Mitchell (1990): the free space $\Omega \setminus \mathcal{H}$ is grass over which fire travels at speed 1. All the holes are highly flammable, that is, once they are ignited, the fire moves through them with infinite speed. We start setting the bottom on fire. The *wavefront* at time τ is the boundary of the burnt grass by time τ. Whenever the fire has not hit a hole after burning for 2 time units, we can route a thick path through the burnt grass: Arkin et al. showed that a thick path does exist when the fire has burnt for 2 time units, and that routing a thick path at that point does not hamper the construction of any of the other paths in a maximum set of such paths. Once a thick path has been routed, we use the wavefront as the new bottom, and start over.

If a hole H is hit after $\tau < 2$, Arkin et al. define e to be the segment of length τ connecting H to B (or, to T, as they started the fire at the top, we will use the bottom in our application and adapted the description accordingly). We split the free space along e and around $\delta(H)$ (thus, the hole is no longer a hole), and glue a Riemann sheet to each copy of e, where we place a circular segment of radius 2 with e as chord in each. Then we continue the grass fire by igniting H and a belt of thickness τ around it, as the fire flips to the other side of H and runs there.

Polishchuk and Mitchell (2007) proved a continuous version of the discrete network Flow Decomposition Theorem, the Continuous Flow Decomposition Theorem (CFDT),



(a) (b)

Figure 1: (a) A polygonal environment with two obstacles (gray), and the source and sink shown in bold. (b) shows the wavefronts after 2 time units each (which induce the x-monotone paths). Waterfalls are depicted in green.

which states that the support of a minimum-cost flow can be decomposed into a set of thick paths; the size of the decomposition is linear in the size of the description of the flow (Theorem 5.5. in Polishchuk (2007)). This enables the proof that the above algorithm actually routes the maximum number of thick non- crossing paths. Its runtime is $\mathcal{O}(nh + n \log n)$.

Monotone Thick Paths Polishchuk also aimed for x-monotone thick paths, where a thick path Π is x-monotone if its reference path π is x-monotone (each vertical line intersects π in at most one point). Each x-monotone thick paths is a monotone simple polygon.

To compute the maximum number of monotone thick non-crossing paths, Polishchuk extended the algorithm for the maximum number of thick non- crossing paths. First, we need a monotone $\delta(P)$. Hence, we need to add "waterfalls" (following notation from Arkin et al. (1989)): the inner x-monotone hull of P is the largest x-monotone polygon that is contained in P; to compute it (see Polishchuk (2007)), we can sweep a vertical line in x direction, for every vertex $v \in P$, connect v to the first point of P hit when going up from v, and when going down. Whenever we hit a hole (with the wavefront of our burning fire), we use the waterfalls to outer-monotonize the hole. See Figure 1(a) for an exemplary polygonal domain, and Figure 1(b) for the waterfalls of the obstacles (note that P was already monotone) and the maximum number of thick paths.

4 Inserting a Maximum Number of Trains in a Timetable

We consider the existing trains as fixed, are given a time-window for possible departure on the start station and a time-window for possible arrival at the end station, and a maximum speed for the trains to be inserted. Moreover, we are given a minimum temporal distance, d, that we need to keep between consecutive inserted trains on any track.

The idea is that thick paths through the timetable reflect train paths with a certain temporal distance to neighboring trains. That is, we still use the algorithms with the "grass fire" analogy and route trains when the fire burnt without hitting obstacles—given by other trains–for d time units. However, to be able to do so, we need to complete two steps:

1. The timetable is given by line segments for trains and stations, we need to make

certain adaptations to generate our polygonal domains: stations are not obstacles and we need to remove these lines; we need to keep a required safety distance to the existing trains, hence, we will extend the line segments to polygonal obstacles, such that a train path passing this new obstacles keeps the safety distance to the existing train, et cetera. We describe the construction in Subsection 4.1.

2. The algorithm presented by Arkin et al. (2010) is for x-monotone paths. For example, applying it to our problem, this would allow for paths parallel to the y-axis, that is, our trains would run with infinite speed. We are given a maximum speed, and this limits the slope of the feasible train paths. Hence, we need to extend the algorithm for x-monotone thick paths to compute thick paths of a limited slope, see Subsection 4.2.

Finally, we can combine this general geometric algorithm with our specific polygonal domain to compute the maximum number of trains in Subsection 4.3.

4.1 Construct Polygonal Domain from Timetable

We are given: a starting station s_0 and an end station s_M for the trains to be inserted; time windows $w_s = [w_s^a, w_s^e]$ for earliest arrival and latest departure of the trains at station s for all, or some of, the stations; the train-specific running times $t_{i,i+1}$ for the trains from station i to $i+1$ $\forall i \in \{0, \ldots, M-1\}$, given by a maximum possible speed (defining the maximum slope for our thick paths); the timetable of all trains in the set $\mathcal{T}$: all trains that run in $[w_0^a - \varepsilon_1, w_M^e + \varepsilon_2]$, where ε_i is defined such that the trains that depart before or arrive after a possible path for new trains at any station are included; the required temporal safety distances d_s, d_o between any other train τ and inserted trains. For $s = 0$ the time window describes all possible departure times from the origin, and for $s = M$ the time window describes all possible arrival times at the destination. A time window at an intermediate station may also be given, e.g., due to staff schedule or wagon coupling/uncoupling, here we concentrate on the case with time windows at $s = 0, s = M$, the other case can easily be integrated in the construction, by using the algorithm between any consecutive time windows, given that the intermediate station with a time window has enough side tracks. The minimum number of train paths over all consecutive sections will determine the total number of additional train paths to be inserted.

The following construction of the polygonal domain Ω depends on d and the cone for the allowed slope (that is, different values will result in different polygonal domains for the same timetable). See Figure 2 for an exemplary construction, the time-space diagram of the considered timetable is given in Figure 2(a) (where the bold lines denote the given departure and arrival time windows).

1. **Extend the time windows by $\frac{d}{2}$ to both sides to create Γ_s and Γ_t, let $\Gamma_s = [p_1, p_2], \Gamma_t = [p_3, p_4]$, see Figure 2(b).** The train path, our reference path π, can depart anyway in $w_0 = [w_0^a, w_0^e]$ and arrive anyway in $w_M = [w_M^a, w_M^e]$, if we cut our polygon at the end of edges that represent these time windows any thick path would be restricted to that interval, hence, the train path could only depart in $w_0 = [w_0^a + d/2, w_0^e - d/2]$ and arrive in $w_M = [w_M^a + d/2, w_M^e - d/2]$. This would not be the correct solution, hence, we add $\frac{d}{2}$ to both ends of the time windows to create our source and sink edge.

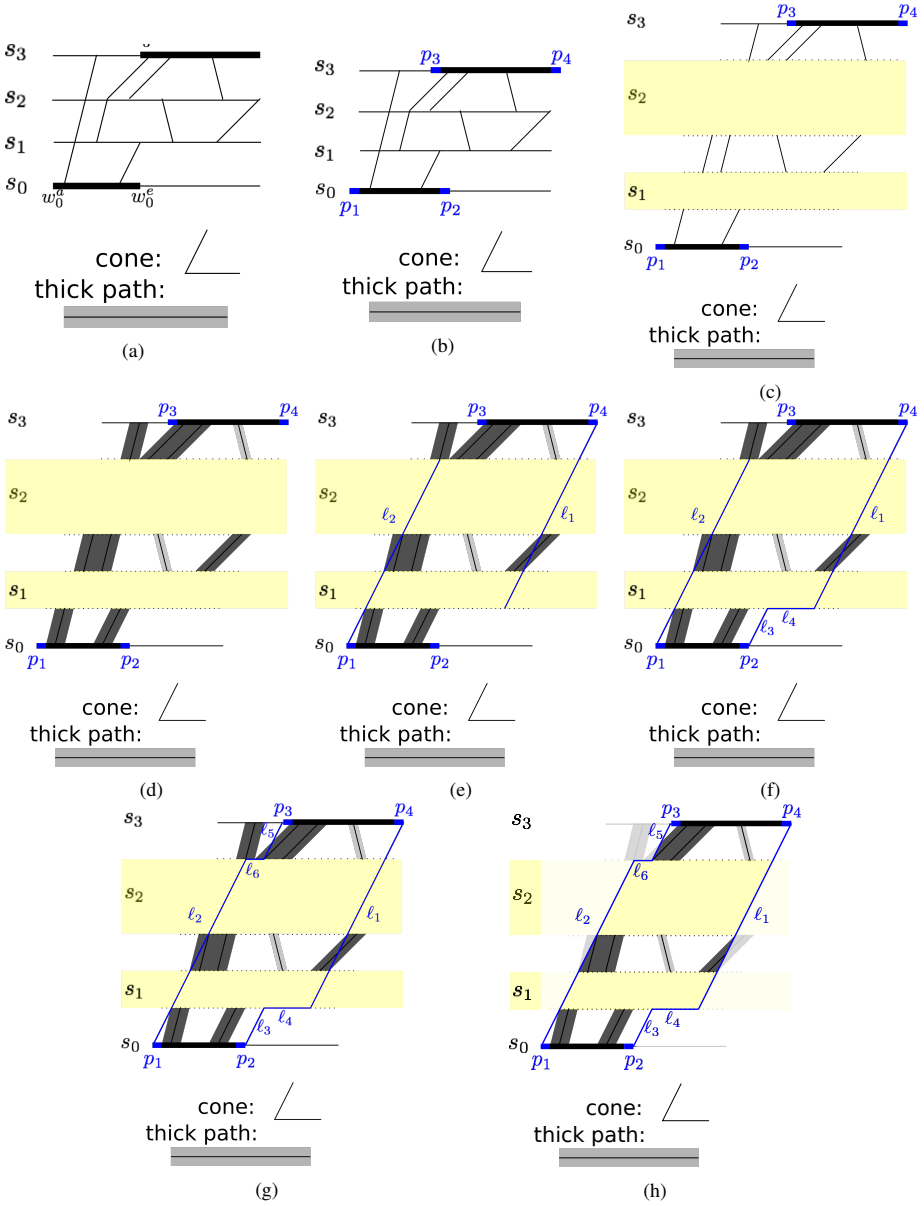


Figure 2: Example for the construction of the polygonal domain. The dotted lines and yellow blocks are given just for visual help and are not present in the domain. The width of the thick paths is shown in light gray below the diagrams. (a) Time-space diagram for the timetable we consider with given time windows $w_0 = [w_0^a, w_0^e]$ and $w_M = [w_M^a, w_M^e]$, $M = 3$. (b) Extension of the time windows by $\frac{d}{2}$ to both sides to create Γ_s and Γ_t (the bold blue and black line segments together constitute Γ_s and Γ_t). (c) "Cut open" (intermediate) stations s_1, s_2, insert a vertical distance (yellow): for s_1 we assume two side tracks, for s_2 no such limit exists. Moreover, the stations are shifted horizontally according to the procedure described in Figure 3. (d) Construction of the set of potential holes $\mathcal{H}_p$ by inserting the security distance (d_s, d_o) around the existing trains (d_s is shown in dark gray, d_o in light gray). (e) Insert ℓ_1, ℓ_2, (f) insert ℓ_3, ℓ_4, (g) insert ℓ_5, ℓ_6. (h) The polygonal domain Ω obtained by intersecting the potential holes from $\mathcal{H}_p$ with $\delta(P)$ to construct the holes $\mathcal{H}$, and P.

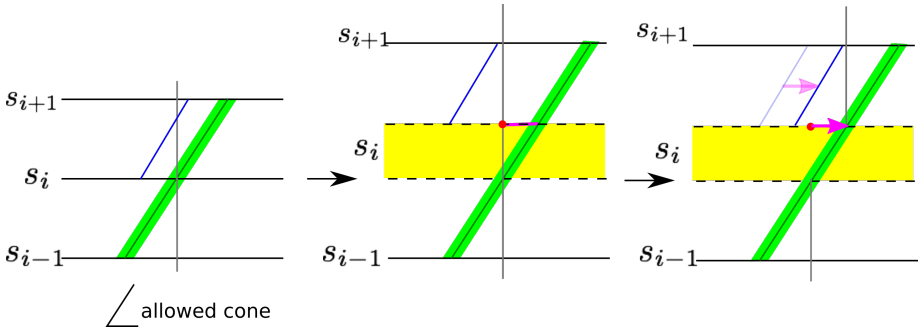


Figure 3: Time-space diagram with three stations s_{i-1}, s_i, s_{i+1}. As we cut the diagram open in step 2 (insert the yellow vertical distance), and we have a limited slope for allowed paths, we also need to shift consecutive stations horizontally. The green train arrives and departs s_i at the point in time denoted by the gray vertical line. With the added vertical distance, the red point of departure cannot be reached, hence, all stations above are shifted by the pink distance.

2. **"Cut" the diagram "open" at intermediate stations ($s_1, \ldots, s_{M-1}$), delete the vertical line for the station, insert an appropriate vertical distance ("blow up" each station) and shift the next stations according to the inserted vertical distance, see Figure 2(c).** Consider Figure 3: If we keep the stations as is, the lines would block any trains, and no train can stay at a siding at a station (and hence switch from the "left" of a train to the "right" of a train). To allow this, we cut each station open, and insert a vertical distance between arrival at station s and departure from station s (shown in yellow in Figure 3). If the station s has exactly k sidetracks, we insert a vertical distance of $k \cdot d$, if no such limit exists, we can insert a vertical distance of $\min\{|\Gamma_s|, |\Gamma_t|\}$ (which would allow the maximum possible number of additional trains to stay at a station). For the case of k sidetracks, if one or several tracks is occupied, we insert a height d rectangles to block the according height. Just inserting a vertical distance is not enough: We shift consecutive stations horizontally, as we do not allow the paths to run in parallel to the y-axis, which they would have to do when just passing through a blown-up station. So, we shift the consecutive station to the right, such that this path can be reached with limited slope. Our new trains, one of which is shown in green in Figure 3 has a limited velocity, which for us translates to the limited slope of our path. The green train arrives at station s_i at the point in time denoted by the gray vertical line. However, when we insert the yellow vertical distance, the limited slope path cannot reach the same point in time at which it should depart s_i, marked by a red point. Hence, we need to shift all stations above, and all existing trains departing s_i by the pink distance in Figure 3.

3. **Insert the security distance (d_s, d_o) around the existing trains, that is, construct the potential holes $\mathcal{H}_p$, see Figure 2(d) (for the example, we use $d_s = d$, and $d_o = \frac{d}{2}$ and denote these in dark and light gray, respectively).**

4. **Construct a line segment of maximal slope from p_4 down to s_1, ℓ_1, and from p_1 up to s_{M-1}, ℓ_2, see Figure 2(e).**

5. **Construct a line segment of maximal slope from p_2 up to s_2, ℓ_3, and a horizontal**

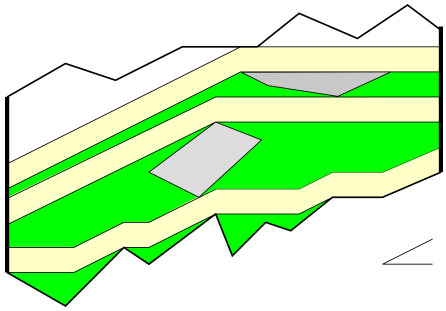


Figure 4: Paths with a limited slope that are restricted to directions in the cone shown in the lower right corner. Waterfalls are depicted in green, the wavefronts after d time units each (which induce the $\mathcal{C}$-respecting paths) are shown in light yellow. The polygonal domain is the domain from Figure 1(a).

line segment ℓ_4 from the end of ℓ_3 to the intersection with ℓ_1, see Figure 2(f).

6. **Construct a line segment of maximal slope from p_3 down to s_{n-1}, ℓ_5, and a horizontal line segment ℓ_6 from the end of ℓ_5 to the intersection with ℓ_2, see Figure 2(g).**

7. **The edges $\Gamma_s, \Gamma_t, \ell_1, \ldots, \ell_6$ define $\delta(P)$.** If we would have no other trains, we could route any thick path within the constructed polygon P.

8. **Intersect the potential holes from $\mathcal{H}_p$ with $\delta(P)$, the part of the potential holes in the interior of P determines the holes $\mathcal{H}$, together with P they constitute Ω, the polygonal domain, see Figure 2(h).**

4.2 Thick Paths with Limited Slope

If we think of our train paths (with temporal buffer around them) as thick paths, we do not just aim for thick paths, but for a path with a limited slope, that is, within a cone limited by the x-axis and a line somewhere between the x- and the y-axis.

We show that we can adapt the waterfall construction from Subsection 3.2, shown in Figure 1, to restrict our paths to the given cone, see Figure 4. Note that, while P in the example is already x-monotone, we need to use waterfalls from the bottom to make the first path feasible.

Let $\mathcal{C}$ be a cone limited by the x-axis and a line somewhere between the x- and the y-axis. We call a thick path $\mathcal{C}$-respecting if its reference path π is $\mathcal{C}$-respecting, i.e., if every line that is orthogonal to a half-line within $\mathcal{C}$ intersects π in at most one point. Accordingly, we call a flow $\mathcal{C}$-respecting if each of its streamlines is $\mathcal{C}$-respecting.

We extend the algorithm from Polishchuk for monotone thick path to find $\mathcal{C}$-respecting paths: First we make B $\mathcal{C}$-respecting: We use a different type of waterfalls than Polishchuk (2007), Arkin et al. (2010) and Mitchell (1990). First, we sweep a horizontal line in the y direction. For every vertex $v \in B$, we connect v to the first point of P hit when going right from v. Additionally, we sweep a line of the maximum slope in $\mathcal{C}$ orthogonally to this direction. For every vertex $v \in B$, we connect v to the first point of P hit when going left from v. After this procedure B is $\mathcal{C}$-respecting.

Then we run the shortest path maxflow algorithm from Mitchell (1990) to fill the free space with flowlines. As the new bottom B is $\mathcal{C}$-respecting, all flowlines are also $\mathcal{C}$-respecting.

When a hole is hit by a wavefront, we make the hole outer-$\mathcal{C}$-respecting. Essentially, this means that we outer-monotonize a hole w.r.t to two directions, see Arkin et al. (1989) for efficient algorithms for monotonization of the holes. If a waterfall during this process hits another hole, this holes is also made outer-$\mathcal{C}$-respecting. We assign the wavefront and the boundaries of the new, outer-$\mathcal{C}$-respecting holes to the new bottom B. We the make the new bottom $\mathcal{C}$-respecting and continue the grass fire. See Figure 4 for an example of this process.

Theorem 4.1. *A representation of the maximum number of $\mathcal{C}$-respecting thick non-crossing paths can be found in $\mathcal{O}(nh + n \log n)$ time.*

4.3 Maximum Number of Trains

Now, we can compute the maximum number of trains to be inserted into a timetable by computing the maximum number of $\mathcal{C}$-respecting thick non-crossing paths in the polygonal domain constructed in Subsection 4.1. See Figure 5 for an example for the construction of the maximum number of $\mathcal{C}$-respecting thick non-crossing paths for the polygonal domain constructed in Figure 2, and the resulting maximum number of train paths.

Let s denote the number of stations in our domain, and t the number of existing trains. We have $\mathcal{O}(ts)$ holes, and $\mathcal{O}(ts)$ vertices (because we have at most four vertices per train in between two stations). This yields (using Theorem 4.1):

Corollary 4.1.1. *A representation of the maximum number of train paths can be found in $\mathcal{O}(t \cdot s \cdot t \cdot s + t \cdot s \cdot \log(ts)) = \mathcal{O}(t^2 s^2)$ time. (Or, if we consider the number of times some train departs from some stations, x, in $O(x^2)$.)*

Paths of Different Thicknesses. Note that if we want to compute paths of different thickness, that is, train paths with different temporal buffers, we can use the same algorithm if the order of the trains is given. If the order of the paths is not given, Kim et al. (2012) showed the problem to be NP-hard.

5 Conclusion

We showed how to convert the time-space diagram of a timetable into a polygonal domain Ω, such that finding a maximum number of $\mathcal{C}$-respecting thick non-crossing paths in Ω gives the maximum number of additional trains that can be inserted into the timetable. To compute this, we extended a known algorithm to compute the maximum number of x-monotone thick non-crossing paths to an algorithm that can compute the maximum number of $\mathcal{C}$-respecting thick non-crossing paths in a polygonal domain. In general, this provides an application of using the geometric representation of a timetable with a geometric algorithm. In the future, it would of course be interesting to study what other geometric concepts could be extended to this geometric representation, that is, which railway problems could be solved using geometric algorithms. Moreover, future work will include the application of our algorithm to real-world scenarios. This includes the more general problem of converting headway based capacity measures on macro level to real blocking times.

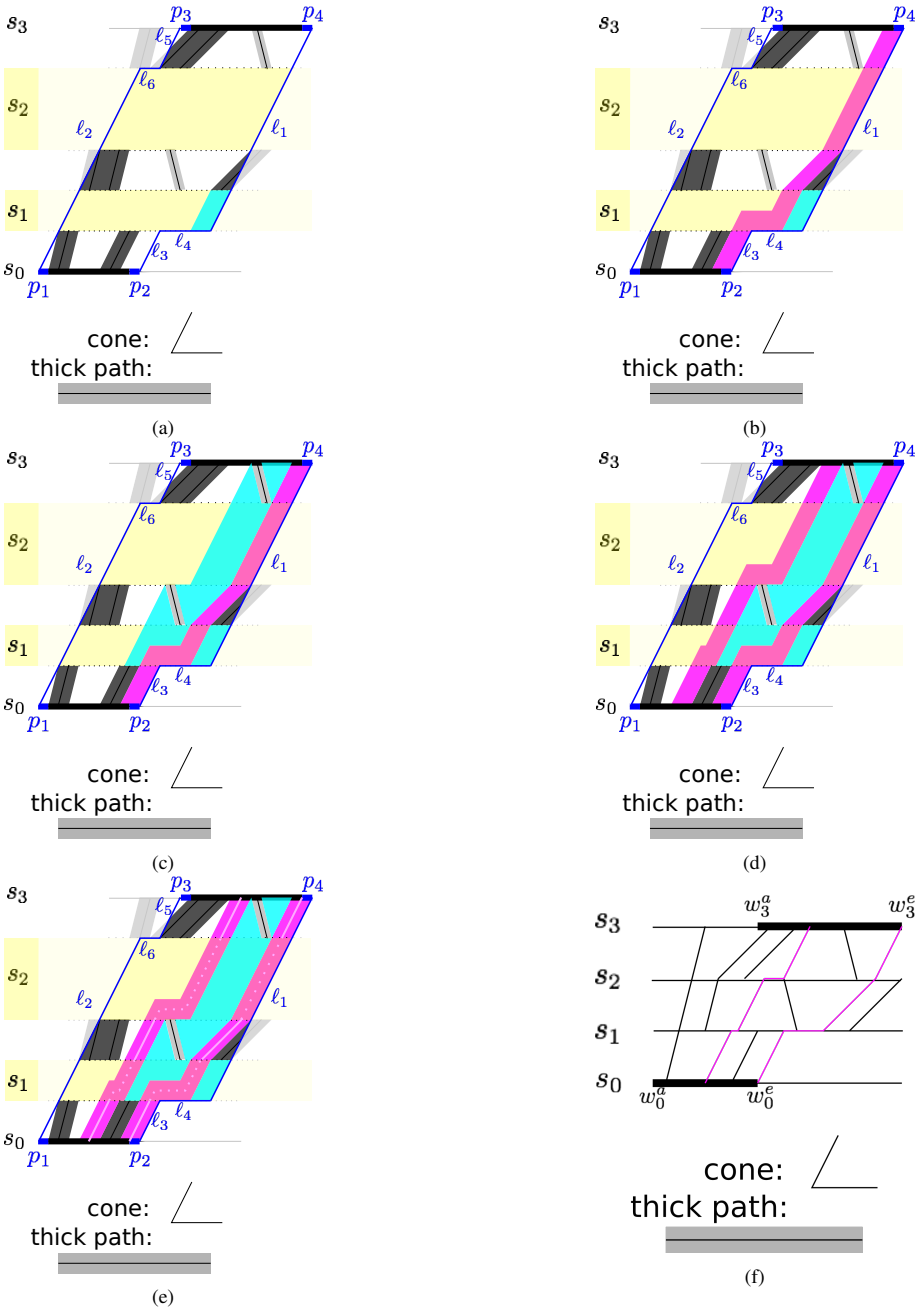


Figure 5: Example for the construction of thick paths with limited slopes for the polygonal domain constructed in Figure 2. Waterfalls are shown in turquoise, the wavefront covered after d time units each in pink. (a)-(d) Waterfalls and wavefronts/thick path construction. (e) The light pink line is the (thin) reference path (the dotted part will not be considered, as the blown-up stations are not part of the original timetable). (f) The obtained train paths inserted in the original timetable Figure 2(a).

Acknowledgements
This research is a result of a collaboration between Linköping University and Trafikverket, and part of the EU H2020 project Shift2Rail/FR8Hub (Grant Agreement No. 777402), and partially funded by Trafikverket (Dnr TRV 2016/75881). The authors are grateful to Magnus Wahlborg and Fredrik Lundström (Trafikverket) for helpful discussions.

References

E. M. Arkin, R. Connelly, and J. S. B. Mitchell, 1989. On monotone paths among obstacles with applications to planning assemblies. In *Proceedings of the Fifth Annual Symposium on Computational Geometry, Saarbrücken, Germany, June 5-7, 1989*, pages 334–343.

E. M. Arkin, J. S. B. Mitchell, and V. Polishchuk, 2010. Maximum thick paths in static and dynamic environments. *Comput. Geom.*, 43(3):279–294.

R. Burdett and E. Kozan, 2009. Techniques for inserting additional trains into existing timetables. *Transportation Research Part B: Methodological*, 43(8):821 – 836.

V. Cacchiani, A. Caprara, and P. Toth, 2010. Scheduling extra freight trains on railway networks. *Transportation Research Part B: Methodological*, 44(2):215 – 231.

D. Z. Chen, O. Daescu, and K. S. Klenk, 2001. On geometric path query problems. *Internat. J. Comput. Geom. Appl.*, 11(6):617–645.

S. D. Eriksson-Bique, V. Polishchuk, and M. Sysikaski, 2014. Optimal geometric flows via dual programs. In *30th Annual Symposium on Computational Geometry, SOCG'14, Kyoto, Japan, June 08 - 11, 2014*, page 100.

H. Flier, 2011. Optimization of railway operations: Algorithms, complexity, and models.

H. Flier, T. Graffagnino, and M. Nunkesser, 2009. Scheduling additional trains on dense corridors. In *8th International Symposium on Experimental Algorithms (SEA 2009), Dortmund, Germany, June 4-6, 2009*, pages 149 –160.

M. Grimm, 2012. The analysis of congested infrastructure and capacity utilisation at Trafikverket. In *8thWIT Transactions on the Built Environment*, 127:359-367.

I. A. Hansen and J. Pachl, 2014. *Railway Timetable & Traffic: Analysis - Modelling - Simulation*. Eurailpress in DVV Media Group, 2nd edition edition.

T. Hu, 1969. *Integer programming and network flows*. Addison-Wesley, Reading, MA.

T. C. Hu, A. B. Kahng, and G. Robins, 1992. Solution of the discrete plateau problem. *Proceedings of the National Academy of Sciences*, 89(19):9235–9236.

L. Ingolotti, F. Barber, P. Tormos, A. Lova, M. A. Salido, and M. Abril, 2004. *An Efficient Method to Schedule New Trains on a Heavily Loaded Railway Network*, pages 164–173. Springer Berlin Heidelberg, Berlin, Heidelberg.

J. Kim, J. S. B. Mitchell, V. Polishchuk, S. Yang, and J. Zou, 2012. Routing multi-class traffic flows in the plane. *Comput. Geom.*, 45(3):99–114.

A. Landex, A.H. Kaas, B. Schittenhelm, and J. Schneider-Tilli, 2006. Practical use of the UIC 406 capacity leaflet by including timetable tools in the investigations. *WIT Transactions on The Built Environment*, 88:643-652.

C. Liebchen, 2008. The first optimized railway timetable in practice. *Transportation Science*, 42(4):420–435.

T. Lindner, 2011. Applicability of the analytical uic code 406 compression method for evaluating line and station capacity. *Journal of Rail Transport Planning & Management*, 1(1):49–57.

Y.-H. Liu and S. Arimoto, 1995. Finding the shortest path of a disc among polygonal

obstacles using a radius-independent graph. *IEEE Trans. Robot. Autom.*, 11(5):682–691.
F. Ljunggren, K. Persson, A. Peterson, and C. Schmidt, 2018. Maximum robust train path for an additional train inserted in an existing railway timetable. In *CASPT 2018*.
L. Lucchini, A. Curchod, R. Rivier, 2001. Transalpine rail network: a capacity assessment model (CAPRES). In *Swiss Transport Research Conference (STRC) 2001*.
J. S. B. Mitchell, 1990. On maximum flows in polyhedral domains. *J. Comput. Syst. Sci.*, 40(1):88–123.
P. Pellegrini, G. Marlière and J. Rodriguez, 2017. RECIFE-SAT: A MILP-based algorithm for the railway saturation problem In *Journal of Rail Transport Planning & Management*, 7(1):19–32.
V. Polishchuk, 2007. Thick non-crossing paths and minimum-cost continuous flows in polygonal domains.
V. Polishchuk and J. S. B. Mitchell, 2007. Thick non-crossing paths and minimum-cost flows in polygonal domains. In *Proceedings of the 23rd ACM Symposium on Computational Geometry, Gyeongju, South Korea, June 6-8, 2007*, pages 56–65.
G. Strang, 1983. Maximal flow through a domain. *Math. Program.*, 26(2):123–143.
Trafikanalys, 2017a. Rail traffic 2016.
Trafikanalys, 2017b. Railway transport 2017 quarter 3.
Trafikverket, 2018. Järnvägens kapacitet 2017 (in Swedish). Report, publication number 2018:050.
UIC, 2004. Capacity. leaflet 406, International Union of Railways, 1st ed.
UIC, 2013. Capacity. leaflet 406, International Union of Railways, 2nd ed.

A New Approach to Strategic Periodic Railway Timetabling

Gert-Jaap Polinder [a,1], Marie Schmidt [a], Dennis Huisman [b,c]
[a] Rotterdam School of Management, Erasmus University
P.O. Box 1738, 3000 DR Rotterdam, The Netherlands
[1] E-mail: polinder@rsm.nl, Phone: +31 (0) 10 40 88041
[b] Econometric Institute, Erasmus School of Economics, Erasmus University
P.O. Box 1738, 3000 DR Rotterdam, The Netherlands
[c] Department of Process quality and Innovation, Netherlands Railways
P.O. Box 2025, 3500 HA Utrecht, The Netherlands

Abstract
One of the criteria to judge a timetable is what passengers think of it, and an operator has to take this into account when designing a timetable. We study this problem in a case study from the Netherlands, where on part of the network the frequency of trains has increased recently. We formulate a model that integrates passenger routing and timetabling in order to find timetables that are good for passengers. This can be used for studies by railway operators, and by infrastructure managers to decide where to invest in new infrastructure.

Keywords
Periodic Timetabling, Periodic Event Scheduling, Passengers

1 Introduction

The so-called 'A2-corridor' between Amsterdam and Eindhoven is one of the most densely used parts of the Dutch railway network: Recently, the frequency on this corridor was increased from four to six intercity trains per hour, to decrease passenger waiting times at their origin station. Ideally, passengers who travel on this corridor come to the station at any time without checking the timetable and should be able to take a train shortly after they arrive at the station.

The demand for travelling on this part of the Dutch railway network is the highest between Amsterdam and Eindhoven (the corridor itself), but there is also a significant number of passengers traveling on the 'branches' of the corridor, that are the parts of the network shown in Figure 1 that are only served by one or two lines and hence by two or four trains per hour. When making a timetable for this network, we thus observe a trade-off between regularity of the trains on the corridor, and regularity of the trains on the branches. Especially if the trains on the branches have a frequency of four per hour, and on the corridor a frequency of six per hour (not a multiple of four), the timetable can only be regular on both corridor and branches if trains wait relatively long on the stations where they enter and leave the corridor. So in this case we trade a regular service and thus short waiting times of passengers at the origin station for longer in-train waiting times.

In this paper, we formulate a model that optimizes a timetable structure. This timetable

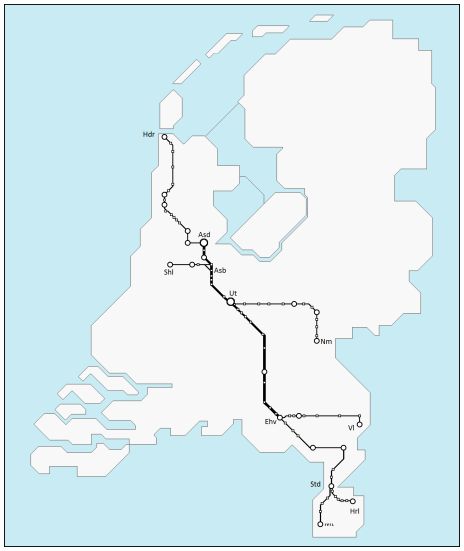


Figure 1: Overview of the geographical network of the A2 corridor instance. The corridor is highlighted.

is based on a input line plan. We build the model by extending the well-known PESP model (Serafini and Ukovich, 1989) for periodic timetabling to include passengers' route choice. In the optimization, we minimize the sum of the perceived travel times for all passengers, which is a weighted sum of waiting time at the origin station and in-train time. Our model can be used as a strategic planning tool, e.g., to evaluate line plans based on the timetable structure they allow. By assuming that we have unlimited infrastructure, we can determine what is an ideal timetable, and hence we can investigate to what extend this timetable fits on the currently existing infrastructure. This can thus also support decision making regarding infrastructure investments.

Our contribution in this paper is twofold. First of all, we propose a quadratic integer programming model to integrate passenger routing and periodic timetabling, where we explicitly take the waiting time at the origin station into account. This model is linearized to a linear mixed integer program. Secondly, we demonstrate the viability of our method on the Dutch A2-corridor instance to advice on the optimal regularity of train lines on this corridor, and the possible benefit of infrastructure investments.

The remainder of this paper is organized as follows. In Section 2 we state the problem we are solving. Section 3 describes background information and literature that is relevant for our study. A quadratic integer programming model is formulated and linearized in Section 4. Computational results on the A2-corridor are provided in Section 5. Finally, we conclude the study and mention future research in Section 6.

2 Problem Statement

In this paper we address a strategic timetabling problem. The goal is to find an ideal timetable structure, that can help us evaluate line plans and advise on infrastructure investments.

The timetable is made based on a line plan. A line plan is a set of train lines that are to be operated on the network. Each of these lines consists of a geographical route through the rail network, a list of stations where the train has to stop, and a frequency by which the train line is to be operated per time period. This line plan serves as input for the timetabling problem.

In this research, we require periodic timetables, i.e., timetables that repeat themselves every time period, say every hour. This type of timetables is often used in European countries.

To find good timetables, we take into account passenger demand. The demand is given in terms of numbers of passengers that want to travel between each pair of stations. We assume that the demand is uniformly distributed over the cycle period, i.e., every minute the same number of passengers want to depart. Often passengers arrive at their origin station shortly before their train departs (Zhu et al., 2017; Ingvardson et al., 2018). However, as the timetable is not yet known, we assume that the demand is evenly distributed over time to find a timetable that best matches this demand assumption, and in the actual operation passengers will adapt their arrival times based on such a timetable.

Our problem can be stated as follows: Given an input line plan and an estimate of passenger demand, find the timetable structure which minimizes the sum of the perceived travel times for all passengers. The perceived travel time is a weighted sum of waiting time at the origin station and in-train time.

3 Literature Review

In this section, we place the problem we study in the context of existing literature. Section 3.1 describes the problem of periodic timetabling and research that is related to this. Section 3.2 describes how passenger routing can be combined with timetabling and how this is done in existing literature.

3.1 Periodic Timetabling

The periodic timetabling problem is commonly modelled as a Periodic Event Scheduling Problem (PESP) (Serafini and Ukovich, 1989). The task here is to assign event times for all arrivals and departures of the trains in the line plan. As the timetable is periodic, these events are periodic as well, i.e., they re-occur every cycle period, e.g., every hour. This cycle time is denoted by T. The event times have to satisfy several restrictions in order to guarantee a reasonable timetable. These restrictions are generally referred to as *activities*. Each activity is a relation between a pair of events, stating that the time difference between these events should be in a given (periodic) time interval. Examples of these activities are drive, dwell and transfer activities. It is also possible to include headway activities, ensuring a certain time distance between trains. Overviews on how to model timetabling constraints and what can be included in a PESP framework can be found in Odijk (1996); Peeters (2003); Liebchen and Nachtigall and Möhring (2007).

The essence of PESP is to find *any* periodic timetable satisfying all activities. Approaches to find a feasible solution to PESP include constraint programming (Kroon et al., 2008), the modulo-simplex heuristic (Nachtigall and Opitz, 2008; Goerigk and Schöbel, 2013), or using a SAT solver after applying a polynomial transformation from PESP to SAT (Grossmann et al., 2012). If a feasible solution exists, one can be found rapidly.

3.2 Passenger Routing

If many feasible timetables exist, one can distinguish between them by adding an objective function to the PESP model, by which *good* timetables can be found. A definition of a good timetable will consist of several aspects, but at least one of the aspects has to do with *efficiency*. By an efficient timetable we mean that the timetable is optimized with respect to passenger travel times. This is achieved by giving each activity a weight and then minimizing the weighted sum of all the activity durations. To solve such a model, a Mixed Integer Programming formulation can be used. More details about such a modelling approach are provided in Section 4. Examples of successful applications in practice can be found in (Liebchen, 2008).

In the case of efficient timetable, the activity weights are chosen such that they represent the (relative) importance of the activities. For example, the weight for an activity can represent the number of passengers using this activity in their route. In this case, the passenger flows have to be known and the timetable can be found based on these flows. However, as a timetable can be suboptimal for certain passengers, they might choose a new route if the timetable is known, thus changing the weights. This in its turn can again influence the optimality of a timetable, which can be changed based on this. Several approaches exist in which an iterative approach is taken to find good a timetable (Kinder, 2008; Lübbe, 2009; Siebert, 2008; Siebert and Goerigk, 2013; Sels et al, 2016). In these approaches, passenger flows are determined by routing passengers through the network on for example shortest paths. After this, the timetable is optimized (retiming) and passenger are rerouted (reflowing), until a stopping criterion is reached.

Another option is to integrate the passenger routing and timetabling problems, which can provide the optimal timetable for the passengers, although the model is more complex. Here the timetable and the passenger routes are chosen simultaneously. This is the approach we take, where we, additionally to most existing literature, also explicitly take the waiting time at the origin station into account. By assuming that passenger demand follows a uniform distribution, good headway times between trains are found, such that the total experienced travel time of passengers is minimized. The integration of passenger routing and timetabling is not a new field of study, as this already has been applied for the aperiodic case (Schmidt and Schöbel, 2015; Schmidt, 2012) and for the periodic case (Schöbel., 2015; Borndorfer et al., 2017; Gattermann et al., 2016; Schiewe and Schöbel, 2018). However, the approach we take is that we do not assign passengers to a specific departure event before solving the model, but that we allow this freedom in the model. This is most closely related to Schiewe and Schöbel (2018). However, we also assume that demand is uniformly distributed over time and we determine all headway times between consecutive trains based on this, which, to the best of our knowledge, has not yet been applied in literature. Burggraeve et al. (2017) also integrate timetabling and line planning to achieve better results. However, in this approach infrastructure is taken into account in a very detailed manner, while we discard as much of the current infrastructure as possible, in order to determine long term

strategic timetables.

In the past, attention has been given to the gap between line planning and timetabling, and that integrating these two problems does not solve everything. Goerigk et al. (2013) propose several evaluations of line plans and determine the influence of a line plan on the resulting timetable. This is done by computing several characteristics of line plans, and by finding a passenger-oriented timetable. However, the passenger routing and timetabling problems are not integrated, but passengers are routed along shortest paths, which is one of the main differences with our approach.

4 Integer Programming model

In this section, we present a mathematical programming model for passenger routing and timetabling. We start by introducing the necessary notation, after which we present the mathematical model. We conclude the section by linearizing the proposed model.

4.1 Notation

The model that we introduce consists of a timetabling part and a passenger routing part. In the following sections, we introduce these two parts separately.

Periodic Timetabling
First of all, we assume that an Event-Activity network $G = (V, A)$ is given, with events V and activities A. Based on this network, we develop a model to find a periodic timetable with cycle time T. A commonly used model for periodic timetabling is based on the Periodic Event Scheduling Problem (Serafini and Ukovich, 1989). In PESP, next to the network $G = (V, A)$, we are given lower and upper bounds ℓ_{ij} and u_{ij} for each $(i, j) \in A$ and a cycle time T. The task is to find an assignment $\pi : V \to \{0, \dots, T-1\}$, such that all activities are satisfied. In PESP, each activity $(i, j) \in A$ is of the form

$$y_{ij} = \pi_j - \pi_i + Tp_{ij} \in [\ell_{ij}, u_{ij}], \tag{1}$$

where p_{ij} is an integer variable accounting for the shift from one cycle to another, it acts as a modulo operator. Each activity states that the time difference between events i and j should be within the T-periodic interval $[\ell_{ij}, u_{ij}]$. The additionally introduced variable y_{ij} represents the *activity duration* for activity $(i, j) \in A$.

Without loss of generality, the timetable is planned in full minutes, but any other time grid can be chosen as well. The rationale behind this assumption is that we want to find a timetable for the long future and there is no need for a detailed timetable in this case.

Passenger Routing
Next to timetabling, we have variables and constraints dealing with the routing of passengers. Suppose that passenger demand is given in terms of an OD-matrix. This provides for each origin-destination combination k in the set $\mathcal{OD}$ the number of passengers d_k that want to travel in the cycle period from their origin to their corresponding destination. We assume that this demand is uniformly distributed over the time period.

For each OD-pair $k \in \mathcal{OD}$, we pre-determine a set of possible routes, which we denote by $\mathcal{R}^k$. In our computations, this set consists of all direct travel options for this OD-pair, but

this can be extended to routes containing transfers as well. The set of all routes is denoted by $\mathcal{R}$ and is determined as

$$\mathcal{R} = \bigcup_{k \in \mathcal{OD}} \mathcal{R}^k. \tag{2}$$

We assume that these sets are given as input.

A route $r \in \mathcal{R}$ is a path through the Event-Activity Network. It consists of a sequence of trip and dwell activities, so $r \subseteq A$. The total (timetable-dependent) duration Y_r of such a route is determined as the sum of the duration of all activities it uses:

$$Y_r = \sum_{a \in r} y_a. \tag{3}$$

The task in our model is to assign the passengers of every OD-pair to a relevant departure event, and route them all together. For each OD-pair $k \in \mathcal{OD}$, the set of relevant departure events (V^k) can be determined by

$$V^k = \bigcup_{r \in \mathcal{R}^k} j(r), \tag{4}$$

where $j(r)$ is the first event of route $r \in \mathcal{R}$.

In our model, we make the simplifying assumption that every passenger departs with the first train towards his destination. Note that in practice this may not be the best traveling option, since a later train may overtake the first train. However, since in our case study we consider only intercity trains which travel at approximately the same speed, this assumption seems appropriate.

We group all passengers of an OD-pair k for who, due to their arrival time, departure event $v \in V^k$ is the next possible departure event, together and assume they make the same route choice. This assumption is valid since the perceived passenger travel time is minimized and we assume that all passengers have the same perception of travel time, and that there are no capacities on the routes.

In order to compute the number of passengers for who v is the next possible departure event, we determine the time period before event $v \in V^k$, in which no other event $v' \in V^k \setminus \{v\}$ takes place. This is denoted by A_v^k. Note that this variable is timetable dependent, and can be determined by the following set of equations:

$$A_v^k = \min_{v' \in V^k} \{\pi_v - \pi_{v'} + T\alpha_{v',v}\} \tag{5a}$$

$$\alpha_{v,v'} + \alpha_{v',v} = 1. \tag{5b}$$

In the first equation, the time difference between all other relevant departure events is determined, and the minimum is taken. In order to determine an implicit order between events happening at the same time, which is needed to determine how many passengers take a certain train, we add the second set of restrictions.

The number of passengers for event $v \in V^k$ can then be calculated as $A_v^k \cdot d_k/T$. Once event $v \in V^k$ takes place, all these passengers choose a route. The set of all routes, starting at this departure event, is denoted by $\mathcal{R}_v^k \subseteq \mathcal{R}^k$. If a passenger departs with event $v \in V^k$, he will choose exactly one of these routes to use. The duration of the journey, starting from event v, is denoted by Y_v^k, and can be determined as

$$Y_v^k = \min_{r \in \mathcal{R}_v^k} Y_r. \tag{6}$$

Note that this assumes that passengers use shortest paths, which is true since the perceived passenger travel time is minimized.

The expected waiting time for each group of passengers at the origin station is denoted by W_v^k. As we assume that passengers arrive according to a uniform distribution, this value is calculated as $W_v^k = A_v^k/2$.

To compute the total perceived travel time of a passenger, we weight the waiting time at the origin station with a factor γ_w and add it to the in-train time or duration of the journey. E.g., a factor $\gamma_w = 3$ means that a passengers perceives one minute waiting at the station as bad as three minutes traveling on the train. We can then compute the average perceived travel time of each passenger k for who departure event v is the next possible departure as $\gamma_w W_v^k + Y_v^k$.

4.2 Mathematical Program

Using the notation and constraints introduced above, a Quadratic Integer Program for timetabling with passenger routing, including waiting times, can now be formulated as follows:

$$\text{Minimize} \quad \sum_{k \in \mathcal{OD}} \frac{d_k}{T} \sum_{v \in V^k} A_v^k \cdot \left(\gamma_w \cdot W_v^k + Y_v^k\right) \tag{7a}$$

$$\text{Such that} \quad y_{ij} = \pi_j - \pi_i + T p_{ij} \qquad \forall\, (i,j) \in A \tag{7b}$$

$$\ell_{ij} \leq y_{ij} \leq u_{ij} \qquad \forall\, (i,j) \in A \tag{7c}$$

$$Y_r = \sum_{a \in r} y_a \qquad \forall\, r \in \mathcal{R} \tag{7d}$$

$$A_v^k = \min_{v' \in V^k \setminus \{v\}} \{\pi_v - \pi_{v'} + T\alpha_{v',v}\} \qquad \forall\, k \in \mathcal{OD},\, v \in V^k \tag{7e}$$

$$\alpha_{v,v'} + \alpha_{v',v} = 1 \qquad \forall\, k \in \mathcal{OD},\, v \in V^k,\, v' \in V^k \setminus \{v\} \tag{7f}$$

$$W_v^k = \frac{1}{2} A_v^k \qquad \forall\, k \in \mathcal{OD},\, v \in V^k \tag{7g}$$

$$Y_v^k = \min_{r \in \mathcal{R}_v^k} \{Y_r\} \qquad \forall\, k \in \mathcal{OD},\, v \in V^k \tag{7h}$$

$$A_v^k \in [0, T] \qquad \forall\, k \in \mathcal{OD},\, v \in V^k \tag{7i}$$

$$W_v^k \in [0, T/2] \qquad \forall\, k \in \mathcal{OD},\, v \in V^k \tag{7j}$$

$$Y_r, Y_v^k \in [0, \infty) \qquad \forall\, r \in \mathcal{R},\, k \in \mathcal{OD},\, v \in V^k \tag{7k}$$

$$\pi_v \in \{0, \ldots, T-1\} \qquad \forall\, v \in V \tag{7l}$$

$$p_{ij} \in \mathbb{Z}_{\geq 0} \qquad \forall\, (i,j) \in A \tag{7m}$$

$$\alpha_{v,v'} \in \{0, 1\} \qquad \forall\, k \in \mathcal{OD},\, v \in V^k,\, v' \in V^k \setminus \{v\}. \tag{7n}$$

The task in this formulation is to minimize the perceived travel time for all passengers (7a). This is composed of waiting time, plus the actual travel time. Constraints (7b) and (7c) are the timetabling constraints. Constraints (7d) determine the length of each route. Constraints (7e) and (7f) determine the time between trains, and define the expected waiting times in (7g). The actual perceived travel time durations are determined in (7h). Constraints

(7i)–(7n) state the domains of the variables. Note that this is a quadratic model. In the next section, we show how this model can be linearized.

4.3 Linearization

The model in (7) contains a quadratic objective and two minima in the formulation. In the following sections, we linearize each part of this.

Objective
The objective in (7) can, by using (7g), be written as

$$\text{Minimize} \sum_{k \in \mathcal{OD}} \frac{d_k}{T} \sum_{v \in V^k} \frac{\gamma_w}{2} \left(A_v^k\right)^2 + A_v^k \cdot Y_v^k. \tag{8}$$

For the linearization we define new variables

$$x_{v,d}^k = \begin{cases} 1 & \text{if } A_v^k \geq d \\ 0 & \text{else} \end{cases} \qquad \forall\, k \in \mathcal{OD},\, v \in V^k,\, d \in \{1, \ldots, T\}. \tag{9}$$

Note that these variables satisfy the following restrictions:

$$x_{v,d}^k \leq x_{v,d-1}^k \qquad \forall\, k \in \mathcal{OD},\, v \in V^k,\, d \in \{2, \ldots, T\}. \tag{10}$$

Using these new variables, we can write

$$A_v^k = \sum_{d=1}^{T} x_{v,d}^k \tag{11a}$$

$$\left(A_v^k\right)^2 = \sum_{d=1}^{T} (2d-1) \cdot x_{v,d}^k. \tag{11b}$$

Substituting this in (8) leaves a multiplication of binary variables $x_{v,d}^k$ by bounded variables Y_v^k, which we substitute by $R_{v,d}^k = Y_v^k \cdot x_{v,d}^k$. The objective then becomes

$$\text{Minimize} \sum_{k \in \mathcal{OD}} \frac{d_k}{T} \sum_{v \in V^k} \sum_{d=1}^{T} \left[\frac{\gamma_w}{2} (2d-1) \cdot x_{v,d}^k + R_{v,d}^k \right], \tag{12}$$

with the additional restrictions that

$$R_{v,d}^k \leq u_v^k \cdot x_{v,d}^k \tag{13a}$$

$$R_{v,d}^k \geq l_v^k \cdot x_{v,d}^k \tag{13b}$$

$$R_{v,d}^k \leq Y_v^k - l_v^k \cdot \left(1 - x_{v,d}^k\right) \tag{13c}$$

$$R_{v,d}^k \geq Y_v^k - u_v^k \cdot \left(1 - x_{v,d}^k\right), \tag{13d}$$

where l_v^k and u_v^k are the lowest and highest possible values for Y_v^k respectively.

Minima
Constraints (7e) and (7h) both contain a minimum. We replace (7e) by

$$A_v^k \leq \pi_v - \pi_{v'} + T\alpha_{v',v} \qquad \forall\, k \in \mathcal{OD},\, v' \in V^k \setminus \{v\} \tag{14a}$$

$$\sum_{v \in V^k} A_v^k = T. \qquad \forall\, k \in \mathcal{OD} \tag{14b}$$

(14a) represents the minimum and (14b) ensures that all passengers are assigned to a group.

The linearization of (7h) is done by replacing the following set of restrictions for every $k \in \mathcal{OD}$ and every $v \in V^k$ by the following:

$$Y_v^k \leq Y_r \qquad \forall\, r \in \mathcal{R}_v^k \tag{15a}$$

$$Y_v^k \geq Y_r - M_v^k \cdot \left(1 - z_{v,r}^k\right) \qquad \forall\, r \in \mathcal{R}_v^k \tag{15b}$$

$$\sum_{r \in \mathcal{R}_v^k} z_{v,r}^k = 1. \tag{15c}$$

We introduced new binary variables $z_{v,r}^k$, which correspond to the route chosen. That means, if $z_{v,r}^k = 1$, passengers use route r, starting at event v. The newly introduced constants M_v^k are to be chosen large enough to make the second set of constraints redundant, but still as small as possible. Therefore, we can set

$$M_v^k = \max_{r \in \mathcal{R}_v^k} \{\overline{Y_r}\} - \min_{r \in \mathcal{R}_v^k} \{\underline{Y_r}\}. \tag{16}$$

Here, $\overline{Y}$,$\underline{Y}$ denote respectively the highest and lowest possible value variable Y can take.

To summarize, in the linearization several steps are taken. The objective (7a) is replaced by (12). Here, additional variables $x_{v,d}^k$ and $R_{v,d}^k$ are introduced, with additional restrictions (10) and (13). Next, the minima are replaced by linear restrictions, (7e) is replaced by (14), and (7h) by (15).

5 A Case Study: A2 Corridor

5.1 Description of the Case

The case we consider in our study is the so called 'A2-corridor', which is part of the Dutch railway network between Eindhoven and Amsterdam Centraal. It is named after the highway A2 which runs next to it for a long part of the track. The most interesting feature of this instance is that it is the first corridor in the Netherlands where the frequency of the intercity trains was increased from four to six intercity lines per hour. An overview of the lines in the network is shown in Figure 2. In this Figure, the main stations are shown, with corridors connecting them. For each corridor, the train lines that uses these corridors are shown. For example, it shows that line 3000 travels between Nijmegen and Den Helder. The full names and abbreviations of the involved stations are shown in Table 1.

Each train line is operated in both directions with a frequency of two trains per hour. Note that not all of these trains use the full corridor between Eindhoven and Amsterdam Centraal, line 3000 only uses Utrecht-Amsterdam Centraal, and line 3500 only uses Eindhoven-Utrecht. Line 3100 does not use the corridor itself at all, but it interacts with other lines on the branches of the network as depicted in Figure 1.

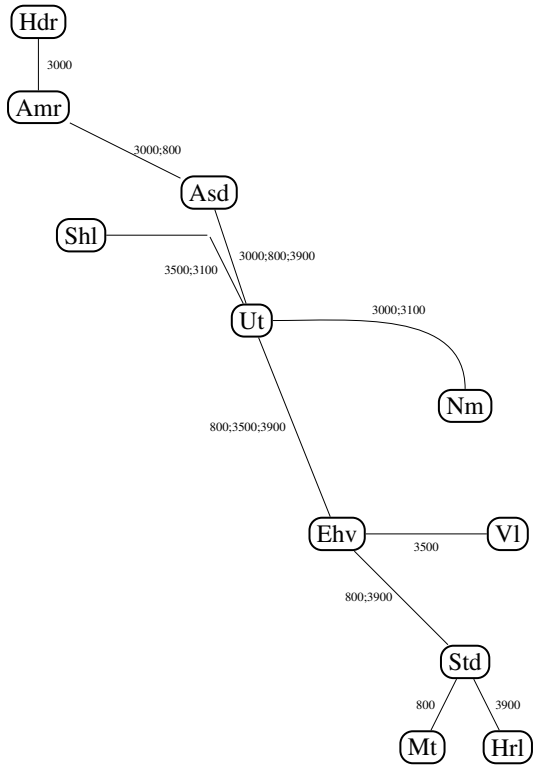


Figure 2: Overview of the train lines in the A2-corridor

Abbreviation	Name	Abbreviation	Name
Hdr	Den Helder	Sgn	Schagen
Amr	Alkmaar	Asd	Amsterdam Centraal
Shl	Schiphol	Asb	Amsterdam Bijlmer ArenA
Ut	Utrecht Centraal	Ah	Arnhem
Nm	Nijmegen	Ht	's Hertogenbosch
Ehv	Eindhoven	Vl	Venlo
Std	Sittard	Hrl	Heerlen
Mt	Maastricht		

Table 1: Abbreviations of the stations

In Section 5.2, we investigate how regular the train lines should run on the corridor and on the branches of the network, if we assume unlimited infrastructure. We also investigate how this result changes when we adjust the weighting parameter γ_w which controls the impact of the waiting time at the origin.

As a reference, we also compute a timetable in Section 5.3, where we assume that we have the current infrastructure available.

In our computations, we only consider direct travel options for all passengers, so the set of possible routes can be easily determined. For the OD-matrix, we make use of an OD-matrix containing the number of all passengers that travelled between two stations in the year 2015. Hence, this is based on historic data and is an aggregated matrix. The flows will be different in peak and off-peak hours, and in the weekends. For our study that is no problem, as we are doing strategic studies and only the estimated magnitude of the flows matter.

5.2 Unlimited Infrastructure

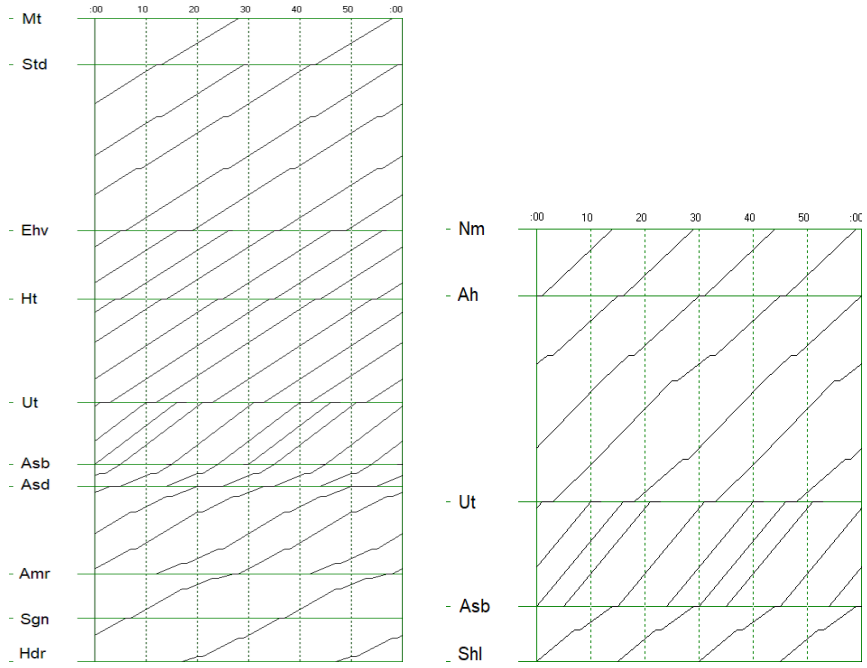
As the main goal of this research is to find out how an optimal timetable from a passengers perspective looks like, we assume there is sufficient infrastructure available to operate any timetable. We compute a timetable by solving (7) with a time limit for the computation of one hour. As travel options we only consider direct connections for all passengers. We distinguish between two cases. First, we assume that waiting at the origin station is considered to be less pleasant than sitting in a train. Second, we compare this to the situation where waiting time is equally pleasant to in-train time.

Waiting Time is Less Pleasant

When passengers travel, they often arrive shortly before their train departs, in order to minimize their waiting time. However, as no timetable is available yet, these arrival times can not be determined, and therefore we assume that passengers arrive according to a uniform distribution. In order to deal with the situation that passengers generally do not like to wait, we set the coefficient for waiting time relatively high. In-train time has a coefficient of 1, so we set the waiting time coefficient higher to $\gamma_w = 3$.

The results of our computations are shown in Figure 3, showing the results in terms of two time-space diagrams. Only the southbound trains are shown to make the picture more clear. The first picture shows the tracks between Maastricht and Den Helder (and hence

includes the full corridor), the other shows the tracks between Nijmegen and Schiphol. The vertical axes denotes space and the labels show the different stations where a trains stops. The horizontal axis shows time, so the timetable is depicted for one cycle period of one hour.



(a) Maastricht-Den Helder (b) Nijmegen-Schiphol

Figure 3: Timetable with unlimited infrastructure, $\gamma_w = 3$.

The objective value and a lower bound for this timetable can be found in Table 2. The objective is split in two parts to better distinguish what the contribution is of the individual components.

Interesting to note in this Figure, is that the trains run very regular between the large stations on the main corridor, i.e., between Eindhoven and Utrecht, and between Utrecht and Amsterdam. Between Amsterdam and Alkmaar, the pattern becomes slightly less regular. Here a choice has to be made for a service between these stations and hence a longer waiting time at the Amsterdam Central station at the border of the corridor (because the frequency decreases from 6 to 4 trains per hour), or a less regular service and short waiting times. Apparently it seems to be better not to stop too long at the border of the corridor to obtain the regular pattern between Amsterdam and Alkmaar.

Another interesting thing to note here, is that the train paths of lines 3000 and 3500 coincide. It is hard to see in Figure 3, but can be deduced together with Figure 2. Line 3000 comes from Amsterdam Central and passes Amsterdam Bijlmer Arena (Asb) at .45. Also line 3500, coming from Schiphol, leaves Amsterdam Bijlmer ArenA at .45, and it travels to

Utrecht at the same time as line 3000. From Utrecht onwards, line 3000 leaves the corridor for Nijmegen, while line 3500 takes over the position of line 3000 in the pattern on the corridor and drives towards Eindhoven. So we can say that these trains replace each other in the pattern on the corridor. Passengers traveling from Amsterdam to Eindhoven could hence use also this route with a transfer. This happens even though we do not allow passengers to transfer in our model - it is a consequence of the fact that regular train services are preferred by the model due to the high impact of the waiting time.

Waiting Time and In-Train Time is Equal
We now compare the findings of Section 5.2 with a setting of $\gamma_w = 1$, i.e., the in-train time and waiting time are perceived equally. The resulting timetable for the corridor is shown in Figure 4.

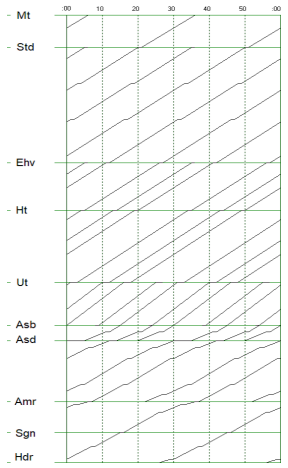


Figure 4: Maastricht-Den Helder with unlimited infrastructure and $\gamma_w = 1$

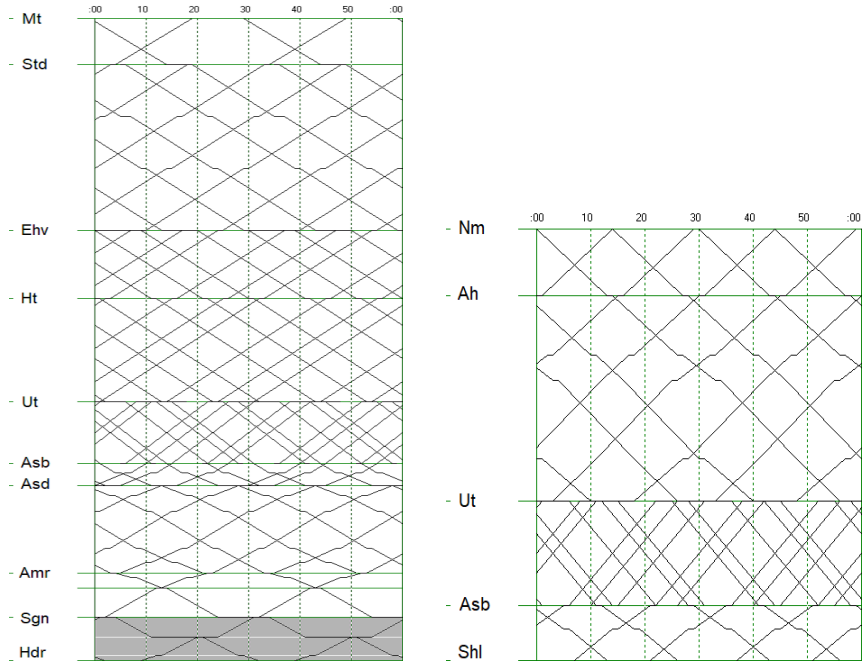
In this timetable, the trains have a less regular pattern. As the waiting time is less important, a smaller focus is on the regularity of the trains. Also in Table 2, this effect can be seen. The contribution of in-train time decreases, whereas the contribution of waiting time is increased.

5.3 Current Infrastructure

If we compare the timetables determined in the previous sections with the current infrastructure on the A2 corridor and its branches, we observe that the current infrastructure is not sufficient. In particular, between Den Helder (Hdr) and Schagen (Sgn), the network is currently single-track, so crossings of trains from two directions, as we see them in Figure 3a, are not possible with the current infrastructure. Furthermore, currently the trains from Utrecht to Amsterdam Centraal (Asd) and from Utrecht to Schiphol (Shl) use the same infrastructre until Amsterdam Bijlmer Arena (Asb), thus a headway of three minutes has to be respected between them if no additional infrastructure would be provided here. To see the benefit of providing extra infrastructure, we now compute the ideal timetable

structure taking current infrastructure restrictions into account by adding them as headway constraints to our optimization model, and comparing the timetable achieved in this way to the timetable from Section 5.2.

More precisely, the headways we consider state that between every pair of trains running in the same direction, a headway time of 3 minutes has to be respected. Furthermore, between trains entering and leaving a station in opposite directions on single track regions, a time difference of at least 1 minute has to be respected. Finally, on single track regions trains have to wait for each other before the single track can be entered. Again, we set $\gamma_w = 3$. Within one hour of computation time, several solutions can be found. However, due to the increased complexity caused by the added headway activities, these solutions are not very good. Therefore, we set a time limit of three hours for these computations. The resulting timetable is shown in Figure 5 and the objective values are shown in Table 2. The shaded area in Figure 5a denotes that this part has only 1 track, and trains can only pass each other at the stations.



(a) Maastricht-Den Helder (b) Nijmegen-Schiphol

Figure 5: Timetable with current infrastructure, $\gamma_w = 3$.

A few differences can be noted with respect to the case with unlimited infrastructure and $\gamma_w = 3$. First of all, trains dwell for a long time on stations around the single-track area. Especially the northbound trains have long dwell times. These trains come from a region that has a very regular timetable, and then they have to wait for the southbound trains because these are using the tracks. Based on passengers' demand, a choice can be

made whether the southbound of the northbound trains have to wait. Also at other stations trains dwell for a longer time. As an example, trains coming from Amsterdam and going to Nijmegen have to wait at Utrecht from .58 to .03, in order have a regular service on the branch Utrecht-Nijmegen.

For this line plan, the cost of taking infrastructure into account in terms of experienced travel time is not much more than in the case with unlimited infrastructure, the objective increases by only 1.18% (Table 2).

	γ_w	In-train time	Waiting Time	Objective value	Lowerbound
Unlimited infrastructure	3	2047.004	592.576	3824.733	3808.588
Unlimited infrastructure	1	2042.174	597.926	3835.953[1]	
Current infrastructure	3	2083.382	595.510	3869.912	3808.547

Table 2: Objective values for the timetables

6 Conclusion and Future Research

In this paper, we have developed a mathematical model that allows us to determine timetable structures based on passenger demand, explicitly taking into account waiting time of passengers. By applying the model to the A2 corridor in the Dutch railway network, we could observe that the inclusion of passenger waiting time leads to regular timetable structures, although regularity was not imposed as a constraint in our model. This effect is particularly strong when the waiting time at the origin station plays a prominent role in the objective function. When the weight of waiting time at the origin is smaller, the regularity diminishes to improve dwell times in stations. By applying our model also to a case with infrastructure restrictions, we could furthermore quantify the impact the added value of infrastructure investments for passengers.

In the future, we will apply our approach to other cases from the Dutch Railway network to see whether these findings are instance-specific or can be generalized. In both cases, we think that our model is a valuable tool for strategic timetabling, in the sense that it can help to evaluate line plans and infrastructure and help to make improvements.

At this moment, we have several ideas for extensions to our model and to our solution approach. Currently the model deals with direct travel options for passengers only. This is a limitation that we are improving, such that any travel option can be included. Furthermore, we want to include the option of not taking the first departing train, but to take a later train instead, and to investigate to what extend this improves the solutions. Thirdly, as PESP is NP-complete, computation times rapidly grow when more activities are taken into account. Especially headway restrictions can make it hard to find a feasible solution, let alone a good solution. Therefore we have set a higher computation time limit for these experiments. In order to deal with this increasing complexity, we developed an algorithm that first finds a good timetable as is done in this paper, with the assumption of unlimited infrastructure. In a

[1] In order to compare the objective value of this timetable with the other timetables, we have evaluated this timetable with $\gamma_w = 3$.

second step, we use an algorithm based on Cacchiani et al. (2013) that updates this timetable considering a given infrastructure network, such that all trains can be safely operated on this network, and if necessary cancels trains.

References

Borndorfer, R., Hoppmann, H., Karbstein, M., 2017. "Passenger routing for periodic timetable optimization", *Public Transport*, vol. 9(1), pp. 115-135, https://doi.org/10.1007/s12469-016-0132-0

Burggreave, V., Bull, S.H., Vansteenwegen, P., Lusby, R.M., 2017. "Integrating robust timetabling in line plan optimizationfor railway systems", *Transportation Research Part C: Emerging Technologies*, vol. 77, pp. 134-160.

Cacchiani, V., Caprara, A., Toth, P., 2010. "Scheduling extra freight trains on railway networks", *Transportation Research Part B: Methodological*, vol. 44(2), pp. 215-231.

Gattermann, P., Großmann, P., Nachtigall, K., Schöbel, A., 2016 "Integrating Passengers' Routes in Periodic Timetabling: A SAT approach", In: Goerigk, M., Werneck, R. (eds.), *16th workshop on Algorithmic Approaches for Transportation Modeling, Optimization, and Systems*, vol. 54, Dagstuhl, Germany.

Goerigk, M., Schöbel, A., 2013. "Improving the modulo simplex algorithm for large-scale periodic timetabling", *Computers & Operations Research*, vol. 40(5), pp. 1363-1370. https://doi.org/10.1016/j.cor.2012.08.018.

Goerigk, M., Schachtebeck, M., Schöbel, A., 2013. "Evaluating line concepts using travel times and robustness", *Public Transport*, vol. 5(3), pp. 267-284. https://doi.org/10.1007/s12469-013-0072-x.

Grossmann, P., Hölldobler, S., Manthey, N., Nachtigall, K., Opitz, J., Steinke, P., 2012. "Solving periodic event scheduling problems with SAT", In: Jiang, H., Ding, W., Ali, M., Wu, X. (eds.), *Advanced Research in Applied Arti*
cial Intelligence, vol. 7345, pp. 166–175, Springer Berlin Heidelberg.

Ingvardson, J., Nielsen, O., Raveau, S., Nielsen, B., 2018. "Passenger arrival and waiting time distributions dependent on train service frequency and station characteristics: A smart card data analysis", *Transportation Research. Part C: Emerging Technologies*, vol. 90, pp. 292-306.

Kinder, M., 2009. *Models for periodic timetabling*, Diploma thesis, Technische Universität Berlin.

Kroon, L., Huisman, D., Abbink, E., Fioole, P.J., Fischetti, M., Maróti, G., Schrijver, A., Steenbeek, A., Ybema, R., 2008. "The new Dutch timetable: The OR revolution", *Interfaces*, vol. 39(1), pp. 6-17.

Liebchen, C., 2008. "The First Optimized Railway Timetable in Practice", *Transportation Science*, vol. 42(4), pp. 420-435. https://doi.org/10.1287/trsc.1080.0240.

Liebchen, C., Möhring, R., 2007 "The Modeling Power of the Periodic Event Scheduling Problem: Railway Timetables – and Beyond", In: Geraets, F. and Kroon, L. and Schöbel, A. and Wagner, D. and Zaroliagis, C. (eds.), *Algorithmic Methods for Railway Optimization. Lecture Notes in Computer Science*, vol. 4359, Springer, Berlin, Heidelberg

Lübbe, J., 2009. *Passagierrouting und taktfahrplanoptimierung*, Diploma thesis, Technische Universität Berlin.

Nachtigall, K., Opitz, J., 2008 "Solving Periodic Timetable Optimisation Problems by Mod-

ulo Simplex Calculations", In: Fischetti, M., Widmayers, P. (eds.), *8th workshop on Algorithmic Approaches for Transportation Modeling, Optimization, and Systems*, vol. 9, Dagstuhl, Germany.

Odijk, M., 1996. "A constraint generation algorithm for the construction of periodic railway timetables", *Transportation Research Part B: Methodological*, vol. 30(6), pp. 455-464. http://dx.doi.org/10.1016/0191-2615(96)00005-7

Peeters, L., 2003. *Cyclic Railway Timetable Optimization*, PhD thesis, Erasmus University Rotterdam.

Schiewe, P., Schöbel, A., 2018. *Periodic timetabling with integrated routing: An applicable approach*, Preprint nr. 17, Preprint-Serie des Instituts für Numerische und Angewandte Mathematik, Universitat Göttingen. http://num.math.uni-goettingen.de/preprints/files/2018-17.pdf.

Schmidt, M., 2012. *Integrating Routing Decisions in Public Transportation Problems*, PhD thesis, Georg-August-Universität Göttingen, Springer.

Schmidt, M., Schöbel, A., 2015. "Timetabling with Passenger Routing", *OR Spectrum*, vol. 37(1), pp. 75-97, https://doi.org/10.1007/s00291-014-0360-0

Schöbel, A., 2015 In: Proceedings of the Conference on Advanced Systems in Public Transport. Rotterdam, The Netherlands.

Sels, P., Dewilde, T., Cattrysse, D., Vansteenwegen, P., 2016. "Reducing the passenger travel time in practice by the automated construction of a robust railway timetable", *Transportation Research Part B: Methodological*, vol. 84, pp. 124-156. https://doi.org/10.1016/j.trb.2015.12.007

Serafini, P., Ukovich, W., 1989. "A Mathematical Model for Periodic Scheduling Problems", *SIAM Journal on Discrete Mathematics*, vol. 2(4), pp. 550-581, http://dx.doi.org/10.1137/0402049

Siebert, M., 2011. *Integration of routing and timetabling in public transportation*, MSc thesis, Georg-August-Universität Göttingen.

Siebert, M., Goerigk, M., 2013. "An experimental comparison of periodic timetabling models", *Computers & Operations Research*, vol. 40(10), pp. 2251-2259, https://doi.org/10.1016/j.cor.2013.04.002

Zhu, Y., Meo, B., Bai, Y., Chen, S., 2017. "A bi-level model for single-line rail timetable design with consideration of demand and capacity", *Transportation Research Part C: Emerging Technologies*, vol. 85(1), pp. 211-233, https://doi.org/10.1016/j.trc.2017.09.002

A Method of Generating Energy-efficient Train Timetable Including Charging Strategy for Catenary-free Railways with Battery Trains

Takuya Sato [a,1], Masafumi Miyatake [a]
[a] Department of Engineering and Applied Science, Sophia University
7-1 Kioi-cho, Chiyoda-ku, Tokyo 102-8554, Japan
[1] E-mail: takuya-sato@eagle.sophia.ac.jp, Phone: +81 (0) 3 3238 3008

Abstract
Battery train is in development as train which can travel in non-electric section using power supplied from onboard storage system such as lithium-ion battery. However, as this type of train has some problems. First, battery train needs to take charging time when it runs a long distance more than its maximum cruising distance. Second, the energy consumption of battery train depends on the state of charge of the storage system. Needless to say, the energy consumption also depends on running time. Moreover, in catenary-free transportation system, battery train can't give regenerative energy to other trains in acceleration through the catenary. Hence, it's important how much the energy storage system can absorb regenerative energy. In consideration of these characteristics, it can be said that the energy-efficiency of catenary-free transportation system with battery train is in a complicated situation because those factors affect each other. When we design this system efficiently, we have to consider simultaneously "at which station the train should charge the battery", "how fast the train should run", and "how long the train should dwell at each station".

In this research, we propose a method of generating the timetable which is the most energy-saving when a single battery train travels on a route section containing multiple stations. The route section is assumed which distance is longer than the maximum cruising distance and the battery train needs to charge at any station. Although this optimization seems to be defined as a nonlinear programming problem, we use linear approximation to the energy consumption characteristic and ease to solve this problem as a MILP (Mixed Integer Linear Programing). In the end, the proposed methodology and customization analysis are applied on a real case study of the route section of Japan.

Keywords
battery train, lithium-ion battery, energy efficiency, train scheduling, optimization, MILP

1. Introduction

Catenary-free transportation system is tendency and attractive alternative to traditional fuel cars and railways with overhead contact systems. It can realize a beautiful urban landscape by removing catenary and reduce CO2, exhaust gas, and noise of engine. This kind of vehicles is generally powered through the onboard ESS (Energy Storage System) which supply electric energy and absorb regenerative energy. Commonly, regenerative brakes of rail vehicles are used under DC electrification and contribute to enhance the energy-efficiency. Especially, battery train which uses battery as onboard ESS become practical as their capacity increases. When we design a catenary-free transportation system with battery

train, an important point is the selection of the ESS based on the cruising distance and its capacity. Although some researches (Ishino et al. 2012) that design optimally the capacity of ESS or charging infrastructures considering route conditions or rapid charge has been conducted, the operation that uses the limited stored energy efficiently has not been focused on. A related work shows that the difference of the SOE (State of Energy) of the ESS effects on the energy consumption of battery train (Noda et al. 2015). One reason of this difference is caused by that the ESS cannot absorb the regenerative energy when SOE is too high. Another reason is voltage drop of ESS. This voltage drop causes a decrease of the traction force and an increase of the energy consumption. The change of SOE is determined by how fast the train run. Hence, the effective strategy of both SOE and running time for the battery train is necessary. The allocation of running time have been studied so far. In this paper, we regard the characteristic of the storage system as constant and investigate the SOE-characteristic of battery train. After that, we present a mathematical programming and propose a method of generating the schedule to minimize the total energy consumption of battery train by allocation of running time, SOE and the location of charging points.

2. Related Work

In the literature, several ways to reduce the energy consumption are proposed and discussed. These ways are classified broadly into two groups. One way proposes the energy-saving train speed profiles (Licheng et al. 2017). Although this study focused on the traditional railway system with the overhead contact system, in the field of catenary-free transportation system, the optimized speed profile for the battery train is also proposed (Noda and Miyatake 2016). Another way optimizes the timetable (Pena-Alcaraz et al. 2012) (Sicre et al. 2010) (Li et al. 2014). They propose the method of generating energy-efficient timetable by adjusting the runtime of trains. There are also some papers which targets the energy-saving of the catenary-free transportation (Ishino et al. 2012). But in these studies, the target of optimization is mainly the speed profiles or the capacity of the energy storage system. It has been rare to optimize the timetable of catenary-free transportation. As it was mentioned before, the earlier study in this field deal with the ordinally train with the overhead contact system. Miyatake, Kuwahara and Nakasa (2012) proposed a comprehensive mathematical formulation as a linear/nonlinear programming for considering energy-saving train scheduling. They use the relation between runtime and energy consumption. It is used for the smooth railway operation, such that absorbs delays and prevent from propagating them, to make a mathematical programming model to optimize the timetable (Andersson et al. (2015)). Furthermore, these solution methods other than linear/nonlinear programming such as Genetic Algorism is also proposed for this kind of problem (Arenas et al. 2015) (Chao et al. 2016).

Based on the earlier study (Miyatake 2012), we make mathematical formulation as a linear/nonlinear programming for the catenary-free transportation system with battery trains to optimize their timetable. However, when we consider the appropriate programming for battery trains, we also need to consider the SOE (State of Energy) of the battery as the second parameter because the energy consumption of battery trains depends on SOE. When the SOE is low, the terminal voltage drops, and the energy consumption increases. Contrarily when the SOE is too high, the energy consumption also increases because the battery can't absorb all regenerative energy. Taking this characteristic into account, it's important for battery trains to keep the SOE within appropriate range. Not only running time, the location of charging spot when stopping at the station is also variable. It is possibility to change the energy consumption to change the charging point because the SOE

while running recovers and differs from the same section. As the final goal of this study, we make a mathematical programming to optimize running time, SOE control and charging spot simultaneously for catenary-free transportation with battery train.

3. The Characteristic of the Energy Consumption

Modeling of energy storage and battery train

In this research, we deal with a battery train which uses high power lithium-ion battery as the onboard ESS. Generally, super capacitor is often used as the onboard ESS of battery train. We can make a circuit equation between the current capacity and the terminal voltage for super capacitor. We use the indicator SOC (State of Charge) which represent the current capacity of the battery. However, lithium-ion battery has a nonlinear drooping characteristic between its SOC[%] and its no-load open voltage v[V]. We identify the voltage characteristic of LIM-30H made by GS YUASA Corporation (Seyama et al. 2007) in Fig.1. We approximate this characteristic by the following quadratic equation (1).

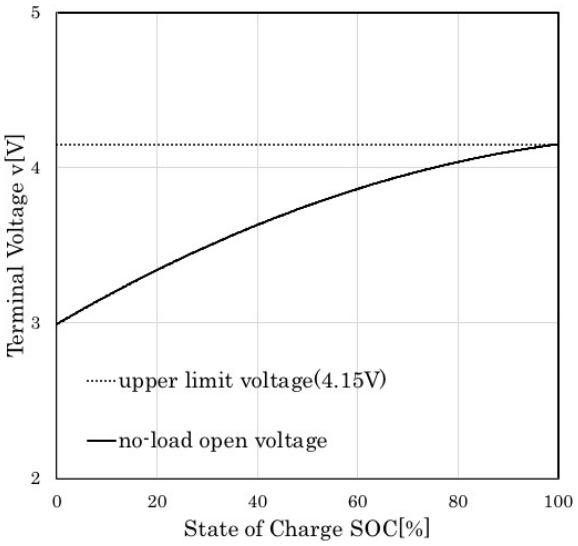


Fig.1 No-load open voltage of LIM-30H made by GS Yuasa Co. (per a cell)
1 unit is composed of 8 cells.

$$v = a_1 SOC^2 + a_2 SOC + a_3 \\ = f_1(SOC) \quad (1)$$

Lithium-ion battery is discharged when the battery train accelerates and charged when it uses its regenerative brake. It is necessary to be careful that the battery has the internal resistance. Because of this resistance, the terminal voltage of the battery increases when using the regenerative brake. For example, when the braking train can be regarded as just a current source, it is connected simply with the battery like the circuit of Fig.2. The terminal voltage V[V] is represented by the equation (2).

$$V = v + ri = f_1(SOC) + ri+ \quad (2)$$

Hence, when the SOC is high, a reducing current have to be done because we keep the terminal voltage under the upper limit voltage. This reducing current means that the regenerative energy decreases. We call it "restriction of regenerative power". This restriction effects the total energy consumption of the battery train.

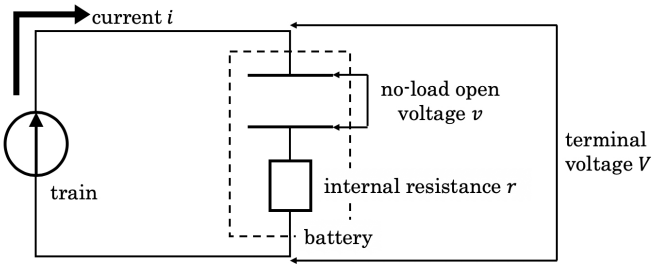


Fig.2 Simple electrical circuit between the train and the ESS (regenerative brake)

Charging infrastructure

When charging battery at the station, the circuit is almost same as the Fig.2. Although lithium-ion battery has high power, it is dangerous with high voltage. Therefore, when we charge lithium-ion battery, we apply CCCV (Constant Current Constant Voltage) control. We have identified the charge characteristic as equation (3) by the charging simulation. In consideration of the situation, we set the charging current at the constant current charging mode $i_c = 90A$.

$$i(t)[\mathrm{A}] = \begin{cases} i_c & (SOC \leq 81.4) \\ i_c \exp\left(-\dfrac{t-t_0}{\tau}\right) & (SOC > 81.4) \end{cases} \tag{3}$$

We got the current capacity $I(t)$ by integrate current. SOC can be calculated from current capacity easily.

$$I(t)[\mathrm{Ah}] = \int_0^t i(t)dt = \begin{cases} i_c t & (SOC \leq 81.4) \\ i_c t_0 + \tau i_c \left\{1 - \exp\left(-\dfrac{t-t_0}{\tau}\right)\right\} & (SOC > 81.4) \end{cases} \tag{4}$$

Where, t_0 is the end of time constant current charging mode and τ is time constant which express how fast the current is reduced in constant voltage mode.

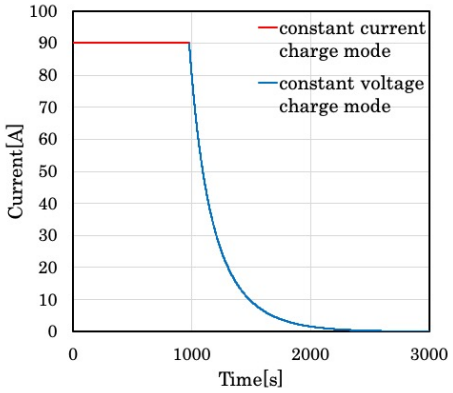


Fig.3 Current characteristic of charging to represent in (3)

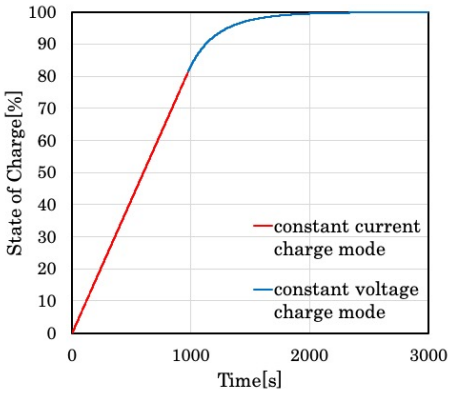


Fig.4 SOC characteristic of charging to represent in (4)

Running Simulation

Now, we conduct the running simulation to investigate how much the energy consumption changes with the SOC and restriction of regenerative power. Table1 shows the status of the battery train which we use in the running simulation.

Table1 Status of battery train

Composition	two-cars
Weight (vehicle)	80.0 ton
Speed control	variable voltage variable frequency
Braking	regenerative brake / mechanical brake
Rated Voltage of the circuit	633.6 V
Capacity of the battery	300 Ah
Numbers of battery units	22 series / 10 parallel
Weight (battery)	4.3 ton
Maximum acceleration (at the rated voltage)	2.0 km/h/s
Maximum deceleration	3.6 km/h/s
Efficiency (acceleration)	90 %
Efficiency (braking)	80 %

The simulation goes following procedure.

(i) Maximum acceleration up to the max speed
(ii) Coasting at the point from where the train can stop at the target point using the maximum deceleration
(iii) Maximum deceleration using the regenerative brake as much as possible (If the terminal voltage reached at the upper limit voltage, the current of train should be reduced and the mechanical brake is used which makes the same deceleration as the regenerative brake.)

This operation of speed control is based on the general principle of the energy-saving driving (Licheng et al. 2017). We conducted the running simulation on the typical route changing the Initial SOC of the battery. This route is 3km long and has -3‰ gradient. The result of energy consumption is showed in Fig.5. As showed in Fig.5, the characteristic of the energy consumption draws a downward curve line to the initial SOC. When the SOC is high, a reduction of regenerative power occurs and the total energy consumption increase. Contrary, the traction force of battery train fall when SOC is low. It takes longer time to accelerate to the same max speed than the high SOC. As a result, the energy consumption increases with the low SOC. Needless to say, the energy consumption is inversely proportional to running time. Then, we can plot the characteristic of energy consumption as the two-variable (initial SOC and running time) function in Fig.6. For ease of viewing, the axis of energy consumption is inverted. This function draws downward curve surface in a three-dimensional space. This characteristic is used as a constraint condition in the definition of MILP from the next chapter.

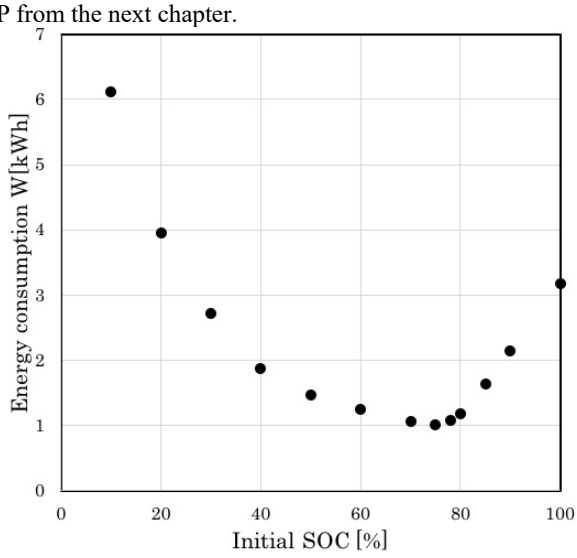


Fig.5 Relevance between Initial SOC and energy consumption

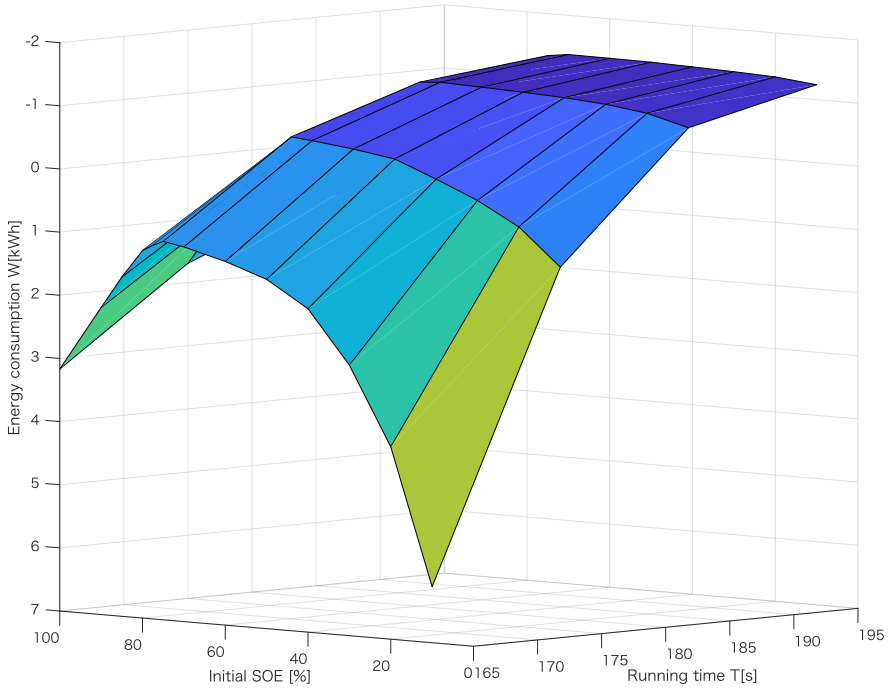


Fig.6 Characteristic of energy consumption among running time and Initial SOC

4. Model to Minimize the Energy Consumption on a Route Section Containing Multiple Stations

As described in chapter 3, we got the characteristic of energy consumption among running time and Initial SOC. The plan which gives an appropriate allocation of running time and SOC of battery train is necessary to enhance energy efficiency for catenary-free transportation system. One well documented method to solve planning problems is to use mathematical programming. Optimization is an often-used method in previous literature to create timetables. In this chapter, we present an optimization model in which allocate running time, control SOC, and also determine the charging point to minimize the energy consumption when a single battery train run a route section containing multiple stations and charging point. This model is originally based on the optimization model for allocating running time presented in a previous research (Miyatake et al. 2012).

Introduction State of Energy:
As an indicator of the state of the ESS, we used SOC which is the value of current capacity as a percentage. It is because that current capacity is generally used to express the voltage characteristic. From here, to make ease the relationship between the state of the ESS and

energy consumption, we use SOE (State of Energy) which is the value of energy capacity as a percentage instead of SOC. Energy capacity is simply calculated from current capacity to multiple the terminal voltage of the battery.

Definition as a Mathematical Programming

In the model we assume the route model shown in Fig.7. The number of stations is N. Variables are divided into two groups. Each station has variables in the first group. SOE_k represents how much energy is remained in the battery at station k. p_k is the binary. It shows that whether the battery train charge at the station or not. D_k is dwell time at each station. In this programming, D_k is not variable but given value. The second group of variables is about each section between stations. There is energy consumption W_k and running time T_k. It seemingly unnatural to set W_k to variables because we illustrate energy consumption depends on running time and SOE. The reason of this will be explained later.

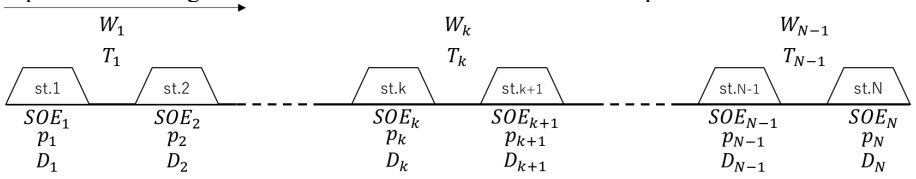


Fig.7 Route model and variables

Parameters:

N	= number of stations
T	= sum of running time of each station
M	= maximum number of the sum of charging points.
D_k	= dwell time at the station k
SOE_{min}^k	= minimum value of SOE at station k.
SOE_{max}^k	= maximum value of SOE at station k.
T_{min}^k	= minimum value of running time at station k.
T_{max}^k	= maximum value of running time at station k.

Variables:

SOE_k	= SOE at station k
p_k	= indicates if the battery train charge at station k (=1) or not (=0)
W_k	= energy consumed between station k and station $k+1$
T_k	= running time between station k and station $k+1$

Objective function:

The objective function (5) is the sum of the energy consumption at each station and the margin of SOE between the starting station and the terminal station. To add this margin to the objective function, the energy for using charging can be considered.

$$Minimize \sum_{k=1}^{N-1} W_k + c_1(SOC_1 - SOC_N) \quad (5)$$

Constraints:

The following constraints are used in the optimization model to restrict the train running, energy consumption and the SOE at each station. The constraint (8) represents the transition of SOE between stations. The SOE at the station $k+1$ is calculated like that the SOE at the

station k minus the energy consumption between station k and station $k+1$, and plus the charging energy at the station k. If the battery train charge at the station k ($p_k = 1$), SOE recovers in proportion to dwell time D_k. It is meaningless to charge at the starting station and the terminal station. Hence, the binary variable p at these two stations should be zero by the constraint (11).

$$\sum_{k=1}^{N-1} T_k = T \tag{6}$$

$$\sum_{k=1}^{N-1} p_k \leq M \tag{7}$$

$$SOE_{k+1} = SOE_k - c_2 W_k + c_3 p_{k+1} D_{k+1} \tag{8}$$

$$SOE_{min}^k \leq SOE_k \leq SOE_{max}^k \tag{9}$$

$$T_{min}^k \leq T_k \leq T_{max}^k \tag{10}$$

$$p_1 = p_N = 0 \tag{11}$$

Constraints (characteristic of energy consumption)

As the constraints, we use the characteristic of energy consumption among running time and Initial SOC got in chapter 3. We suppose the characteristic is two-variable function which determine W_k from T_k and SOE_k. But it is difficult to express this function as a explicit function. Therefore, we include the relationship among the W_{k+1}, T_{k+1} and SOE_k as the implicit function to constraints.

$$f(W_k, T_k, SOE_k) = 0 \tag{12}$$

This is the reason why we deal with W_k as the variable. The function is different by each section. We firstly conduct running simulations in the all section in the route model in the same way in chapter 3. After that, characteristics of energy consumption at each section draw downward curved surfaces like Fig.4. As shown in Fig.7, the controversial point is that this characteristic is nonlinear and cannot be included in the MILP. If we use this characteristic as it is to make a nonlinear programming, there is a possibility that the calculation time to obtain the optimum solution is too huge with respect to the scale of the problem. Therefore, in this paper, we use a method of dividing energy consumption characteristics defined on the space as curved surfaces into fine lattice shapes and approximating it as a polyhedron composed of minute triangles, thereby creating a linear condition.

Linear approximation is done in the following way. First, we discretize continuous and smooth curved surfaces $W(T_k, SOE_k)$ by m samples in the T-axis direction and n samples in the SOE-axis direction. In this way, linear approximation is performed by complementing between the grid points of $m \times n$ created on the surface with a plane. As shown in Fig.6, (13) is the equation of the plane which is created by selecting three adjacent points in the lattice points.

$$a_1 T_k + a_2 SOE_k + a_3 W_k = 1 \tag{13}$$

When we choose any lattice point $(T_k, SOE_k, W(T_k, SOE_k))$, we can make 2 planes (14) and (15). Note that $W_{i,j} = W(T_i, SOE_j)$

$$① \cdots \begin{pmatrix} T_i & ISOE_{j,} & W_{i,j} \\ T_i & ISOE_{j+1,} & W_{i,j+1} \\ T_{i+1} & ISOE_{j+1,} & W_{i+1,j+1} \end{pmatrix} \begin{pmatrix} a_{11} \\ a_{21} \\ a_{31} \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \tag{14}$$

$$② \cdots \begin{pmatrix} T_i & ISOE_{j,} & W_{i,j} \\ T_{i+1} & ISOE_{j,} & W_{i,j+1} \\ T_{i+1} & ISOE_{j+1,} & W_{i+1,j+1} \end{pmatrix} \begin{pmatrix} a_{12} \\ a_{22} \\ a_{32} \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \quad (15)$$

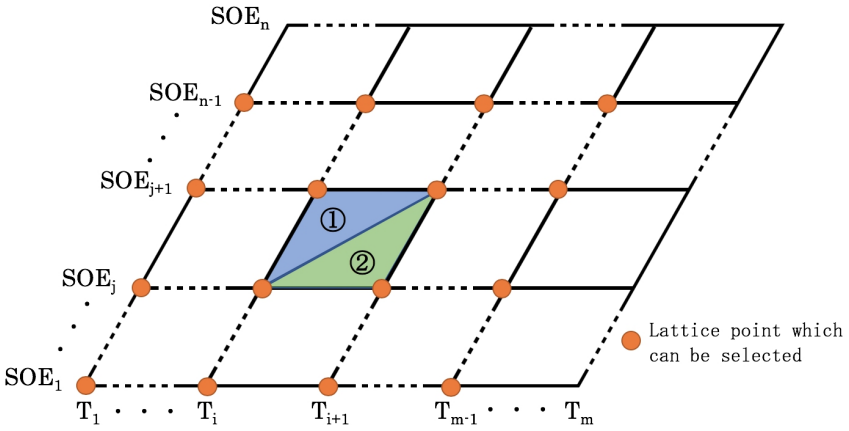


Fig.8 Polyhedron approximation of curved surface

In this way, $(m-1) \times (n-1)$ grids points are selected. Linear Constraints made by this approximation are added to equation (8). Furthermore, to regard the characteristic as a polyhedron, there is no problem to replace the equation (12) to the inequality (16).

$$f(W_k, T_k, SOE_k) \leq 0 \quad (16)$$

Even in this way, if energy consumption is minimized inevitably it gives the same result as the equation constraint, so there is no problem as inequality constraint. By doing this, we can get even more advantage. It is not necessary to define the domain of the approximated plane with respect to the convex constraint, it is possible to prevent the constraint expression from becoming complicated.

Coefficients:
Coefficients c_1, c_2, c_3 has each physical meaning and should be set properly. Mainly, these coefficients convert units between some values (e.g. SOE and W).

5. Case Study

A numerical experiment is performed for real-world case with data from the Japanese suburban line, see Fig.9. The model line is that cannot run without the train battery charging at a station. The total length of the line is 90km assuming an actual local line. We intend to optimize running time and make an energy-efficient timetable for battery train using MILP defined in chapter 4.

5.1 Default timetable

Firstly, we conducted running simulation to determine the default running time at each section and make the default timetable. For initial running time is that the battery train runs this section with a maximum speed 90km/h. Speed profile is determined by the general energy-efficient theory shown in the chapter 3. However, there are some section where the train stop at the front of the next station. In this case, traveling with maintaining the

maximum speed is inserted before the battery train decelerate as shown in Fig.10. Moreover, dwell time at each station is set with reference to the actual operation. Dwell time at each station is shown in Table2. The blue line in Fig.11 shows the default timetable. As the feature of the default timetable, there is a comparatively long dwell time at the station H. This is because this route section has only a single track and station H is equipped for two-way transportation.

Secondly, we investigate the characteristic of energy consumption by running simulation. We set the maximum speed 120km/h and minimum running time with this maximum speed. The simulation is performed while changing running time to 10 second increments. Maximum running time is set so that the difference from the default running time is equal to the difference between the minimum running time and the default running time. For example, if the default running time is 200 second and the minimum running time is 170 second, the maximum running time should be 230 second.

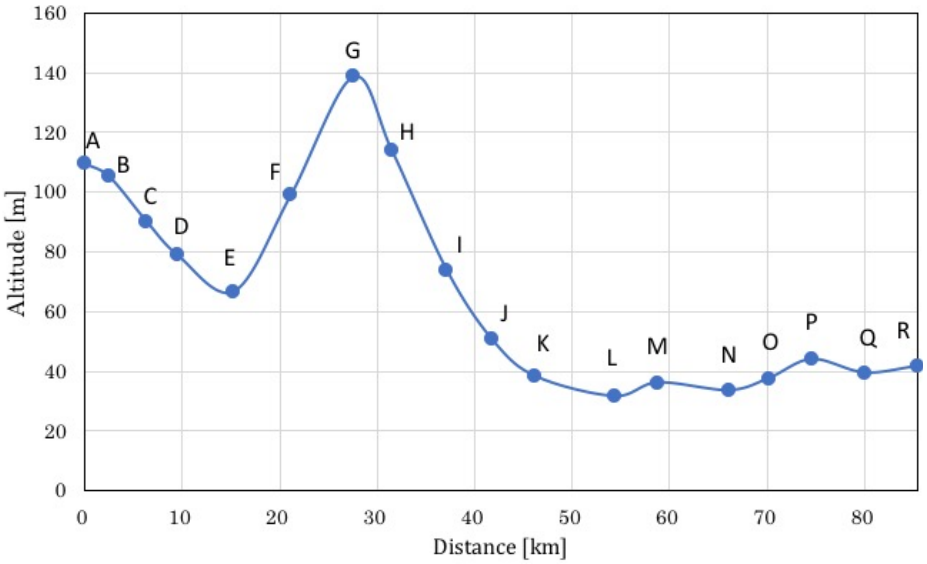


Fig.9 Model of line, the point A, B, C … means the stations

Table2 Dwell time at each station

Station	A	B	C	D	E	F	G	H	I
Dwell time[s]		90	30	50	110	40	40	430	30
	J	K	L	M	N	O	P	Q	R
	30	60	40	40	30	30	80	60	

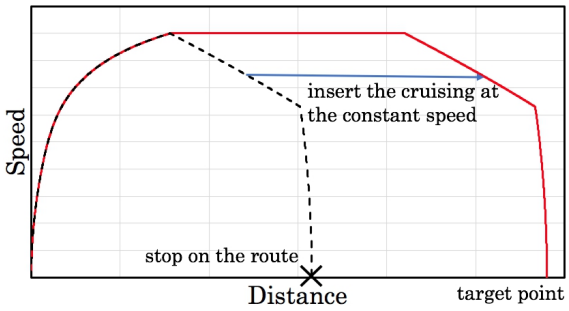


Fig.10 Operation when the train stop at the front of the next station

5.2 Optimization and result

With these conditions, the optimization is defined as a MILP in the way shown in chapter 4. As an additional condition, the upper limit of the number of charging point is set 2. The upper limit of SOE is set 95% and the lower limit 20%. This MILP can be solved by the solver: “intlinprog”. This is included in “Optimization Toolbox” of the MATLAB.

The calculation finishes in the several seconds. We show the optimized timetable in Fig.11. Furthermore, according to the allocating running time, second running simulation is conducted to get the speed profile of the battery train and more accurate transition of SOE and energy consumption. We show the speed profile in Fig.12, the transition of SOE in Fig.13 and the transition of the integrated value of the energy consumption in Fig.14.

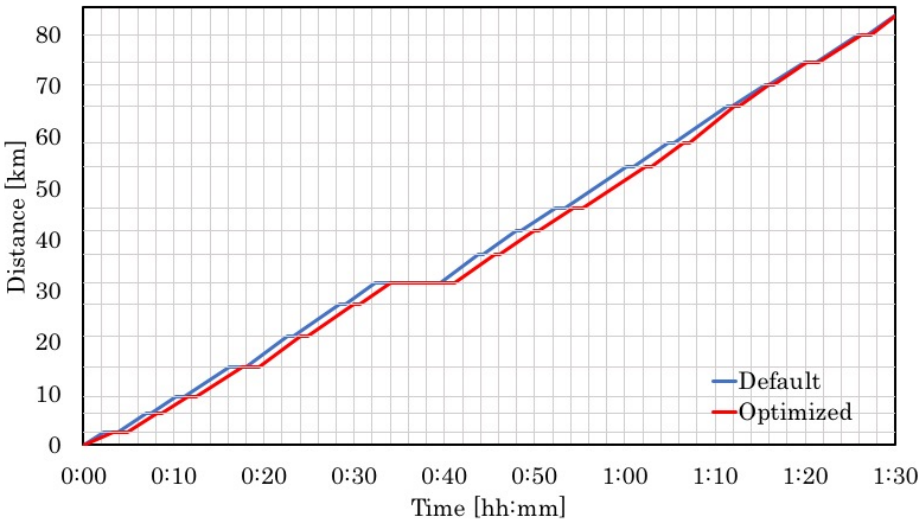


Fig.11 Default and optimized timetable

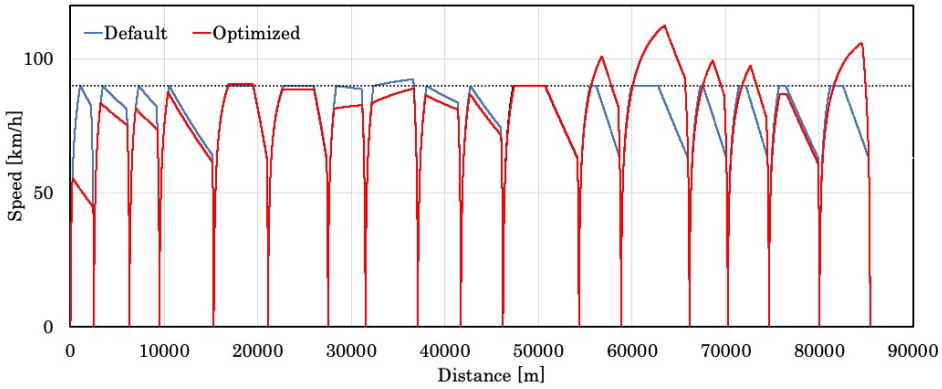


Fig.12 Speed profile according to the default and the optimized timetable

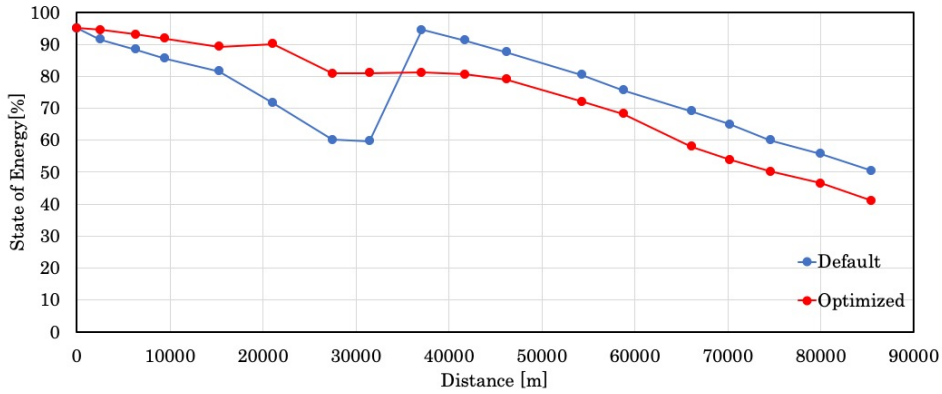


Fig.13 Transition of SOE of the default and the optimized timetable

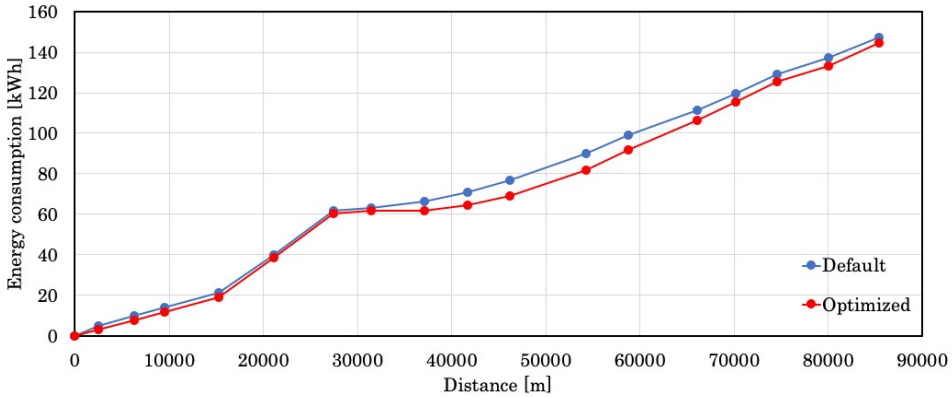


Fig.14 Transition of integrated value of the energy consumption of the default and the optimized timetable

Table3 Station performed charging

Default timetable	H (the 8th station)
Optimized timetable	E (the 5th station)

Table4 Total energy consumption of the default and optimized timetable

Default timetable	147.1kWh (100%)
Optimized timetable	144.6kWh (98.3%)

6. Discussion

By the optimization, running times are adjusted up to 60 seconds at each section. The location of the charging point is also changed. The total energy consumption is reduced about 1.7% by the adjustment. In this chapter, we discuss this adjustment and energy-efficient timetable for catenary-free transportation with battery train.

Location of charging point
In the default timetable, charging is performed at the station H. We can see a big recovery of SOE in the blue line of Fig.13. On the other hand, in the optimized timetable, charging point moves to the station E. The station E is located in the valley on the route model. Usually, trains consume much more energy on the uphill section than the downhill section. It can be said that the energy consumption has reduced because battery had been charged before uphill section that needs much electric energy. This change has the similar feature to the related study (Noda et al. 2015). It can be referred for the detail explanation.

Allocation of running time
See Fig.11. In the optimized timetable, running times in sections of the first half are extended. Contrary, running time in sections of the last half are shorten. The reason of this adjustment is SOE dependence as shown in Fig.5. Although the SOE characteristic depends on the route condition, it has the convex characteristic and the most energy-efficient point is at 75%. This is the highest point where the battery train can absorb all regenerative energy without restriction of regenerative power. On the red line (optimized timetable) in Fig.13, the battery train runs fast when the SOE is around 60-80%. Hence, we got the acknowledge that it's necessary to keep SOE around the most energy-efficient point to minimize the energy consumption. We calculate the average of SOE while running between the starting station and the terminal station. The average value is showed in Table5.

Table5 Average value of SOE while running

Default timetable	75.76% (+0.76% from 75%)
Optimized timetable	74.91% (−0.09% from 75%)

As shown in Table5, the average value of SOE of the optimized timetable is nearer than the default's one. This control to keep SOE around the most energy-efficient point is important to make a time table efficiently for catenary free transportation with battery train. Still, in this case study, it is not sufficient to evaluate the energy saving effect of this optimization because the effect depends on the initial timetable. Currently, catenary-free transportation is not popular in Japanese railway. Therefore, in this paper, we make the initial timetable by ourselves, and the it is difference from the timetable used in the actual operation. Moreover, the consideration of dwell time is necessary. If we add dwell time to

variables this optimization program is not defined as a linear programming but a nonlinear programming. However, there is a possibility to get a new acknowledge if we can adjust simultaneously running time and dwell time so that the sum of those is constant. The effect of the number of the charging points is also considerable.

7. Conclusion

In this paper, we focus on the energy-efficient timetable for catenary-free transportation system with battery train. It is significant for catenary-free transportation system that how much regenerative energy can be absorbed in the ESS. Firstly, we investigate the characteristic of SOE by running simulation. It is clarified that the characteristic draws a convex curved line to SOE. We can draw the convex curved surface to add running time as the second parameter in the three-dimensional space. Secondly, we make the model to minimize the energy consumption and make the energy-efficient timetable. Although the characteristic of the energy consumption is nonlinear, we define this optimization programming as MILP to use a polyhedron approximation. Finally, we perform a numerical experiment using the data from real-world route model. Consequently, the optimized timetable reduced the total energy consumption by 1.7%. The optimization result of the location of charging point is at the station with a low altitude, and this is similar result to the related study. As a characteristic of the running time allocation, running time in the section where the SOE is around the most energy-efficient point is shorten. In future work, it should be more case studies in the more variable conditions to enhance this method.

There are many studies to design a suitable and energy-efficiently transportation system in the field of railway. It is expected further reduction of electric energy in catenary-free transportation and battery train.

References

Licheng, T., Tao, T., Jing, X., Shuai, S., Tong, L., 2017. "Optimization of train speed curve based on ATO tracking control strategy", In: *Chinese Automation Congress (CAC2017),* Jinan*,* China.

Noda, Y., Miyatake, M., 2016. "Methodology to Apply Dynamic Programming to the Energy-Efficient Driving Technique of Lithium-ion Battery Trains", In: *Electrical Systems for Aircraft, Railway, Ship propulsion, Road Vehicles, International Transportation Electrification Conference, International Transportation Electrification Conference*, *(ESARS-ITEC 2016),* Toulouse, France.

Pena-Alcaraz, M., Fernandez, A., Cucala, A, P., Ramos, A., Pecharroman, R, R., 2012. "Optimal underground timetable design based on power flow for maximizing the use of regenerative-braking energy" The Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit July 2012 vol. 226 no. 4 397-408.

Sicre, C., Cucala, P., Fernandez, A., Jimenez, J. A., Ribera, I. and Serrano, A., 2010. "A Method to Optimise Train Energy Consumption Combining Manual Energy Efficient Driving and Scheduling", *Computers in Railways XII*, WIT Press, pp. 549–560.

Li, X., Lo, H, K., 2014. "Energy minimization in dynamic train scheduling and control for metro rail operations" Transportation Research Part B: Methodological, Volume 70, pp.269–284.

Ishino, K., Sakamoto, K., Miyatake, M., 2012. “Energy-saving Operating Strategy of a Catenary Free Light Rail Transit”, In: *15th International Conference on Electrical Machines, Systems (ICEMS 2012)*, Hokkaido, Japan.

Miyatake, M., Kuwahara, R., Nakasa, R., 2012. “A simple adjustment of runtimes between stations for saving traction energy by means of mathematical programming”, In: *Computers in Railways XIV (COMPRAIL 2012), New Forest, UK.*

Andersson, E, V., Peterson, A., Krasemann, J, T., 2015. “Improved Railway Timetable Robustness for Reduced Traffic Delays – a MILP approach”, In: *6th International Conference on Railway Operations Modelling and Analysis (RailTokyo2015)*, Chiba, Japan.

Arenas, D., Chevrier, R., Hanafi, S., Rodriguez, J., 2015. “Solving the Train Timetabling Problem, a mathematical model and a genetic algorithm solution approach”, In: *6th International Conference on Railway Operations Modelling and Analysis (RailTokyo2015)*, Chiba, Japan.

Chao, F., Liao, Z., Liu, S., Xun, J., Luo, X., 2016. “A Two-Layer Construction of Energy Optimization Approach for Timetable” In: *18th IEEE Mediterranean Electrotechnical Conference (MELECON 2016)*, Limassol, Cyprus.

Noda, Y., Miyatake, M., 2015. “Rational placement of charging-station for Lithium-ion battery onboard of rail vehicles” In: 7th International Symposium on Speed-up and Sustainable Technology for Railway and Maglev Systems (STECH 2015), Chiba, Japan.

Seyama Y., Nakamoto T., Nishiyama K., and Sonoda T., “Development of Forced-air-cooling Lithium-ion BatteryModule "LIM30H-8R" for Railway Applications”, *GS Yuasa Technical Report, Vol. 4, No 2, pp.24-29 (2007) (in Japanese).*

Train Slots: A Proposal for Open Access Railways

Martin Scheidt
ORCID: 0000-0002-9384-8945
TU Braunschweig, Germany,
Institute of Railway Systems Engineering and Traffic Safety

Abstract
This paper seeks a concept to include fixed-interval paths with manageable train slots to satisfy the flexible needs of freight traffic in a strict fixed-interval passenger timetable. The primary method is constituted by literature review and theoretical slot construction. The terms timetable and railway operation are specified and illustrated. Different levels of timetables will be discussed and further developed. Two concepts, slots and pulses, are described together with precondition and modelling to accomplish a mixed timetable level for flexible freight traffic and fixed passenger traffic. Finally, a comparison of the timetable levels with the rail freight corridor Rhine-Alpine is presented. In conclusion, three points for further research are made, and an experiment is suggested to validate the result in the future.

Keywords
railway operation, railway timetable, timetable path, timetable slot, transit layer

1 Introduction

Timetables are a vital component in satisfying transport needs in railway systems. This paper discusses timetable concepts and their dependencies on railway operation. It is part of an ongoing study about how to determine infrastructure from a given timetable. In Scheidt (2018), an approach in layers was presented to reduce the solution space for the transition between infrastructure and timetable. Two layers contain timetable information to organise traffic on the infrastructure: the transit and the transport layers. The transit layer contains the paths, and the transport layer contains the rail services that use the paths. While there is a suitable model for the transport layer, there is no viable model description for the transit layer.

The current lack of the model results from the representation of the variability of train movements in combination with the actuality of the operational disposition. The variability of train movements was the focus of several works. Pöhle (2016) points out that different interval and stopping patterns need to be taken into consideration for the operating varieties. Medeossi, Longo, and Fabris (2011) extends the standard blocking time with performance parameters to include these variations. Hertel and Steckel (1992) illustrates that these parameters also vary significantly between passenger and freight trains. Roos (2006) introduces a dispatching concept to dispatch trains through a junction. Caimi et al. (2009) shows that compensation zones can help to stabilise variations in a train movement along a line section before the train enters a junction. However, typical timetables from the field of operations research form the conditions of the railway operation inadequately. The impact on the railway operation and the ability to guide traffic flow have to be considered to develop timetable concepts further.

The conditions of railway operation include that planned train movements have different probabilities for how they occur. Passenger trains will most likely run as planned, whereas freight trains can have a shifting occurrence. The reaction from railway operation is to use dispatching extensively for freight trains while trying to enable passenger trains to be as punctual as possible. The dual nature of traffic, together with the planning of train archetypes for timetables, shape the problem set that railway operation has to overcome. This leads to the research question (Q_0): How to handle variations in the traffic mix with a timetable as a working base for railway operation?

The goal is to evaluate the impact of timetable concepts down to specific train movements on routes and tracks. Several subquestions can be formulated. The following three questions were used to guide the answer:

(Q_1) Are there different kinds of timetables, and can they be structured?

(Q_2) What is railway operation, and how does a timetable affect it?

(Q_3) What types of paths in a timetable can be used to accommodate variations?

The principal method consists of a literature review and the theoretical construction of train movements. The goal of the construction is to identify use cases for operators. These use cases have to be visualised in such a way as to improve the ability of an operator to capture the situation. A proposed timetable with new types of paths has to be able to contain the duality of the traffic without favouring one over the other.

This analysis will describe what will be necessary for such a timetable to be included in the previously mentioned layer model and to enhance planning without capacity impact on the infrastructure. The main result is a formulation for the specifications and data required for the transit layer and the impact on the transport layer. The specifications result from the connection between timetables and railway operation, and to comprehend the connection further, both terms must be specified.

This paper is structured as follows: Section 2 describes and defines the relation of railway operation and timetables. In Section 3 categories including the need for further development are concluded from current timetables and the timetable life cycle. From there, concepts for line segments (Section 4) and junctions (Section 5) are formed into a future timetable category and followed by the preconditions for such a timetable category in Section 6. Finally, an exemplary use case is indicated in Section 7; followed by the conclusion.

2 Timetable and Railway Operation

A timetable is the sum of all planned train movements. The planning of the train movements is known as scheduling and is considered a large field for operations research. Scheduling has to fulfil requirements such as synchronising between trains, preventing severe congestion including deadlocks (stability) and being without conflict (feasibility). The requirements are satisfied by the structure of the timetable. For instance, to prevent severe congestion, the allocated timetable capacity is usually less than the actual line capacity (Pachl 2018, p. 179).

The timetable presents access rights to the infrastructure by means of the timetable authority. Furthermore, it also has to provide a working base for railway operations (Pachl 2018, p. 27). Different views on the fully scheduled timetable are used to fulfil traffic demands by passengers or freight and guide the traffic flow by the operators (see Figure 1).

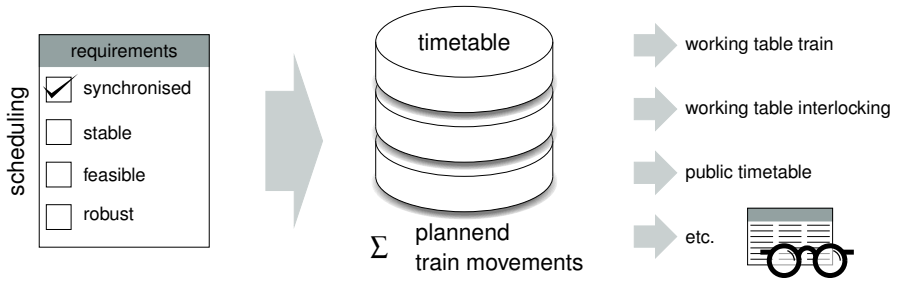


Figure 1: Requirements and views of a timetable.

The operation control of an operator is considered a part of the railway operation. Railway operation can be described as a collection of processes and functions with a focus on enabling train runs (see Figure 2). To do so, the operator observes the current operational status and places the movement requests for a train according to the dispatching rules in compliance with the timetable authority from the working timetable.

Consequently, the timetable plays an essential role in railway operation. The operator is the second tier in preventing deadlocks and coordinates the current train movements within the normal fluctuation in operation by amending the train order, if necessary. The operator can also remove the bonding between fast and slow trains by slowing down the fast one in the event of extensive disruptions and, therefore, increase the usable capacity (Pachl 2018, p. 181). He can also improve synchronisation between trains by adapting their priorities. Disruption management by the operator plays an essential part in the efficiency and effectiveness of the real-time rescheduling and in compensating the disruption. For this, the operator needs freedom of action in the timetable contributed by margins.

There are several methods to integrate margins for railway operation in a timetable. A widely adopted one is documented by the UIC in the leaflets code 406 and code 451 (UIC 2000; UIC 2004). These codes suggest adding margins to a train run for circumstances like general recovery, hub margins, or track work (see Figure 3c). The UIC codes distinguish between different types of passenger or freight traffic and different types of infrastructure. However, these codes do not consider that at the time of planning the margins, little is known about the freight traffic as opposed to the passenger traffic. Nevertheless, the very essence of competitive freight traffic is that it is flexible, it cannot be planned like passenger traffic, and it should be regarded differently in the UIC codes. Consequently, a timetable should satisfy the different environments of well-planned passenger traffic and flexible freight traffic.

In timetables, path requests are used to coordinate the use of a timetable's resources, and there are two poles of path requests on open-access railways. One pole consists of path requests with known properties that can be defined as long-term. These requests are typically for paths for passenger trains with fixed-interval timetables. The other pole consists of ad-hoc traffic in need of paths where properties are only known on short notice. These requests come mainly from freight operators, where the business model depends on flexibility. These two poles represent the contradictory nature a timetable has to overcome: to generate ad-hoc paths for short-term freight trains and still enable long-term fixed-interval paths. From the point of view of railway operation, the main challenge is to provide the opposite of a compact timetable, with ample margins for dispatching.

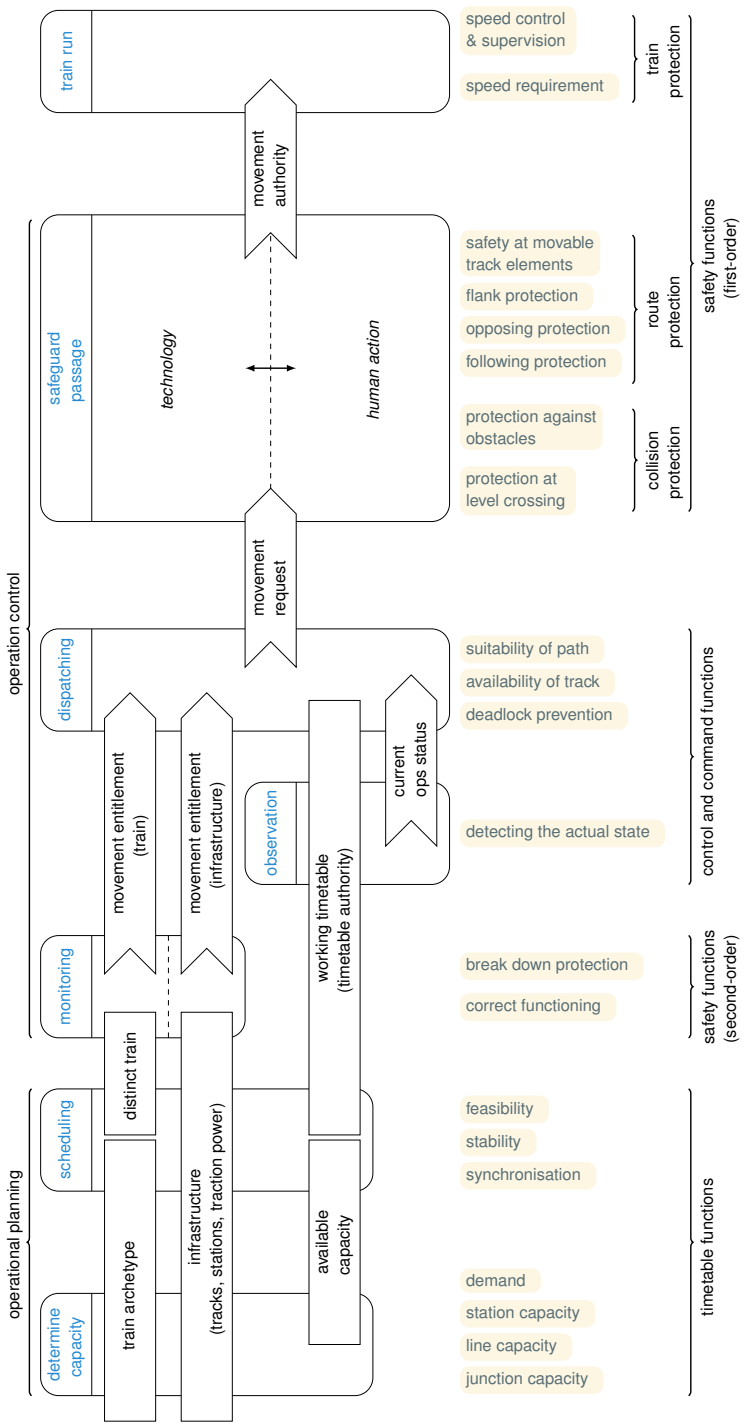
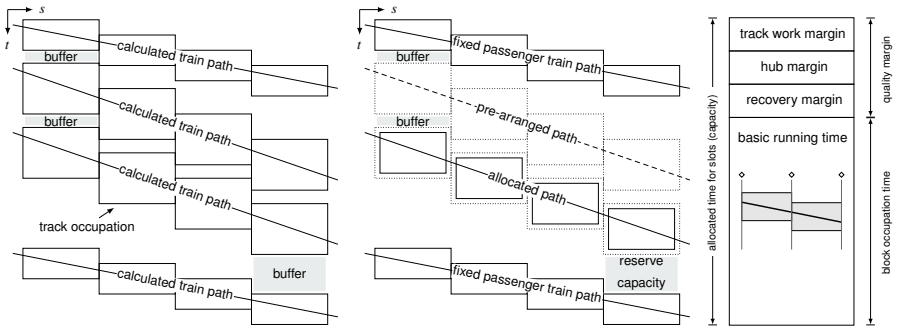


Figure 2: Process map of railway operation (Scheidt et al. 2018). Getting more specific from left (macroscopic network) to right (single train run).



(a) Level 1 timetable.

(b) Level 2 timetable.

(c) Timetable margins (based on UIC 2000; UIC 2004).

Figure 3: Timetable categories and margins

3 Timetable Categories

There are different timetable categories, as railway operation happens in a diverse environment which differs in requirements for the timetable. For some networks, a timetable can be optional. This form can be encountered in the tram system: A timetable still has a purpose for passengers, but it is not needed to prevent deadlocks or severe congestions. This characteristic is due to the comparatively simple network, which is usually not deadlock-prone. A timetable can be categorised into several levels to distinguish between the properties and relevance of different timetable concepts (see Figure 3). Level 0 shows the cases in which timetables are optional.

Level 1 encompasses classic timetables for railway systems (see Figure 3a). Here, the timetable helps with coordinating over an extensive network. In many cases, a timetable may also serve essential safety functions. This type of timetable consists of individual calculated paths. If there are temporary restrictions on the network, paths may require reconstruction. Such Level 1 timetables have existed in Europe for decades and come with a mutually agreed life cycle (RNE 2014). In a Level 1 timetable, all calculated paths have to be included in the scheduling phase, which is closed before a yearly timetable takes over and ends with the timetable handover. This procedure favours the long-term planning of paths and does not favour short-term path requests.

In Level 2, European Union legislation enforced a regulation to overcome favouritism (EU 2010). Here, a timetable consists of common calculated paths for long-term traffic and particular pre-arranged paths for likely traffic (see Figure 3b). The timetable life cycle is amended and includes regular path requests, which have to be made at least eight months in advance of the timetable period. The ad-hoc request phase starts two months before a yearly timetable takes over and ends with the timetable handover (see Figure 4). The timetable also includes reserve capacity to further enable ad-hoc traffic. The organisational structure to manage path request and methods for scheduling can be inherited from Level 1. The operation of such a Level 2 timetable raises questions about how to cope with used or unused paths for operators. Operators have to identify reserve capacity that will not be used for

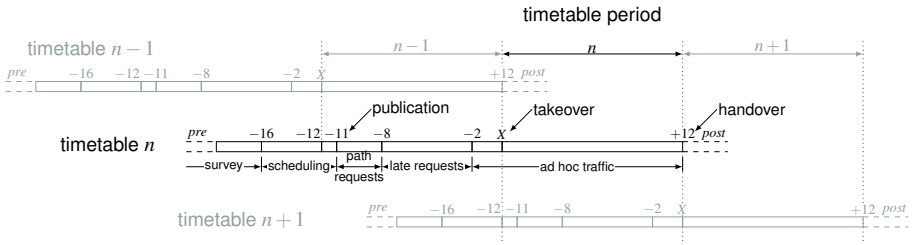


Figure 4: Timetable life cycle (based on RNE 2014).

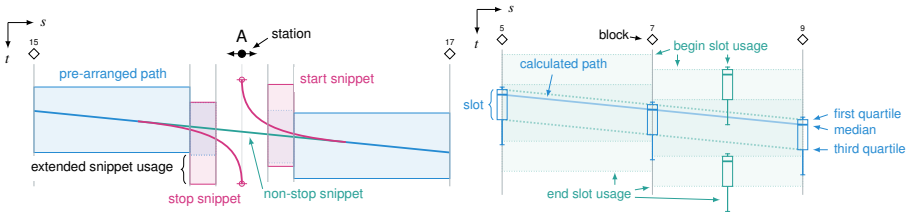
ad-hoc traffic so they can use it for dispatching. They could have to identify trains running with the same train number but different destinations or vice versa.

Currently, the answer to the question mentioned above is to project the Level 2 timetable to a Level 1 timetable in small time slices since most of the railway operation management systems predate the EU legislation. However, if the Level 2 timetable, as a working basis for railway operation, can provide information about the actual use of pre-arranged paths and reserve capacity, there could be the aptitude to predict and reduce route conflicts. This provided information could have an impact on train priorities, dispatching rules, and real-time rescheduling, disturbance, and disruption management.

4 Line Section Slots

The Level 2 timetable has to work with a train archetype since little is known about the trains during the scheduling phase in the timetable lifecycle (Weigand and Heppe 2013, p. 460). These train archetypes could be depicted as cohorts composed of actuals trains from the past. There will be regional differences, and, therefore, the cohorts are best collected subsidiarily. However, cohorts can be an access impediment and will, therefore, need a legal structure to ensure neutrality between infrastructure managers and railway undertakers. Performance parameters can be used to retain the cohorts (Medeossi, Longo, and Fabris 2011). Boxplots can be used to simplify handling in railway operation. The 50% (median) and the 80% (third) quantiles have proven to be useful for representing a time slot for a pre-arranged path (Schittenhelm and Richter 2009, p. 12). There will be different needs for quantiles depending on the archetype train. To distinguish between a regular pre-arranged path and a pre-arranged path with quantiles, here, the terms train slot or slot is used (see Figure 5b).

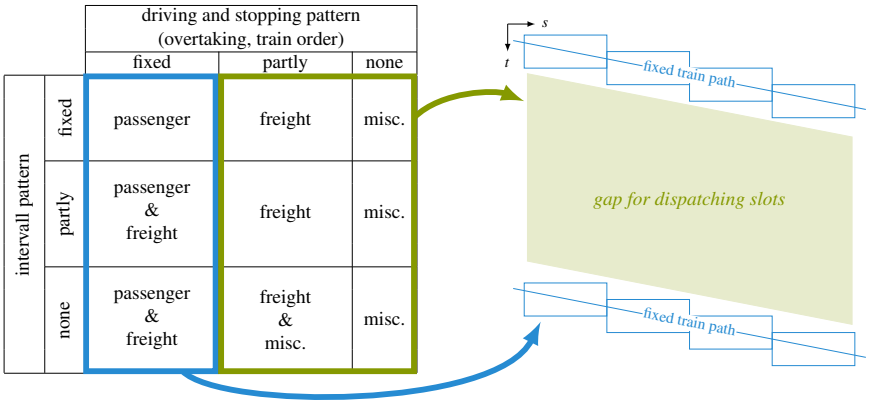
Pöhle (2016, p. 68) shows that different intervals and stopping patterns need to be considered for railway operation, and these two patterns can be attributed again into fixed, partly fixed, and non-existing patterns. Although this leads to nine types of path classifications (see Figure 6a), two concepts can be used to accommodate them: classical fixed train paths and slots. The fixed train paths create a gap in a timetable, which can be used to arrange slots. Instead of filling the gap between the fixed train paths with pre-arranged paths, the gap can be kept open for dispatching and preferably provide further information for the operator regarding how to use the gap efficiently. The concept of a dispatching gap, together with railway operation management for slots, is an advancement to a Level 3 timetable (see Figure 6b). Other concepts for arranging and allocating these slots are needed to take full advantage of slots within the free capacity.



(a) Pre-arranged path with snippets at a station.

(b) Slot, calculated path, and boxplot.

Figure 5: Slots and snippets.



(a) Classifications of paths for central Europe.

(b) Level 3 timetable.

Figure 6: Paths classifications and their accommodation in a Level 3 timetable.

The concepts required for arranging and allocating slots are:

1. Slot parameters for train archetypes to fit distinct trains.
2. A slot must fit several train run events, like acceleration, deceleration, and non-stop train run.
3. Dispatching rules for train archetypes to grant and organise timetable authority.
4. Differences between lines and junctions for the arrangement of slots.
5. Dispatching rules for traffic jams and matching distinct trains to their train archetypes.

A solution to the obstacle to fitting several train run events (like acceleration, braking, and non-stop drive through stations) into slots has been presented by Pöhle (2016, p. 68). He describes the solution as snippets, where each snippet represents one possible train run event (see Figure 5a). Snippets will increase the blocking time usage compared to a non-stop train

run. The increase should not impede the recommendation from UIC code 406 that it is not to exceed more than 60% to 85% for occupation time against a time window (UIC 2004, p. 19). The extended blocking time usage by snippets, together with the cohorts of the slots, could be seen as a kind of buffer time under the UIC code 406.

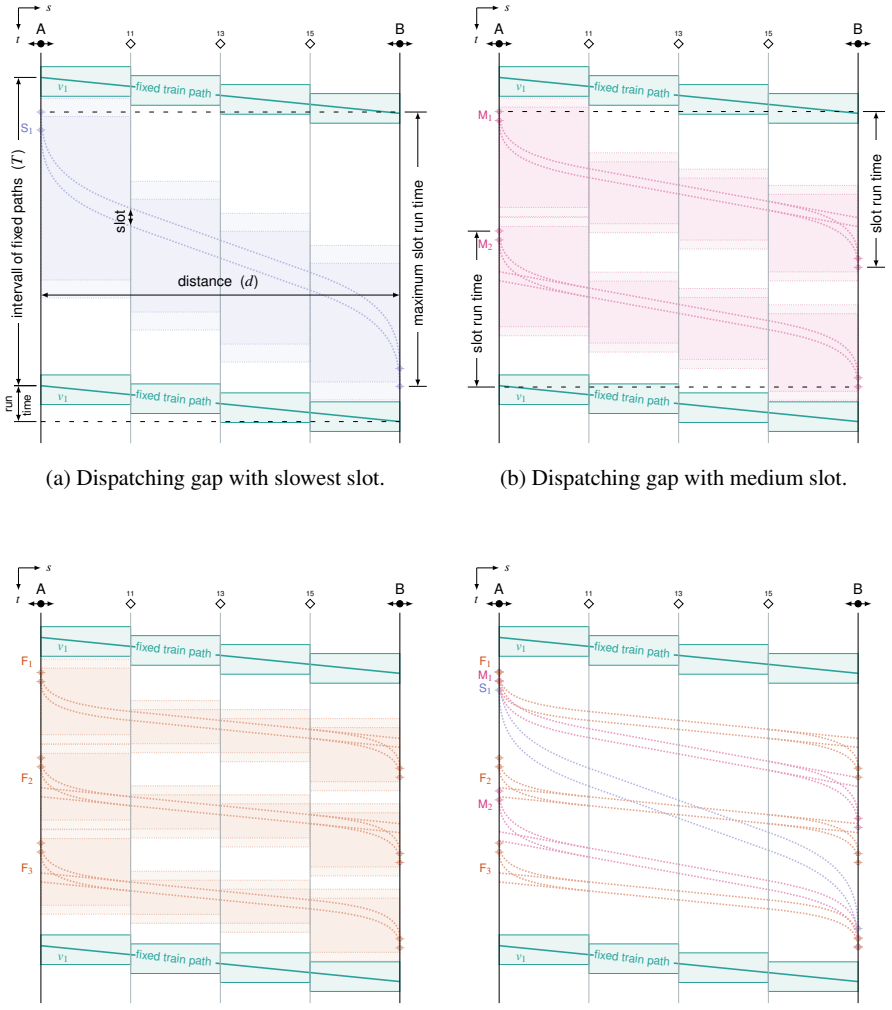
The utilisation of the dispatching gap depends on various factors, such as the snippets, the distance between two stations, the interval between two fixed paths, the slot cohorts, and the run time of the slot (see Figure 7a). The snippets can be used to include different slots with a different driving regime into the dispatching gap. A fast train can vacate the second station early, which permits a second or a third train with the same properties to pass through the line (see Figure 7b and 7c). Slots are supposed to overlap and include a mechanism for preclusion through dispatching rules to utilise the dispatching gap fully. For example, the preclusion mechanism will disable the medium and fast slot if the slow one is used; however, it could permit a combination of a medium and a fast slot (see Figure 7).

It is necessary to consider different numbers of slots for different traffic times during a day in railway operation and hence alter the dispatching gap. Traffic schemes can be used to control the number of available slots per scheme by adapting the size of the dispatching gap and, therefore, be suitable for an adaptable number of fixed train paths. The number of fixed train paths can be increased or decreased for peak traffic times since the passenger traffic usually follows a characteristic pattern (see Figure 8). Similar kinds of traffic schemes can also be implemented for temporal restrictions to automate timetable generation in cases of track work or incidents where the capacity must be restricted, assuming alternative lines with spare capacity.

5 Junction Pulses

Infrastructure limitations have to be considered to utilise the dispatching gap further. For the utilisation, dispatching rules must differentiate between line sections and junctions. Line sections are parts of a network where trains cannot change their order, and signalling is merely used for spacing the trains (i.e., block signals). Junctions are parts of the network where turnouts require interlocking systems with routes. A junction usually stretches from a route signal to the route clearing points. If the junction is signalled bidirectional, the junction can be extended between the opposing route signals for easy recognition. The junction may be called interlocking limit in conformity to North American railway operation rules. The reason to differentiate between line sections and junctions is due to the different properties and restraints in railway operation. A line section can be viewed as a queue that can be used by trains in accordance to blocking time theory, and the trains will only interact with trains on the same line sections in sequence. Junctions, on the other hand, introduce restrictions not only on the sequence of trains but also on the simultaneity of trains in multiple line sections. Consequently, a network with stations and lines can be fragmented into junctions and line sections (see Figure 9a and 9b). Furthermore, a network can be modelled for dispatching rules and a Level 3 timetable into a timetable network equivalent (see Figure 9c).

Line sections can be accommodated with slots. However, bare slots will only partially work for junctions, since slots only account for sequences. Slots have to be supplemented to consider simultanous restraints of junctions during railway operation. Railway operation restraints for junctions are similar to an interlocking matrix (see top of Figure 10a). Two opposing trains will cause railway operation states of restriction on each train: wait/release, pull-in/pull-out, counter movement, and route crossing. The wait/release restriction forces



(a) Dispatching gap with slowest slot.

(b) Dispatching gap with medium slot.

(c) Dispatching gap with fastest slot.

(d) Preclusion for different slots by dispatching.

Figure 7: Utilisation of dispatching gap with slots.

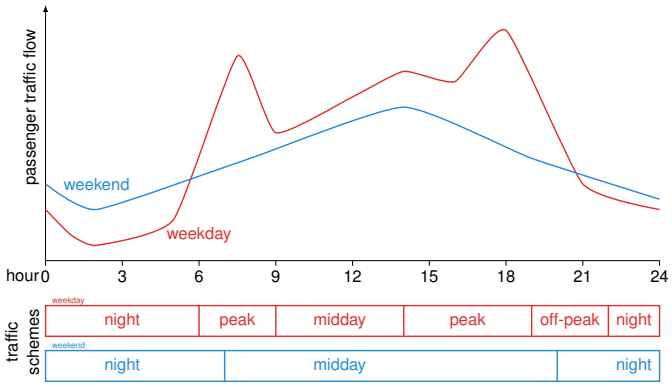


Figure 8: Traffic times and schemes in relation to the traffic flow for a timetable (fictitious).

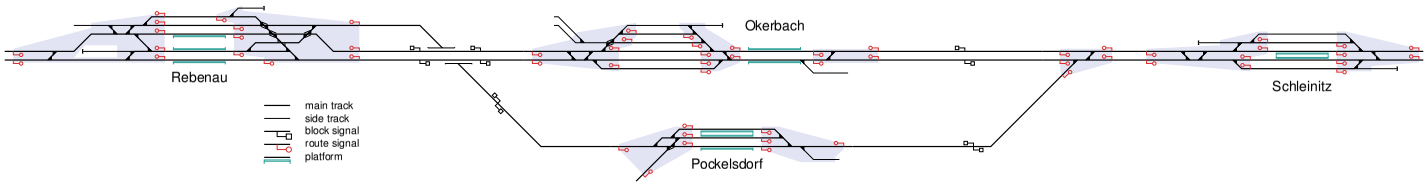
one train to wait until the other has released a track. Compared to an interlocking matrix, the railway operation restraint matrix will not be symmetric with the wait/release restriction. The pull-in restriction terms the dependency of trains if they share the same path after a turnout. The pull-out restriction applies if the trains separate their path. The combination of pull-out and pull-in restrictions at two succeeding junctions could lead to an overtaking manoeuvre. The counter movement restriction is the preliminary stage of a deadlock and has to be avoided. The involved trains for the route-crossing restriction have to be coordinated to not occur at the same time.

A junction can be parted into discrete time tranches to simplify the solution set for the coordination of the restrictions. Roos (2006) introduces the concept of pulses laid upon the junction, and describes pulses as a form of screening regarding possible train runs to reduce the solution set for the scheduling process. The achievement is a smaller solution set for regularly running trains through the junction with pulses. These pulses can be used to match trains quickly and patch them through a junction without, or with limited, conflicts with other trains. Pulses enable a relatively clear possibility for arranging slots. The combination of the railway operation restriction matrix, together with slots in pulses, resemble a possible way to provide tools for dispatching and scheduling for a Level 3 timetable for junctions.

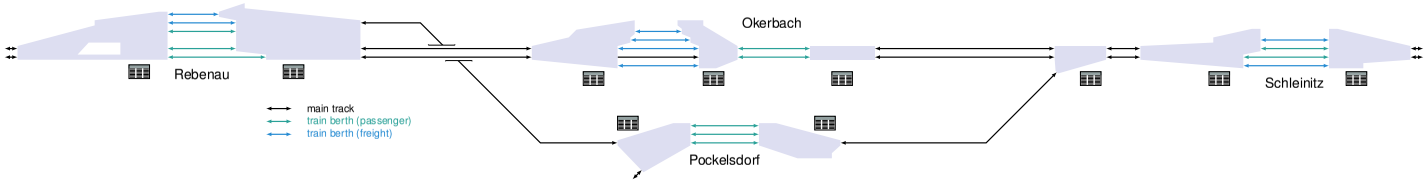
There are two kinds of pulses: common and concurrent (see Figure 10). The mutual exclusion characterises common pulses: only one train can use one pulse. Concurrent pulses can enable more than one train run depending on the railway operation restriction matrix. Figure 10b shows such concurrent pulses: trains 2–5 and 4–1 can run simultaneously while train 3–4 has to wait for the next pulse. Routes via the diverging track of a turnout usually run at a lower speed; therefore, train run 3–4 needs a little more time than the slot envisaged but stays within the limit of the slot.

6 Level 3 Timetable Preconditions

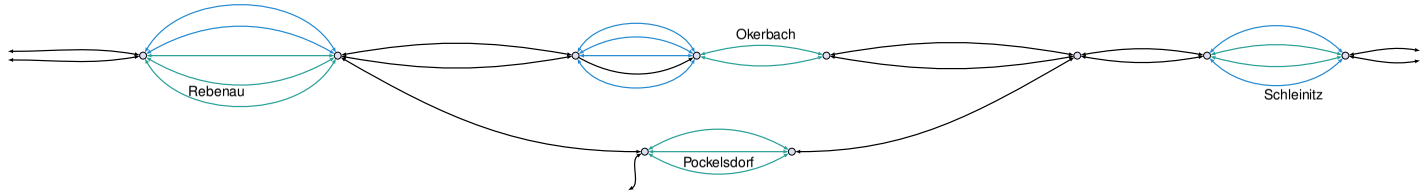
Trains will have to stay as punctual as possible to concatenate slots from line sections and pulses from junctions. Compensation zones could be included in the line sections to add further margins and ensure punctuality. Compensation zones add flexibility of the speed profile in line sections to ensure punctuality at condensation zones like junctions (Caimi et al.



(a) Infrastructure example with main and side tracks.

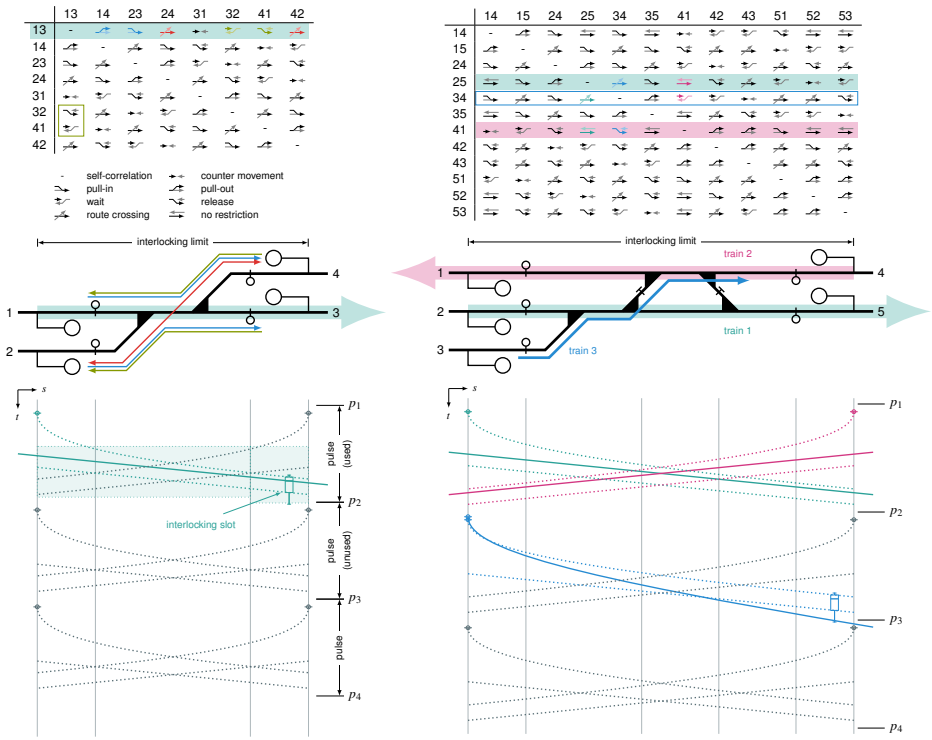


(b) Fragmentation into junctions and line sections.



(c) Timetable network equivalent.

Figure 9: Example for fragmentation of lines with/without platforms, junction with turnouts.



(a) Common interlocking pulses.

(b) Concurrent interlocking pulses.

Figure 10: Pulses at interlocking limits with slots.

2009). Compensation zones will work under the assumption that a train is fitting a slot but still needs a greater degree of freedom for driving.

If a train does not match the properties of a slot, an operator will require means to ensure the quality of a timetable. These means could be to reallocate the train to a different slot type or to demand a change of the train properties, for instance, an additional power unit or a split of the train. Track facilities are required to change the train properties. Facilities for railway operation to react to the current operational status is one of the preconditions for a functioning Level 3 timetable. Other preconditions are:

1. not all slots can be booked to have spare capacity available following UIC code 406; therefore, capacity management for selling rights of usage is required;

2. availability of train berths in front of condensation zones or other bottlenecks (infrastructure for queuing at erratic network parts);

3. sidings or other facilities as a waiting area in order to use the next suitable or the booked slot;

4. alternative lines with spare capacity as a bypass;

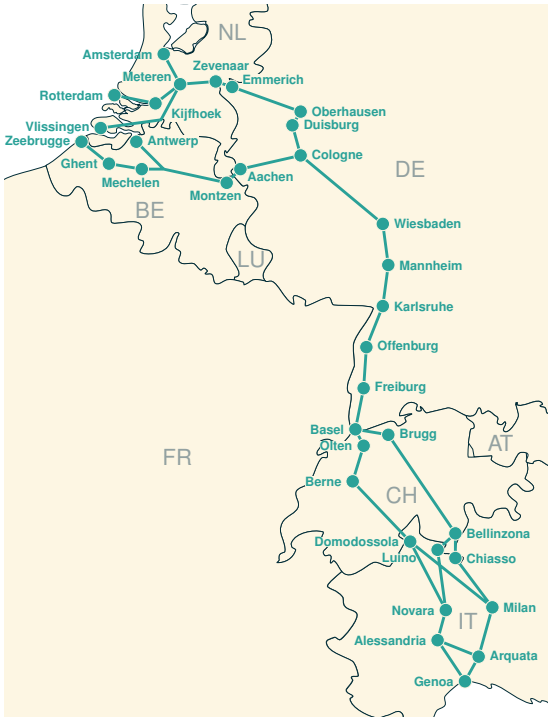


Figure 11: Rail freight corridor Rhine-Alpine.

5. a feedback system for overcrowded lines to modify the future timetable or induce infrastructure enhancements;

6. information management regarding reservation information for actual usage of slots and a dispatching handling system of allocated slots.

Slots, pulses, dispatching gaps and traffic schemes can be integrated into the timetable life cycle currently endorsed by RNE (2014) with these preconditions.

7 Example Corridor Rhine-Alpine

Rail freight corridor Rhine-Alpine (RFC-RA) was created to achieve the goals of the EU 2010. The RFC-RA represents the idea of Level 2 timetable and can hold as an example for the timetable categories from Level 1 to Level 3. The RFC-RA covers railway operation from the Netherlands, Belgium, Germany, Switzerland and Italy (see Figure 11). Their primary purpose is to provide a one-stop shop for path requests for freight trains crossing at least one border along the corridor. They use a path coordination system to manage and allocate path requests for pre-arranged paths and reserve capacity (RFC-RA 2018a).

The path requests for pre-arranged paths have to be submitted at least two months before the start of the timetable period to follow the timetable's life cycle (see Figure 4). After two months, the reserve capacity will be used for the path requests. The objective of the RFC-RA

could be achieved with a Level 1 timetable, but this would mean higher organisational expenses for those railway undertakings using the corridor to book their paths ahead of time. Slots, on the other hand, would mean no difference between path requests for pre-arranged paths and path requests for reserve capacity. The fact that there is a difference can be illustrated by network conditions: if a freight train wants to leave Rotterdam in the direction of the German border, it will get a suitable path for the Betuweroute relatively quickly since it is the only freight route in a homogenous traffic mix. If the same freight train wants to leave Mannheim in any direction, it will matter whether the train has a path or will use reserve capacity because the train will have to share the frequently used network with competing trains in a diverse traffic mix.

The RFC-RA uses data collection from railway undertakings in the last timetable period to form a train archetype. The data collection is done via a spreadsheet called "Expression of Needs" and takes parameters from reference-rolling stock (see RFC-RA 2018b). With the reference-rolling stock, the RFC-RA constructs or allows the construction of parts of pre-arranged paths along the corridor. Data management for trains from the past and procedures for forming cohorts need to be enhanced to further develop the formation of train archetypes for automation and building slots for the corridor.

As shown in section 2, the slots or pre-arranged paths have an impact on railway operation. Therefore, feedback is needed on which path is booked and which is freely available for dispatching. Different infrastructure managers do the railway operation of the RFC-RA in each country, and the booking of the paths are made via the Path Coordination System of RailNetEurope. If the path request is issued in a timely matter (see Figure 4), the railway operation for these paths can be based on the actual usage. However, there are many parties and management systems in various stages involved and, therefore, the consistency and availability of timetable data for the operator can be further improved to gain benefits in dispatching.

Alternative lines are essential for dispatching and slot in a Level 3 timetable. Consequently, the construction or expansion for a second line between Cologne and Basel was created, as the need for alternative lines was recognised for the RFC-RA (see Figure 11). The importance was even further stressed at the collapse of a tunnel during construction in Rastatt (between Karlsruhe and Offenburg) in August 2017. The collapse disrupted the corridor and alternative lines where not suitably equipped to absorb the resulting detour traffic.

8 Conclusion and Further Research

In summary, properties and concepts of timetables can be arranged in levels. This paper proposes timetable levels from 0 to 3, where Level 3 is an advancement to integrate the ad-hoc paths for short-term freight trains. The integration is devised by introducing a dispatching gap between long-term, fixed-interval paths. The dispatching gap needs to be managed and blended into the railway operation to enable the benefit of flexibility for freight trains. For junctions and line sections, different approaches must be used to utilise the dispatching gap. Slots can be used for the line sections and pulses can be used for junctions. Railway operational restrictions at junctions can be stated as a restriction matrix.

Encapsulating infrastructure restraints for railway operation, together with a timetable level with flexible slots, can help to formulate input data to determine infrastructure. Moreover, it will provide a further step for determining infrastructure from a given timetable. This paper also provides further analysis for the transit and transport layer of Scheidt (2018).

Further research is required since slots and pulses for the dispatching gap are only a concept. There are several opportunities to enhance the concept further:

1. Train archetypes form the basis of slots and are crucial for the usability of real trains. Therefore, the estimation of quantiles for each train archetype and line section needs to be further investigated.

2. Dispatching rules for the preclusion of slots and the handling of mismatching trains to their slots should be examined.

3. The impact of the capacity usage of slots compared to the UIC code 406 or other methods can be further reviewed.

It could prove difficult to verify the improvement to a Level 3 timetable for freight traffic experimentally with real trains. Two comparable line sections with railway operators willing to test the Level 2 timetable against a Level 3 timetable would be required. However, current railway operation management systems work in a critical environment where safety is the priority and the time for the development of features for these systems is elongated, which leaves only railway laboratories at universities for the validation of the model concept. Real train data for the train archetypes will be the limiting factor for the transfer of the results from railway laboratories to actual railway operation apart from the artificial interaction with a railway operation management system in a laboratory.

In the future, a level 3 timetable could simplify ad hoc scheduling. A train could immediately be given a conflict-free route based on the real time situation, in which all involved dispatchers and interlockings would know this train. Possibly across infrastructure manager borders, too; a prerequisite is a networked, digital train path management system, which combines operational planning and control (see Figure 2). This is easier said than done: interfaces between components within and between operators as well as suitable algorithms for processing are required.

References

Caimi, G., M. Fuchsberger, D. Burkolter, T. Herrman, R. Wüst, and S. Roos (2009). "Conflict-free train scheduling in a compensation zone exploiting the speed profile". In: *International Seminar on Railway Operations Research*.

UIC – International Union of Railways (2000). *CODE 451-1: Recovery margins*. ISBN: 2-7461-0223-4.

— (2004). *CODE 406: Capacity*. ISBN: 2-7461-0802-X.

EU – European Parliament and Council (2010). *Regulation EU/913/2010*.

Hertel, G. and J. Steckel (1992). "Eine neue Philosophie der Fahrzeitberechnung für Zugfahrten". In: *Wissenschaftliche Zeitschrift Hochschule für Verkehrswesen Friedrich List*, pp. 104–111.

Medeossi, G., G. Longo, and S. de Fabris (2011). "A method for using stochastic blocking times to improve timetable planning". In: *Journal of Rail Transport Planning & Management* 1.1, pp. 1–13.

Pachl, J. (2018). *Railway Operation Control*. 4th ed. Mountlake Terrace (USA): VTD Rail Publishing. ISBN: 978-1-7327310-0-4.

Pöhle, D. (2016). "Strategische Planung und Optimierung der Kapazität in Eisenbahnnetzen unter Nutzung von automatischer Taktfahrplanung". PhD thesis. TU Dresden.

RFC-RA – Rail Freight Corridor Rhine-Alpine (2018a). *PaP Catalogue TT2019*. URL: https://www.corridor-rhine-alpine.eu/downloads.html?file=files/downloads/coss/20180206_Catalogue_RFC1_TT2019.pdf (visited on 2019-01-21).

— (2018b). *TT2020 Corridor Capacity Wish List template Tutorial*. URL: https://www.corridor-rhine-alpine.eu/downloads.html?file=files/downloads/coss/TT2020%20Corridor%20Capacity%20Wish%20List%20template%20tutorial.xlsx (visited on 2019-01-21).

RNE – RailNetEurope (2014). *Guidelines for Pre-arranged Paths*. Vienna. URL: http://www.rne.eu/rneinhalt/uploads/Guidelines-PaP-V3.01.pdf.

Roos, S. (2006). "Bewertung von Knotenmanagement-Methoden für Eisenbahnen". MA thesis. IVT, ETH Zürich.

Scheidt, M. (2018). "Proposal for a Railway Layer Model". In: *COMPRAIL 2018*. Southampton, UK: WIT Press, pp. 157–168. DOI: 10.2495/cr180141.

Scheidt, M., L. Pelster, G. Bosse, and J. Pachl (2018). "Process Map Railway Operation". In: *Universitätsbibliothek Braunschweig*. DOI: 10.24355/dbbs.084-201901211619-0.

Schittenhelm, B. and T. Richter (2009). "Railway Timetabling Based on Systematic Follow-up on Realized Railway Operations". In: *Annual Danish Transport Conference*. Aalborg University. Aalborg, pp. 1–34.

Weigand, W. and A. Heppe (2013). "Spurplangestaltung und betriebliche Infrastrukturplanung". In: *Handbuch Eisenbahninfrastruktur*, pp. 441–494.

Simulation of metro operations on the extended Blue line in Stockholm

Hans Sipilä [a, 1], Anders Lindfeldt [a, 2]
[a] Sweco
[1] E-mail: hans.sipila@sweco.se
[2] E-mail: anders.lindfeldt@sweco.no

Abstract
The current Blue metro line in Stockholm will be extended and connected to a branch on today's Green line in the future. This paper presents a timetable simulation study which was part of a larger study regarding traffic analyses and design of the future Blue and Yellow line metro conducted in 2015–2016. Two timetable cases were considered, the first one gives 4-minute intervals on the branch lines and the second one 5-minute intervals (peak hours). At that time there was a discussion whether to design a new branch station (Sofia) with two or three tracks. The simulation model could not be set up to model all the features of the signaling system that is used today and is planned to be used when the new extensions open. Therefore, a separate model was developed to study the effects of two or three tracks at the branch station on a more detailed level.

The results from this study shows that the 4-minute timetable case is clearly more sensitive to delays. Although the effects of having two or three tracks on the branch station where northbound trains will merge can be seen locally on that and subsequent stations, there is no significant difference further along the line. There exist other operational benefits of having a 3-track design at a branch station and these were also considered, although not discussed in this paper. Later it was decided that the branch station will be designed with two tracks, mainly due to the significantly higher cost for a 3-track design.

Keywords
Metro operation, Timetable, Simulation, Delay, Buffer time

1 Introduction

Stockholm's Metro is about to be expanded. The current Blue line will be expanded with nine stations, where one replaces an existing surface station and one expands an existing surface station with an underground section. One of the southern branches on today's Green line will be connected to the Blue line. The construction start is planned to 2019.

Earlier in the project a study of the train operations on the expanded Blue line was carried through. This study consisted, among other things, of timetable analyses, trip scheduling from depots and simulations.

A new branch station (Sofia), where the line divides into two branches, is planned a few kilometers south of the current terminal station at Kungsträdgården. The design of this branch station was not decided at the time of this study. In short, this station could consist of either one common platform track or separate platform tracks for northbound inbound trains from the two branches. For outbound trains one platform track was considered in both cases.

The station will lie deep under the street level and the platforms will only be accessible by lifts, no escalators are planned.

This paper aims to describe the simulation setup and presents some results from the simulations. Two timetables are considered. The first one has 5-minute intervals on each branch during peak, giving 2.5-minute intervals on the common section. The other alternative has 4-minute intervals on each branch and 2-minute intervals on the common section. Full weekday timetables are designed with empty train runs from and to depots. However, in this paper presented results correspond to morning and afternoon peak periods only.

In addition, a separate model was developed to be able to model the signaling system behavior in a more realistic way and this model was used to study the difference in inbound train flow for the two considered designs at Sofia station. The results from this analysis are presented as well.

Figure 1 shows a schematic track layout for the new Blue line metro. The new sections reflect the track layout as planned in the fall of 2015 when this study was conducted. Figure 1 shows branch station Sofia with a 3-track design. The existing branch station in Västra skogen (VÄS), the future one in Sofia as well as the connections tracks to the two depots (RI and HÖ) have flyovers (grade separation).

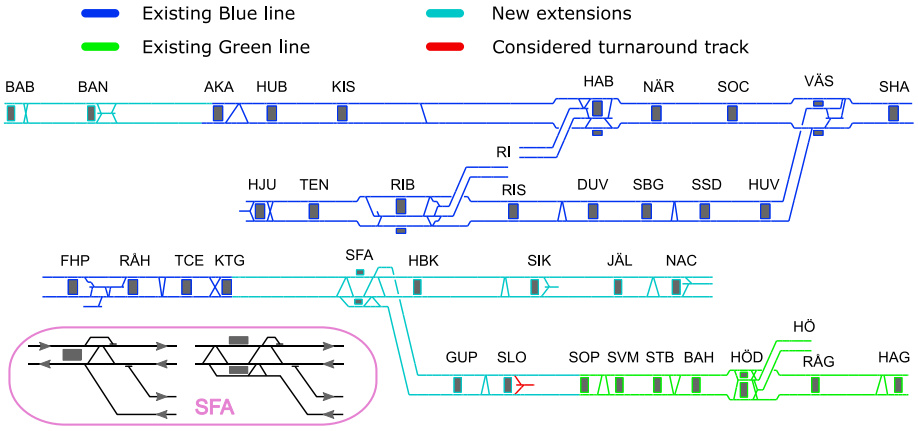


Figure 1: Schematic track layout with station codes for the new Blue line metro as planned in the fall of 2015. The figure shows both station designs for Sofia.

2 Signaling system

There are currently two different signaling systems used in the Stockholm metro. The Red and Blue lines use a relay-based system from Union Switch & Signal, this system was also used on the Green line until replaced with a more modern system by Siemens 20 years ago. The relay-controlled system is planned for installation on the new track sections to avoid different systems when these are opened. However, the section today belonging to the Green line that will be connected to the Blue line will continue to use the Siemens system. Both systems can be handled by the C20 train stock. This is also the plan for the new C30 units delivered in the coming years.

Both systems send information continuously to the train's safety system. In the older system (Union Switch & Signal), the signal is picked up from the rail tracks (Stockholm Public Transport Authority (1964)). Three speed aspects can be transmitted:

- L – Low: 15 km/h
- M – Medium: 50 km/h
- H – High: Line maximum speed, 80 km/h is used on the Red and Blue lines

Speed restriction signs can imply lower speeds than the cab signal speed aspect. If a train exceeds the transmitted speed at any point along the line, the system automatically applies the brakes until the allowed speed is reached. The signal alters successively to Medium and Low when a train approaches another train in front or some other obstruction along the line, e.g. a wayside signal at stop in front of points.

The system allows two trains to come close to each other, the subsequent train has in this case received speed aspect Low at a sufficient distance and can normally continue up to the rear light on the preceding train. If no speed aspect is transmitted the cab signal system interprets this as Low.

The resolution of the system depends on the length of the track circuits, these have been calculated from the braking power of the cars and the interval desired between trains. On straight level track, each track circuit is around 200 meters on central parts of the line. The signal system is designed to give a headway of approximately 90 seconds and 30 second train stops at stations.

In the RailSys-model an ATC system with continuous updating is used since there is no straightforward way to model all the properties of the system used in the Stockholm metro. The block section lengths correspond approximately to the track circuit lengths. This setup is used in the simulations of the Blue line timetables. In the analysis of the effects on train flow if the two branch lines merge before or after Sofia station (i.e. 2-track or 3-track station design), a separate model is developed with the purpose of modeling the signal system properties as closely as possible.

3 Timetable case simulations

The track infrastructure setup for the existing Blue line together with the extensions is created in the simulation software RailSys (see e.g. Bendfeldt et al. (2000)). Management and project data are used. This includes vertical and speed profiles and an approximation of the speed-code signaling system in use today, also planned for the extensions. The vehicle model used is C20, which currently and for years to come most likely represents the largest share in the fleet. Data from the vehicle manufacturer is used for traction diagram, acceleration and braking performance etc. RailSys version 8 is used in this study.

A simplification is that trips are not defined by connecting arriving to departing trains for the whole timetable at terminals. Trains that are late to terminal will not transfer the possible delay for the reversing departure directly (for the same trip). However, crossing conflicts can occur as late arriving and/or departing trains must pass the point area. In metro operations in Stockholm trains can be reversed before the terminal at stations equipped with turnaround tracks. This is done to prevent a delay being passed on in the reverse direction.

Train runs cannot be canceled (in whole or in part) in a RailSys simulation. The advantage of connecting train runs is that delays can be passed on at terminal stations which is realistic. However, the disadvantage is that for larger delays the delay transfer can be unrealistic since a trip could have been reversed before the terminal to counteract the delay

transfer. To model a more conservative behavior at terminals, all inbound and outbound trains must change tracks (cross) through the point area. The departing trains are given initial delays from distributions as well.

3.1 Input data for timetable setup and simulation

To verify the infrastructure and the vehicle model, the running times calculated by the model are compared to running time distributions from recorded train data provided by the metro operator. This is done for the existing sections. The median values are normally chosen as the representative values in the timetable setup. For the new sections, scheduled running times are chosen so that the difference between the minimum technical and the scheduled running times are in the same magnitude as the differences between median and 90-percentile running time values for sections of similar lengths in the recorded data.

The variation in dwell times is modeled by using corresponding recorded data provided by the metro operator as distributions and separated to stations, direction and operational period (e.g. morning, morning peak, mid-day etc.). Similarly, distributions for departure deviations at different stations are provided. These are used in modeling initial delays for trains departing from terminal stations. Dwell and departure deviation data is loaded into a Matlab-database from which further handling in assigning perturbation values and writing the perturbation files for the simulations is done. In this process future stations are mapped to existing stations following rough estimates of expected passenger volumes.

3.2 Approach for emulating the bunching effect

Bunching is an effect that is common in congested light rail, metro and commuter train networks. A train (vehicle) that is already delayed will get more people at the next station and the passenger exchange take longer time. This means that the headway to the train in front increases. The following train will therefore accumulate delays. A method for how this effect can be modeled in the OpenTrack software is presented in Krause (2014) in which this method is used in an analysis of the red metro line in Stockholm. The method uses the OpenTrack API to accomplish this.

In a RailSys simulation there is no possibility to dynamically control dwell times once a simulation is running. In this study, the possibility to emulate this effect is instead defined in the setup scripts run in Matlab. The probability of a train getting an extended dwell, the number of consecutive stations the extended dwell is active on, for which stations in the network this can happen and in what time periods are controlled in the setup. The dwell extension value can be a constant or drawn from a distribution. In this way some trains in a simulation cycle will get extended dwells systematically for several consecutive stations, delay is accumulated and the headway to the preceding train increases. This behavior can of course occasionally occur in a simulation cycle without this additional modeling. This can happen if several consecutive higher dwell values are drawn from the normal dwell variation distributions for a train when the values are assigned prior to a simulation. However, the described approach provides a possibility to control this: frequency, levels of the extended passenger exchange times, when, where etc.

3.3 Results from timetable case simulations

The timetable case simulations are evaluated by comparing the average delays (mean values) and standard deviations. In addition, the potential need for short turnarounds is estimated at different stations where some are hypothetical since there are no separate turnaround tracks planned for these and making train turnarounds on main tracks is not realistic in peak periods.

A train run is marked for turnaround at the considered stations if the sum of arrival delay, remaining scheduled time to terminal and a minimum turnaround time exceeds the scheduled departure from terminal with more than one minute. This limit may sound small but with trains running every 2–2.5 minute on the common section the need for precision is high to avoid passing on delays to trains coming from and going to another branch. The minimum turnaround time in these cases is assumed to be 3.5 minutes. Measurements at Kungsträdgården terminal (KTG), where trains turn around every three minutes (in peak), indicated an almost 95 % fulfillment up to this time.

There are ways for shortening this time a little bit by using a second driver that will help activate a train in the reverse direction or by changing drivers at a turnaround, which would lower the time needed for a turnaround even more. This was not considered in the estimations.

Figure 2 shows simulated mean delays in peak periods for the different lines in both directions. Both timetable cases are combined with the two different track designs at Sofia station. There is no difference coming from the station design in the southbound direction which is reasonable since the different designs only affect the northbound trains. There can of course be an indirect effect through crossing conflicts at terminals, but there is no indication of that in these results.

Although the timetable cases have similar scheduled running times there are differences. This is mainly due to that turnaround times and the time differences between inbound and outbound trains differ. Running time allowances are in some parts of the line distributed differently. In general, the recoverability is better in the southbound direction. The mean delays decrease clearly on the common section between Västra skogen (VÄS) and Sofia (SFA). The standard deviation is not shown in any diagram, but it is also lower in this direction compared to the northbound direction. The standard deviation is higher for the relation Hjulsta–Hagsätra and vice versa compared to the other line. The weakest parts of the system seem to be in the northbound direction from stations Sundbybergs centrum (SBG) and Hallonbergen (HAB).

At some stations it can be observed that the mean delay increases during the station stop. The most likely reason is that the passenger exchange time on average take longer time than the scheduled dwell times used.

The difference between the two station designs at Sofia can be seen in the diagrams. The effect is relatively local since the mean delays coincide further ahead. The standard deviation increases locally with 5–10 seconds as well.

Assuming a minimum turnaround time of 3.5 minutes and relating this to the scheduled turnaround times in the 4-minute timetable case, gives that the additional margin is 1:45 in Hjulsta (HJU), 3:00 in Barkarby station (BAB), 2:00 in Hagsätra (HAG) and 2:00 in Nacka centrum (NAC). Checking the simulated arrival delay distributions at these terminals indicates that around 10 % of the arriving trains to Hjulsta will carry on a delay in the reverse direction. Corresponding values for the other terminals are 2–3 %. In the 5-minute case, the scheduled turnaround times are longer and considered to have enough margins based on these results.

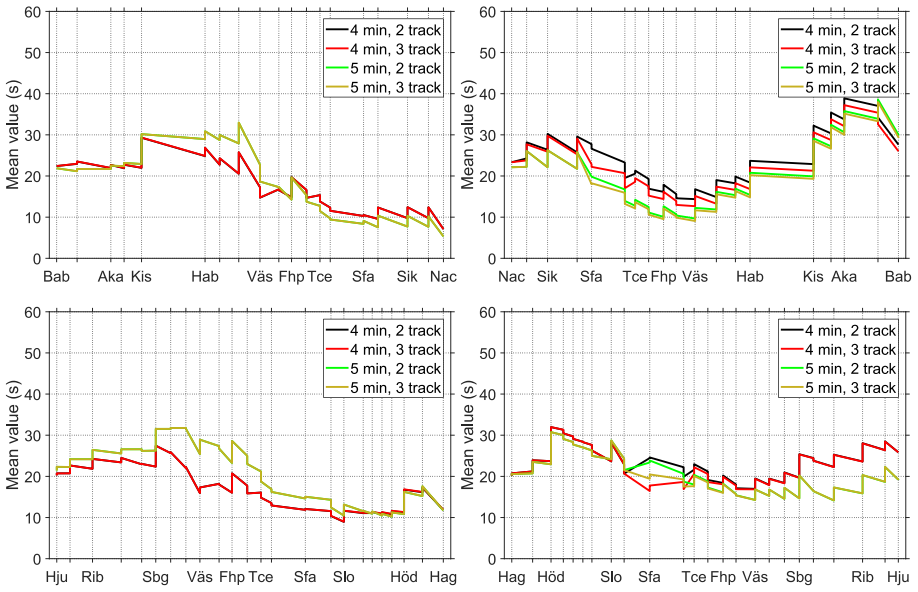


Figure 2: Mean delays in peak periods for the two timetable cases and with different station designs for Sofia station. Trains running from Barkarby station (BAB) to Nacka centrum (NAC) and vice versa on top. Trains running from Hjulsta (HJU) to Hagsätra (HAG) and vice versa on bottom.

Figure 3 shows the estimated number of short turnarounds per day in peak periods at different stations. The relation Hagsätra–Hjulsta and vice versa has in total a higher number of short turnarounds than the other line. It is also clear that the 4-minute timetable case is more sensitive to delays. Akalla (AKA), Rinkeby (RIB) and Högdalen (HÖD) are all situated relatively close to the respective terminals and it makes sense that these would get higher numbers than stations further from the terminals.

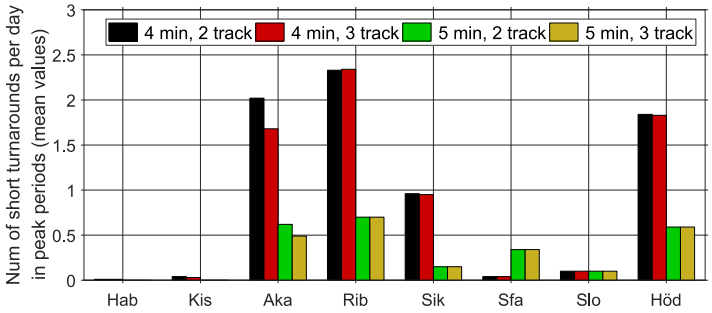


Figure 3: Estimation for short turnarounds per day in peak periods at specific stations. The first four bar groups represent northbound trains (Hab, Kis, Aka and Rib), the remaining four groups represent southbound trains (Sik, Sfa, Slo and Höd).

4 Detailed analysis of train movements through Sofia station

4.1 Method

A new model is developed to determine if it is necessary to have two platform tracks for northbound trains at Sofia station. Northbound traffic from the two different lines merge at Sofia. If the station is built with one platform track for northbound trains, traffic from the two lines must synchronize before arriving at the platform. With two platform tracks, synchronization will occur after the trains depart from the station (see Figure 1).

The reason why the analysis of Sofia station requires a specialized model is that the gradients around Sofia are steep (> 45 ‰, falling for northbound trains) and that the signaling system's characteristics and behavior cannot be fully modeled in RailSys. The steep gradients affect both the braking performance of the trains and the setup of the signaling system. Steep gradients in combination with dense traffic from different branch lines synchronizing at Sofia can mean that the station will become a bottleneck. With two platform tracks for northbound trains, the impact of the steep gradients south of the station will be reduced. However, designing a larger station with one additional platform track implies a more complex construction and a higher cost.

The analysis includes only northbound trains and the infrastructure model includes one stop before Sofia on each branch line (Hammarby kanal/HBK and Gullmarsplan/GUP) and the first stop after Sofia (Kungsträdgården/KTG). The developed model consists of several sub models: infrastructure model, signaling system, driver model and vehicle model. Most of the effort is put into modeling the signaling system's characteristics as accurately as possible.

Modeling the signaling system

As described before, the speed aspect received by a train is transmitted through the track circuits. When a track circuit is occupied by a train, the track circuits behind the occupied track circuit will transmit either one of the three signal aspects (H, M or L). Which speed aspect the track circuits will transmit depends on the location of the track joints and the braking distance calculated from the rear joint of the occupied track circuit. There will always be at least one track circuit transmitting L and one transmitting M. More than one track circuit may be required to transmit the same speed aspects, this depends mostly on the track circuit lengths in combination with gradients. For each track circuit, braking distances are calculated for two different speed aspects, H and M. If a track circuit is occupied, all track circuits behind the occupied track circuit that are located within the braking distance calculated for M will transmit the L-aspect. If the track circuit is within the braking distance of H, it will transmit the M-aspect. Track circuits that are further away than the calculated braking distances, will transmit the H-aspect. Information about which track circuit will transmit which speed aspect is saved and used in the simulation.

The positions and lengths of the track circuits form input to the model. The configuration of the track circuits affects the potential capacity of the system. Hence, the configuration needs to be optimized for each scenario to make results comparable. The adjustments of the track circuit configuration are done manually in several iterations. Each iteration includes recalculation of braking distances and speed aspects as mentioned above. The primary focus for the adjustments is to maximize buffer times at bottlenecks. The bottlenecks of the system are track circuits that are affected when the trains stop for passenger exchange and by the turnout at Sofia station. Several technical limitations, such as for example minimum and maximum permissible length of track circuits, must also be considered in the process.

Braking curves are calculated by means of numerical integration and the effect of varying gradients along the track is considered. The metro train is modeled as a mass band with weight distributed equally along the entire train length. The design guidelines for the signaling system dictates that the calculation of the braking distance shall include a 5.5 second brake reaction time and a 15 % safety margin of the total braking distance.

Deterministic and stochastic simulation

The deterministic simulation aims to analyze the planned situation. The trains in the model run as fast as possible from entry to exit. The trains accelerate, keep constant speed, brakes and stop at stations. No stochastic delays are used in this mode. Trains start with a fixed time interval and station stops (dwell) are performed according to plan. After the simulation, the result is a timetable and stored data about transmitted speed aspects from all track circuits. The information is then used for creating blocking time diagrams and calculate buffer times between trains. The timetable is conflict-free, and the trains will not be affected by other trains (no restrictive speed aspects due to trains in front).

The stochastic simulation is largely a repetition of the procedure in the deterministic simulation. The difference is that the trains are disturbed from their planned timetables by means of stochastic delays (see e.g. Siefer (2008)). The delay distributions determine how often and how much the trains are delayed in different situations. Trains are disturbed when they enter the model by initial/entry delays and when they stop at stations by dwell time extensions. Since the trains are delayed, they will not always run in their planned conflict-free slots. When a conflict occurs with another train, the signaling system will force the train behind to reduce speed according to the speed-code signaling system. The results are delays that are measured relative to the times calculated in the deterministic simulation.

4.2 Results

Buffer times

The distance between trains, buffer time, affects the probability that trains will affect each other in the case of delays. The available buffer time depends on the frequency of the traffic, the minimum headway and the stop times at stations. Minimum headway depends, among other things, on the signaling system (track circuit configuration, reaction times etc.) and the speed and length of the train. Blocking time diagrams (Figure 4) are produced by the deterministic simulation and are used to calculate buffer times. Figure 4 shows the Nacka line. When the turnout at Sofia is not in position for a train coming from the Nacka line (i.e. the optical signal protecting the turnout shows red), the track circuits before the turnout indicate L and M speed aspects. This is the situation when the turnout is either changing from one position to another or when it is in position for trains coming from the Hagsätra line. This is also the reason why some track circuits before Sofia station transmit restrictive speed aspects for long periods of time.

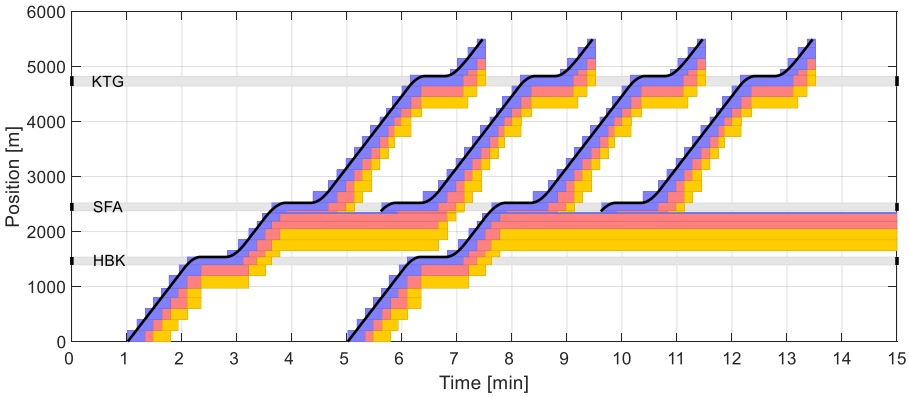


Figure 4: Blocking time diagram with a 2-track station design at Sofia, negative gradients are considered. Red: L-aspect, Yellow: M-aspect, Blue, occupied track-circuit. Station positions for Hammarby kanal (HBK), Sofia (SFA) and Kungsträdgården (KTG) are indicated in the figure.

Figure 5 shows how the buffer time varies along the Nacka line. The figure shows buffer times for each track circuit. Two different buffer times are calculated. The yellow line shows the buffer time to M-aspect and the red line to L-aspect. The reason for why the buffer time to L-aspect is calculated is that the impact on train delays is much higher if a train receives L-aspect (15 km/h) than M-aspect (50 km/h), especially close to stops where trains are not running at full speed.

Figure 5 shows three scenarios. In the first scenario (dashed lines) Sofia station has two tracks in total (one track for northbound trains). The second scenario (dash-dot lines) is the same as the first, but without negative gradients. In the third scenario (solid line) Sofia station has three tracks in total (two tracks for northbound trains). Buffer times to M-aspect are shown in yellow and buffer times to L-aspect in red.

In all scenarios, buffer times clearly decrease between Hammarby kanal and Sofia. The reason is that the traffic frequency doubles at Sofia when trains from the Hagsätra line merge with trains from the Nacka line. In Figure 5, both lines operate trains in 4-minute intervals. Hence, the interval between trains is two minutes from Sofia to Kungsträdgården. The figure shows that buffer times are generally shorter at stations than on the lines. This is due to that trains are standing still at the stations for some time and that they have lower speeds when they decelerate before and accelerate after stops. It is evident from the figure that stops do not only affect buffer times on the platform track circuit(s) but also several track circuits before.

Comparing the dashed end dashed-dotted line reveals the effect of the negative gradients in the scenario where both lines share one platform track at Sofia. Differences are greatest around Sofia station where the negative gradients are located. Without negative gradients, buffer times around Sofia increase from 22 seconds to 27 seconds (M-aspect). For L-aspect, buffer times increase from 38 to 46 seconds. The distance with short buffer times is also about 200 meters longer when negative gradients are considered (M-aspect). It is also worth noticing that the point where the signaling system starts to transmit the M-aspect when a train is at the platform in Sofia, is only about 100 meters after the stop at Hammarby kanal (the preceding station on the Nacka line). If the braking distance had been 100 meters

longer, it would not have been possible to operate trains at 4-minute intervals without conflicts at Hammarby kanal. Operating with conflicts means in this case that trains would get restrictive speed aspects due to other train movements also in a scheduled mode. In practice however, the consequences of such a situation would probably be limited since trains stopping at Hammarby kanal will have lower speeds when entering the platform section.

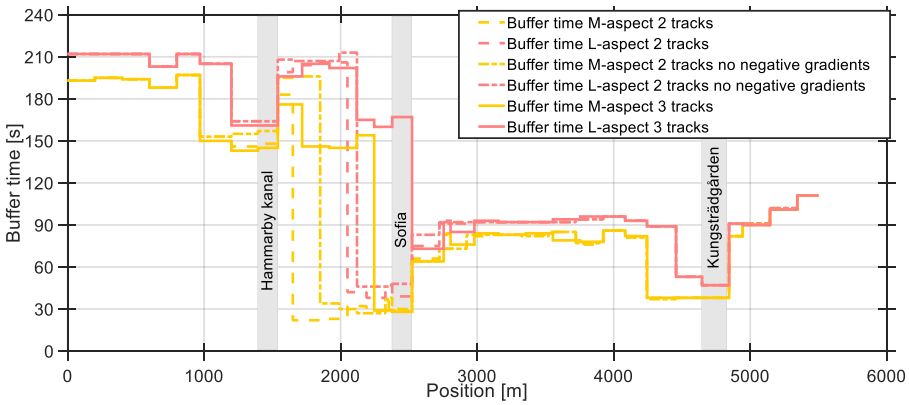


Figure 5: Buffer times for the Nacka line. Station positions for Hammarby kanal, Sofia and Kungsträdgården are indicated in the figure.

In the discussed scenarios, the turnout is located right before the platform in Sofia and both lines share the same platform track. In the third scenario, each line has a separate platform track and the turnout where the tracks merge is located 230 meters after the platform area. Buffer times for scenario 3 are indicated with solid lines in Figure 5. Compared to the scenario with two platform tracks, the buffer time at Sofia increase from 22 to 29 seconds for M-aspect and the distance with short buffer times is reduced with almost 600 meters. The distance between the platform and the turnout is long enough to avoid L-aspect at the platform when the turnout is in the other position (for trains from the Hagsätra line). Hence, the buffer time for L-aspect at the platform is as much as 167 seconds. However, the buffer time between the platform and the turnout is affected by the position of the turnout and it is therefore not independent of the traffic on the other line. It is significantly lower, 73 seconds.

Comparing the buffer times at Sofia with those of the next station, Kungsträdgården, shows that Sofia with a 2-track design (one platform track for northbound trains) have shorter buffer times than Kungsträdgården and might therefore become a bottleneck. Sofia with a 3-track design (separate platform tracks for northbound trains) have longer buffer time to L-aspect but smaller buffer time to M-aspect when compared with Kungsträdgården. How the different scenarios will perform when trains are delayed is not easy to predict based only on the buffer times. For that reason, simulations with delays are performed in the following step.

Delays

Stochastic simulations are performed to determine how the steep gradients and number of platform tracks at Sofia affect the delay sensitivity of the system. Stochastic delay distributions are used to model primary delays and the simulation model is used to determine how the trains will affect each other, secondary (knock-on) delays. Values are drawn stochastically from distributions modeling initial/entry delays and dwell time extensions. The same distributions are used on both lines which are evaluated together. The distributions are compiled from recorded data reflecting years 2011–2015. Distributions used for initial delays have, in this study, relatively high average values and are chosen with the intention to stress the system. Figure 6 shows how the average delay increase from the position where trains are initialized in the model, before Hammarby kanal and Gullmarsplan, until and including departure from Kungsträdgården station.

The average dwell times in the simulation are chosen so that they coincide with the scheduled dwell times. This means that trains, on average, cannot reduce their delays during the station stops. It also means that the delay increase observed in Figure 6 is due to secondary delays only. The figure shows results for the 4- and 5-minute timetable cases. In most cases, the results show a higher increase in delays up to Sofia station, whereas the increase is smaller between Sofia and Kungsträdgården. The difference in delay increase between the 4- and 5-minute cases is significant. In the 5-minute case, delay increases by 3–4 seconds, whereas in the 4-minute case it increases by 9–11 seconds. The impact from the steep down grade (negative gradient) is marginal. In the 4-minute case, the difference when comparing a configuration with and without gradients is around 2 seconds.

Figure 6 shows how the number of northbound platform tracks at Sofia affects the train's average delay. In the scenarios where a 2-track design is used (one northbound platform track) a higher increase is observed in secondary delays up to and including departure from Sofia and less increase thereafter. In the 4-minute case, trains get on average a 10 second delay increase up to and including departure from Sofia and from there up to and including departure from Kungsträdgården a marginal increase. The marginal delay increase on the last section is explained by that trains from the two branch lines have already been synchronized at Sofia which has shorter buffer times than Kungsträdgården. In the simulated scenarios where a 3-track design is assumed (two northbound platform tracks), the increase in secondary delays move from Sofia to Kungsträdgården. The reason is that when Sofia gets larger buffer times, the bottleneck and part of the synchronization effect moves from Sofia to Kungsträdgården.

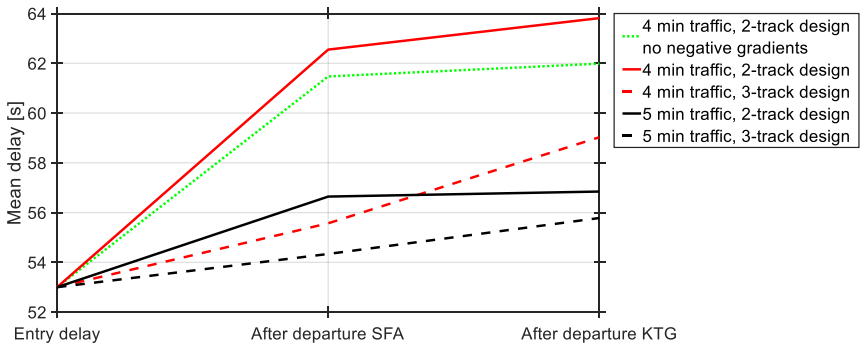


Figure 6: Simulation results (mean delays) for the simulated cases.

The scenarios where Sofia has a 3-track design generate a smaller level of secondary delays in total than in the scenarios where Sofia has a 2-track design. The effect is clearer in the 4-minute timetable. This is expected since the buffer times are smaller. With a 3-track design at Sofia, the delay after Kungsträdgården increases by 3–4 seconds when the train frequency is increased from 5- to 4-minute intervals. The corresponding increase with a 2-track design is 7 seconds.

5 Conclusions

The results from both the timetable case simulations and the detailed analysis shows that the 4-minute timetable case is clearly more sensitive to delays. Although the effects of having a 2 or 3 track design at Sofia where northbound trains will merge can be seen locally on that and subsequent stations, there is no significant difference further along the line. There exist other operational benefits of having a 3-track design at a branch station and these were also considered in other studies, although not discussed in this paper. Later it was decided that the branch station will have a 2-track design, mainly due to a more complex construction and a significantly higher cost for a 3-track station design.

References

Bendfeldt, J.P., Mohr, U., Müller, L., 2000. "Railsys, a system to plan future railway Needs", In: Brebbia, C., Allan, J., Hill, R., Sciutto, G., Sone, S. (Eds.), *Computers in Railways VII: Proceedings of CompRail 2000*, WIT Press, Bologna, Italy.

Krause, C., 2014. "Simulation of dynamic station dwell time delays on high frequency rail transport systems – Representing dynamic station delays with OpenTrack", Master of science thesis, KTH Royal Institute of Technology, Stockholm, Sweden.

Siefer, T., 2008. "Simulation", In: Hansen, I., Pachl, J. (Eds.), *Railway Timetable & Traffic*. Eurailpress. chapter 9, pp. 155–69.

Stockholm Public Transport Authority, 1964. *Technical description of the Stockholm underground railway.*

Implementation of New Timetable Rules for Increased Robustness – Case Study from the Swedish Southern Mainline

Emma Solinen [ab]
[a] Trafikverket, SE-172 90 Sundbyberg, Sweden
[b] Department of Science and Technology, Linköping University,
SE-601 74 Norrköping, Sweden
E-mail: emma.solinen@trafikverket.se, Phone: +46707247038

Abstract
Due to high demand and high capacity consumption, railway timetables often become sensitive for disturbances and there is little time in the timetables for delay recovery. To maintain a high quality in railway traffic it is important that the timetables are robust and there is a need for strategies and rules for how to make them robust without consuming too much capacity. In this paper we present how timetable rules can be implemented to create more robust timetables. The rules are separated into two categories, rules to make the timetable feasible and rules to increase the delay resistance and recovery. The implementation is illustrated in a real-world case from when the timetable for the Swedish Southern mainline was created for 2019. In the paper we describe how new rules can be applied manually and we discuss advantages and disadvantages by using this approach. We also describe how the rules effect the trains, their timetable slots and runtimes. The results from this study show some of the difficulties when moving from theory to practice and what can be done with limited resources in reality. It gives insights to the practical approach of train timetabling problem which can be used to improve optimization models.

Keywords
Railway timetabling, Robustness, Timetable rules, Implementation, Case study

1 Introduction

The demand for railway capacity has over the years increased a lot. In the densest hours it is not unusual that the demand for train slots is higher than the infrastructure admits. Also, there are often several train operators with different needs, running at the same infrastructure at the same time, that need to be scheduled together. Hence, to solve the puzzle it is tempting to use all possible line capacity for trains and neglect time needed for supplements and margins since it also consumes capacity. The timetable then becomes sensitive for disturbances and there is little time for delay recovery. To maintain a high quality in railway traffic it is important that the timetables are robust and there is a need for strategies and rules for how to make them robust without consuming too much capacity.

In Andersson et al. (2013, 2015) the concept of critical points and the related measure Robustness in Critical Points (RCP) are presented. Andersson et al. (2015) illustrate how RCP can be used in a MILP (Mixed Integer Linear Programing) model to improve the overall robustness in a timetable. However, in Solinen et al. (2017) a comprehensive evaluation of a timetable produced by the MILP model is presented, which illustrates that there are some complications when using the produced timetable in a microscopic

environment. There are several simplifications and assumptions made in the MILP model which makes it hard to use straight away and the evaluation shows not only positive results.

The outline of this paper is that the delay problem, with focus on the Swedish Southern mainline, is described in Section 2. In Section 3 the general timetable rules used in Sweden are presented and also the new approach to increase timetable robustness by increasing RCP. In Section 4 the new timetable rules, used for the Swedish Southern mainline 2019, are introduced and we present how the concept of critical points can be combined with other timetable rules and implemented manually in the timetable construction phase. We also describe how the new rules effect the trains, their timetable slots and runtimes, as well as preliminary punctuality effects. In Section 5 we have a concluding discussing on the results where we discuss the advantages and disadvantages by using this approach.

2 Problem Description

Today there are timetable rules for how much time supplements that have to be added to the trains' runtimes, rules for minimum headway times and for some parts of the line, rules for maximum number of train slots per hour, Trafikverket (2016a). These rules are applied most of the time but there are some deviations. Also the timetable tool used by the timetable planners has some shortcomings. For example, the runtimes are only calculated to the centre of the main track for each station and if a train is planned to run on a side track through low-speed switches, the extra runtime needed is missing in the timetable. Until we have a timetabling tool that can identify conflicts and calculate more accurate runtimes there is a need for timetable rules that ensure the feasibility.

To demonstrate the need for new timetable rules a case study from the Swedish Southern mainline is chosen. This is one of the most congested double track lines in Sweden where fast long-distance trains, regional trains, commuter trains and freight trains are using the same infrastructure. The almost 50 km long line goes from Katrineholm, south of Stockholm, to Malmö and further on to Copenhagen in the south, see Figure 1.

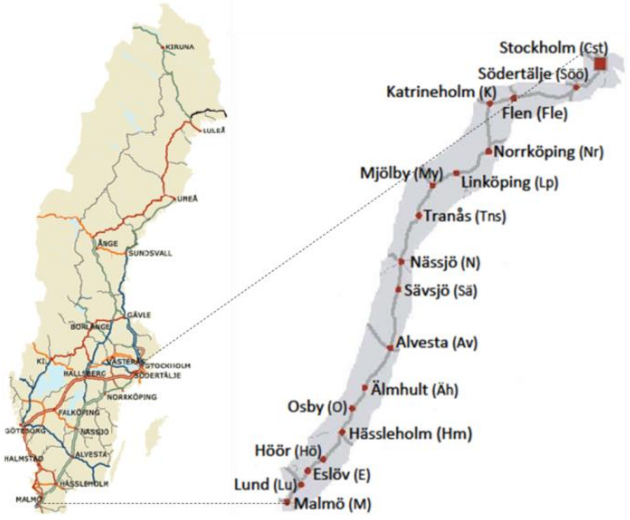


Figure 1. Swedish Southern mainline and the main stations along the line

There is a dense commuter train area between Norrköping–Mjölby and between Hässleholm–Malmö. Over the years, the traffic demand has increased and in the same time the quality of the infrastructure has decreased. This has led to several speed restrictions and maintenance works appearing at the same time, along with a timetable where the trains are scheduled tight, with few margins, to fit them all together. The consequence of this is that the trains operating on the Southern mainline often experience disturbances, they get delayed and have a poor punctuality.

Some trains that are running on parts of the line, such as commuter trains and regional trains, have an acceptable punctuality but long-distance trains that run all the way between Stockholm and Malmö have larger problems. They easily get delayed and since they are running for a long time they often end up disturbing other trains along their way. To get an overall good punctuality in the system, every traffic structure should contribute with the needed robustness and flexibility and we hope to achieve this with new timetable rules.

Several trains get large unexpected delays, over 15-20 minutes, at a short period of time and it is not possible to create a timetable in which these trains can recover from their delays. It would consume too much of the line capacity. Timetable rules to increase robustness should instead be focused on small to medium large delays and increase the possibility for trains with delays around 5-15 minutes to recover from them so these trains can arrive punctual to the end station (i.e. with a delay of 5 minutes at most). A major reason for the medium large delays is that there are a lot of maintenance works and other speed restrictions on the line, to which the timetable is not adapted. If it is not possible to adapt the timetable for all infrastructure variations it is important that the timetable includes enough recovery time so that the delays do not spread too much.

A general finding when studying the timetable and delay statistics is that fast long-distance trains have too short planned dwell times at some stations. In practice, the stops take longer time than the planned 1-2 minutes, which means that time supplement needed for unplanned disturbances in a systematic way will be used for recovering from too long stops instead. This is not a modelling issue but more of a practical problem, not unique for the Southern mainline, since there are no routines for following up actual dwell times today.

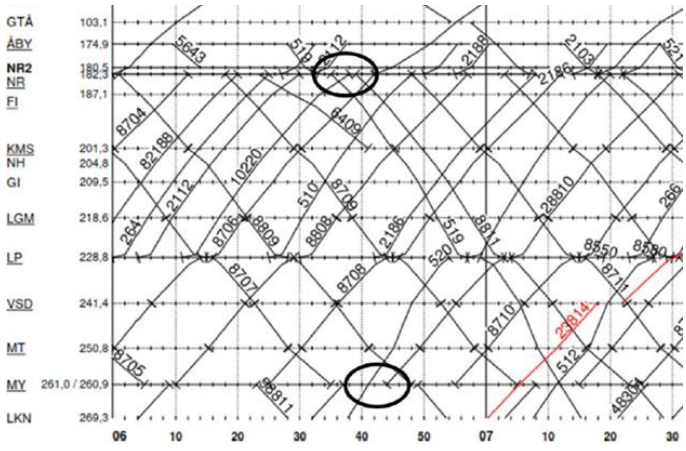


Figure 2. Example of a critical point for southbound trains in Norrköping (NR) with train 519 and 8811 and for northbound trains in Mjölby (MY) with train 520 and 28810.

Another main problem for long-distance trains is that if they get delayed, they often end up after a slower train and have to run with a lower speed for a long time. This increases their delay and also increases the risk for them to delay other trains further on. For example if a southbound fast long-distance train is 5 minutes delayed in Norrköping, it risks to end up after a commuter train and will leave the commuter train area in Mjölby 13 minutes delayed, see train 519 and 8811 in Figure 2. There are several places where this might happen and we refer to them as critical points, which will be described more in section 3.2.

3 Timetable Rules for Creating Feasible and Robust Timetables

When creating a timetable it is important that it is feasible and robust. By a feasible timetable we mean a timetable that is conflict free. It is possible for the trains to run exactly according to the timetable if no unpredicted disturbances occur. By a robust timetable we mean a timetable in which trains are able to keep their originally planned slots despite small disturbances and without causing unrecoverable delays to other trains. A robust timetable should also be able to recover from small delays. Depending on the amount and magnitude of unpredicted disturbances combined with traffic density, the need for robustness differs.

To create a more robust timetable there are two common strategies. The first is to add time supplement to the trains' runtimes. This means that we plan for a longer runtime than it actually takes so that the trains can recover if they get delayed. The downside with this strategy is that it consumes more capacity and that passengers might experience the longer runtime in a negative way. The second strategy is to add headway buffer, which means that we increase the distance between two trains using the same infrastructure. With longer headways, trains do not disturb each other that easily, which prevents delays from spreading. The downside with this strategy is that it consumes line capacity.

3.1 General Timetable Rules used in Sweden

In Trafikverket there are some timetable rules to make the timetables feasible and also to add some time for delay recovery, see Trafikverket (2016a). In this document specific nodes in the railway network are specified together with rules for how much time supplement trains of different categories and speeds should have between these nodes. For example, the nodes on the Southern mainline are Stockholm, Mjölby, Alvesta and Malmö. Passenger trains with a speed above 180 km/h should have a time supplement of 4 minutes between these nodes and passenger trains with a slower speed should have a time supplement of 3 minutes between the same nodes. The timetable planners can place the supplement as they want between the node cities. This strategy was developed nearly 30 years ago and was based on a rough estimation of how much time a train of a certain category needs to recover from typical minor disturbances.

In Trafikverket (2016b) it is specified how closely two trains can be scheduled after each other without causing disturbances, i.e. the minimum headway time between two trains. The headways given in this document is the minimum technical headway rounded up to whole minutes at an aggregated level, resulting in some buffer time between the trains.

3.2 Critical Points

As mentioned before there are points in the timetable that are particularly sensitive to disturbances, referred to as critical points. Critical points appear in a timetable for double track lines where it is planned that a specific train starts its journey after another already

operating train, or where a train is planned to overtake another train. In case of a delay in a critical point, the involved trains are likely to require the same infrastructural resource at the same time which might affect the delay propagation significantly. For more theory about critical points we refer to Andersson et al. (2013) and Andersson et al. (2015).

Each critical point is represented by a specific station and a pair of trains, the *leader* and the *follower*, which interact at this geographic location in such a way that a time-dependency occurs, see Figure 3. The follower refers to the train that starts its journey at the critical point behind another train (denoted the leader), or is overtaken in the critical point by the other train, i.e. the leader. The robustness in critical point p is related to the three margin parts L_p, $F_p$and H_p, and the total robustness for each critical point p, RCP_p, is given by the sum of the parts: $L_p + F_p + H_p$. Below follows a detailed description of the three parts:

- L_p – The available runtime margin time before the critical point for the leader, i.e. the runtime margin for Train 1 between stations A and B in Figure 3. With a large L_p the likelihood of the leader arriving on-time to the critical point increases.
- F_p – The available runtime margin time after the critical point for the follower, i.e. the runtime margin for Train 2 between stations B and C in Figure 3. A large F_p increases the opportunity to delay the follower in favour of the leader, without causing any unrecoverable delay to the follower.
- H_p – The headway margin, or buffer time, between the trains' departure times in the critical point, i.e. the headway margin between Train 1 and Train 2 at station B in Figure 3. In the critical point the trains are separated by the headway margin plus the minimum technical headway. With a large H_p the chance to keep the scheduled train order in the critical point increases, even in a delayed situation.

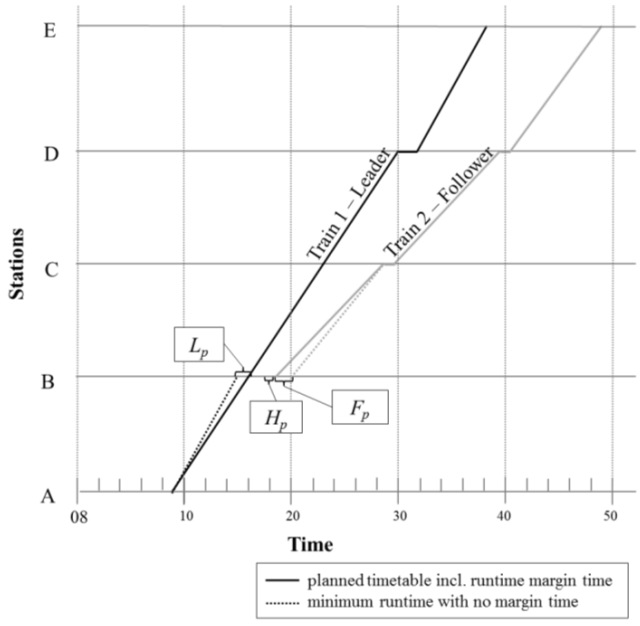


Figure 3: RCP is the sum of the three margin parts: H_p, L_p and F_p.

In Andersson et al. (2015) a MILP (Mixed Integer Linear Programing) model is presented which takes an initial timetable as input, re-allocates the already existing margin time in the timetable to increase RCP and finally returns an improved timetable. The model re-allocates margin time in such a way that the RCP values increase compared to the initial values without increasing the trains' runtimes and the whole timetable gain in robustness. RCP can for example be used in a way to maximize the total RCP for all critical points or as a constraint preventing RCP to be lower than a chosen minimum value.

However, in Solinen et al. (2017) a comprehensive evaluation of a timetable produced by the MILP model is presented, which illustrates that there are some complications when using the produced timetable in a microscopic environment. There are several simplifications and assumptions made in the MILP model which makes it hard to use straight away and the evaluation shows that some trains end up with a lower punctuality even though the overall robustness has improved. This indicates that there are other aspects of timetable planning that needs to be considered in the model as well, so that all trains can benefit from the new timetable, or at least not receive a lower punctuality.

Also, the use of a MILP model requires certain tools and expertise that are not common in current timetabling environments. All of these disadvantages put together makes it hard to use the MILP model in real-world even though the theory behind critical points seems promising.

4 Implementation of RCP and Other Robustness Measures

4.1 New Timetable Planning Rules

To increase the robustness on the Southern mainline new timetable rules has been developed as a complement to the current rules. The rules are separated into two categories, rules to make the timetable feasible and rules to increase delay resistance and recovery.

The rules for feasibility are:

- Time supplement must be added for trains that are planned to use a side track with slow speed switches. A list for all stations and time supplement are included in the new rules as the example in Table 1.
- No planned overtakings on the opposite side of the double track is allowed unless the headway demand in Trafikverket (2016b) is fulfilled also for traffic on the opposite track to prevent trains in the other direction to be disturbed.
- Maintenance works and other speed restrictions that will last for a significant part of the year have to be more carefully planned and sufficient time supplement then has to be calculated and added in the timetable.
- It is not allowed to round off runtimes, time supplements, etc., in a way so that the minimum times are not kept, the maximum deviation is 10 %.

Table 1: Example of how compulsory time supplement for each station is included in the timetable rules. These figures are due for northbound trains that use side track at the stations.

Station ID	Main track nr	Direction of the side track	Time supplement for freight trains (s)	Time supplement for passenger trains (s)
Av	2	North	30	60
Vs	2	North	0	30

The rules for robustness are based on the concept of critical points but with some modifications to make the concept easy to use manually. Not all theoretical critical points are comprised, since they do not always appear in practice. Also, to consider all possible critical points will make it hard to find a feasible timetable unless train slots are removed or significantly modified. Only the most important critical points are selected to be included in the rules and based on real-world circumstances and previous experiences these are critical points where:

- There is a large speed difference between the trains
- The trains interact for a significant amount of time
- There are no overtaking possibility close by
- The points appear in a similar way several times a day
- No freight trains are involved

In the beginning of their run, trains tend to be more on-time than towards the end of their run, which indicates that critical points in the beginning might not need such high RCP values as points in the end. This idea is also discussed in Khoshniyat and Peterson (2017), where the scheduled minimum headway is dependent on the trains' travel times. This idea is not applied for the first version, but it might be an area where the rules can be advanced.

One disadvantage with not using RCP in an optimization model or similar is that it might be complicated for the timetable planners to manually calculate the RCP value for each critical point as they schedule the trains. For that reason each part of RCP, i.e. H_p, L_p and F_p, has been divided into separate timetable rules. Previously, timetable planners have been allowed to place time supplements relatively freely along the line as long as they use the right amount of time supplement between certain nodes as mentioned in Section 3.1. In the new rule the time supplement placement for long-distance trains is more strict and based on the location of critical points. For example, southbound long-distance trains must have the placement of time supplement as presented in Table 2. Here 2 minutes are placed before the most critical points Nr, Av and Hm and 1 minute before the other, not so critical, points. This can be related to L_p in Figure 3.

There is also a new rule to control the minimum headway time at the selected critical points. According to the new rule there must be 6 minutes between the fast and slow southbound passenger trains' departure times in Nr, N, Av and Hm. For northbound trains the limit is also 6 minutes but the locations are instead My, N, Av and Hm. The technical minimum headway time is 2-3 minutes for these stations which then results in a headway margin, H_p, of 3-4 minutes.

Table 2: The placement of time supplements for southbound long-distance trains

Stretch	Time supplement (minutes)
K – Nr	2
Nr – Tns	1
Tns – N	1
N – Av	2
Av – Hm	2
Hm – Lu	1
Lu – M	1

For F_p there is no timetable rule, but there is an operational prioritization rule that states that the follower in a critical point can be delayed up to 3 minutes in favour for the leader.

When combining the three rules this means that there is a total RCP value of around 8 minutes in each critical point. These 8 minutes consist of:

- 1-2 minutes time supplement for the leader before the critical point (L_p)
- 3-4 minutes headway margin time (scheduled headway time - technical minimum time) (H_p)
- 3 minutes for the follower that can be delayed (F_p)

4.2 Timetable Effects of the New Rules

Since the railway market in Sweden is fully deregulated, the timetable is highly dependent on how the different train operators request for train slots every year. The demand for train slots tends to increase and it is a time consuming process for the timetable planners to create a new timetable each year. This means that the timetable will change from year to year to some extent and it is hard to interpret exactly which effects are due to the new rules and which effects are due to changed train slot requests from the operators.

The most obvious difference between the timetable for 2018 (T18) and the timetable for 2019 (T19) is that the runtimes for fast long-distance trains have changed. The average runtime for all trains combined is still around 4 h and 27 min but for southbound trains the runtime has increased by approximate 6 minutes and for northbound trains the runtime has decreased with 5-15 minutes, see Table 3. According to the timetable rules the dwell times at some stations have increased in T19 but the total amount of dwell time is kept almost the same since the operator has chosen to remove one stop for each train. Instead of making 8 stops in T18 most trains only stop 7 times in T19.

The main reason for the changes in total runtime is that a lot of extra time that is needed to fit all trains together in the timetable has to be added or removed compared to previous year. Depending on how the train slots are requested by the operators, it might be necessary to add extra time for one train to make room for another. The amount of time needed differs from year to year. For example, between Malmö and Lund the traffic is very dense which has led to northbound long-distance trains in T18 having a longer stop in Lund than necessary (6 minutes instead of the needed 2). In T19 most of this extra dwell time is not needed which decreases the total runtime. Instead southbound long-distance trains have to have a longer stop in Lund which increases the total runtime with 2-3 minutes.

Table 3: Timetable change (in minutes) from T18 to 19 for some representative long-distance trains. Trains with odd train ID are southbound and trains with even train ID are northbound.

Train ID	Total travel time	Total dwell time	Time supplements			Rounding
			Maintenance work	Timetable synchronization	Delay recovery	
519	+6	+1	-2,4	+5,0	+1,4	+1,5
525	+6	0	-2,4	+8,0	+0,2	+2,5
537	+6	+1	-2,4	+8,5	-0,3	+0,8
522	-11	+1	-0,5	-4,0	+0,7	-3,5
530	-9	+1	-0,5	-1,5	+1,0	-3,7
540	-6	+1	-2,5	-2,4	+1,5	-2,0

If we compare the amount of time supplement needed for timetable synchronisation in Table 3, we can clearly see that southbound trains have an increased amount of time supplement and northbound trains have a decreased amount of time supplement.

In Table 3 we can see that the amount of time supplements differs a lot between T18 and T19 and also between individual trains. For example, in T19 there is less time added to handle maintenance works, but the time added for delay recovery has increased some. However, the main reasons for the change in total travel time are the time supplements not related to delay recovery.

Also other trains in the timetable have been affected by the new timetable rules, some trains that are using side tracks have been given longer runtimes if it is necessary according to Table 1 and some trains have been moved backwards or forwards so that the minimum headway time of 6 minutes is fulfilled. Some of the most affected trains are the extra rush hour commuter trains that run between Norrköping and Mjölby at 6:00-8:30 a.m. and 15:00-18:30 p.m. The main structure is that there are commuter trains running in a periodic 30-min timetable and the operator wants the extra trains to build a periodic 15-min timetable during rush hours. However, due to the amount of traffic, the rule of 6 minute headway in the critical points combined with the trains' different speed and stopping pattern makes this hard to achieve. Sometimes the commuter trains run with 14/16 minute distance, sometimes with 10/20 minute distance and in one occasion the intermediate time is 5/25 minute.

In the southern commuter train area the timetable planners have chosen to deviate from the 6 minute headway rule since the leader has a long dwell time planned before the critical point. This extra dwell time can be seen as margin time of type L_p in Figure 3. The total RCP value stays the same which makes this shift from a longer H_p to a longer L_p acceptable. The deviation from the 6 minute headway makes it easier to combine all train slots and we can avoid to add too much time supplement and increase the runtimes.

For most trains it is possible to follow all timetable rules, but in some rare situations the timetable planners had to deviate slightly from either the node placement or the 6 minute headway to be able to create the timetable.

4.3 Preliminary Punctuality Effects of the New Rules

The timetable with the implemented new timetable rules, T19, was applied the 11th of December 2018, which means that there is not much time to gather statistics. The following reasoning is based on only one month of statistics and can simply give a preliminary indication of the result. The statistics are gathered from all normal weekdays from the same winter period in T18 and T19. All trains with a large disturbance of 15-20 minutes or more are excluded from the results and a train is considered punctual if it is at most 5 minutes delayed at the end station.

When studying fast long-distance trains there is a clear improvement from T18 to T19. Southbound trains, that had the worst punctuality in T18, have an increased punctuality from 74% to 90%. Northbound fast long-distance trains had a better punctuality from the beginning but their punctuality has also improved from 87% to 93 %. The punctuality for regional passenger trains has not changed that much, in the north part of the Southern mainline it has improved slightly and in the south part it is around 96% in both T18 and T19. The punctuality for freight trains has improved around 8 percentage points from T18 to T19 and one reason could be that the other trains are more punctual due to the new rules and do not disturb the freight trains that frequently.

Commuter trains between Norrköping and Mjölby have received a higher punctuality since the risk of being delayed due to an already delayed long-distance train has decreased

with the new timetable rules. The punctuality for commuter trains is measured as a delay of maximum 3 minutes (instead of 5 minutes) and the punctuality has increased from 95% to 97%. For commuter trains in the south of Sweden the timetable has not change that much and the punctuality has decreased with some percentage points due to other reasons than the timetable design.

5 Concluding Discussion

In this paper the use of new timetable rules and their effect on the timetable and punctuality are analysed. The main focus is to study how the robustness measure RCP can be used manually to increase the timetable robustness as a short term solution, until we have a software or optimization model to support the timetable planners. The results show some of the difficulties when moving from theory to practice and what can be done with limited resources in reality. If there is no practical possibility to use an optimization model to increase RCP, the timetable planners have to do it manually when they create the timetable. It soon becomes hard to grasp all consequences of adjusting train paths and it is also hard to know which solution is the most just for all operators. In the presented case study, some fast long-distance trains got longer runtimes and some commuter trains did not get their desired periodic timetable. For an experienced timetable planner this might be hard to overview but it is even harder to find suitable constraints in an optimization model that can handle all aspects and all eventualities in a way that would not distort competition. With a manual approach it is easier to consider the experience of the timetable planners, they have gain knowledge from previous timetables and sometimes know intuitively what is possible to achieve and what is not.

One other finding is that it is not possible to have hard rules for all situations, it is necessary to make deviations from time to time, not to end up with unacceptable consequences for the trains. This is one reason why it is hard to use robustness rules as unbreakable constraints in an optimization model.

The manual implementation shows that there are still some difficulties that need to be solved before RCP and other robustness measures can be applied in a timetable optimization model. However, the preliminary punctuality results indicates that the concept of critical points and RCP can be useful also in a manual way to improve the punctuality. We can for example see a large improvement for southbound long-distance trains who have suffered from a poor punctuality for years.

To support the timetable planners the presented robustness measures could be implemented in a software tool, helping them to get a better overview of the robustness and how their decisions affect the robustness. The future plan at Trafikverket is to, in a few years, start to use a software tool including more microscopic data when creating timetables. This will result in a higher degree of feasibility and the need for timetable rules concerning feasibility decreases. In this future software it could be possible to also include the RCP calculation and the presented robustness rules. To illustrate the robustness with the presented rules in a software tool can be seen as a step towards automatic timetable construction and the results from this study can be used as an input to which rules that need to be implemented.

References

Andersson, E.V., Peterson, A., Törnquist Krasemann, J., 2013. "Quantifying railway timetable robustness in critical points", *Journal of Rail Transport Planning & Management* 3, pp. 95–110.

Andersson, E.V., Peterson, A, Törnquist Krasemann, J., 2015. "Reduced railway traffic delays using a MILP approach to increase robustness in critical points", *Journal of Rail Transport Planning & Management* 5, pp. 110–127.

Khoshniyat, F., Peterson, A., 2017. "Improving train service reliability by applying an effective timetable robustness strategy", *Journal of Intelligent Transportation Systems 21*, pp. 525-543.

Solinen, E., Nicholson, G., Peterson, A., 2017. "A microscopic evaluation of railway timetable robustness and critical points", *Journal of Rail Transport Planning & Management 7*, pp. 207-223.

Trafikverket, 2016a. *Konstruktion av körplaner för tåg* (Railway timetable construction, in Swedish), TDOK 2016:0128, Guideline from the Swedish Transport Administration.

Trafikverket, 2016b. Riktlinjer täthet mellan tåg (Guidelines for headway between trains, in Swedish), TRV 2016/831, Guideline from the Swedish Transport Administration.

Interlocking System Based on Concept of Securing a Train Travelling Path

Tetsuya Takata [a1], Akira Asano [a], Hideo Nakamura [b]
[a] Development Center, Kyosan Electric Mfg. Co., Ltd.
2-29-1, Heian-cho, Tsurumi-ku, Yokohama, 230-0031, Japan
[1] E-mail: takata-t@kyosan.co.jp, Phone: +81 (0) 45 503 8115
[b] The University of Tokyo
5-1-5, Kashiwanoha, Kashiwa, Chiba, 277-8561, Japan

Abstract
In recent years, the environment of railways and the systems such as CBTC (Communication Based Train Control) have been changing. To respond the changes and the needs of customers, a unified train control system (UTCS) has been developed to realize a system that evolves with customers.

Previous type systems consist of independent components such as ATC (Automatic Train Control) system, electronic interlocking system, and facility monitoring system, and there are a complicated overlap of system configurations and functions and difference in concept between the systems. On the other hand, the integrated train control system consists of horizontal layers such as function layer, network layer, and terminal layer. Therefore, the system has been developed to make it simple with no unnecessary redundancy and evolving to meet the needs of customers. In this paper, we explain a method that realizes the interlocking function in the function layer based on the concept of "securing a train travelling path" including path blocking and routing, and evaluate the safety of the method using STAMP/STPA (Systems-Theoretic Accident Model and Processes/System Theoretic Process Analysis).

Keywords
Railway signaling, interlocking System, safety assessment, train control system, CBTC, UTCS, FMEA, STAMP/STPA.

1 Introduction

Interlocking system is a train control system that realizes collaborative control of branching direction or permission for trains to travel, in order to prevent collision or derailment of trains.

As a result of individual development of block system, ATC system, interlocking system, and facility monitoring system, the train control system consists of vertically-divided independent components. Integrated train control system is developed by reorganizing the train control system to have horizontally-divided layers including function layer, network layer, and terminal layer. The reorganization of the system incorporates the control logic into the function layer and therefore the interface between the systems is rational.

Integrated train control system is developed by reorganizing the train control system to have horizontally-divided layers in "hierarchical configuration" including function layer, network layer, and terminal layer. This reorganization of the system not only integrates the functions and reduces on-site facilities, improving the system reliability, but also incorporates all the control logics into the function layer. Therefore, the interface between the systems is rational. Development of the rational interface reduces train accidents caused

by an error of the interface and enhances the safety.

Necessary functions for a train to travel on a track can be roughly classified to the one to "secure a train travelling path" such as blocking and routing functions and the one to "control safety" such as signal and speed control functions. If the entire system is reorganized with the above-mentioned layer components on the basis of the concept of "securing a train travelling path" and "safety control," the "exclusive control" that has been considered necessary and the "overlapping functions" of each system could be eliminated and a simple system can be established.

In this paper, we explain a method to realize an interlocking function based on the concept of "securing a train travelling path" such as blocking and routing and evaluate its safety using STAMP/STPA.

2 Interlocking based on concept of securing a train travelling path

2.1 Concept of securing a train travelling path

Conditions for safe travelling of trains that were indicated before are as follows.

(1) The travelling path shall be fully configured and secured. Namely, the points on the path shall be switched and locked to the travelling direction.

(2) No train or carriage shall exist on the travelling path.

(3) There shall be no possibility of other trains to travel on the path.

(4) The above state shall be maintained until the train passes over the path.

This can be summarized as follows from a viewpoint of securing a train travelling path.

(1) The travelling path shall be fully configured and secured. Namely, the switches (points) on the path shall be switched and locked to the travelling direction.

(2) No train or carriage shall exist on the travelling path occupied by a train.

(3) Other trains shall not be able to travel on the occupied path.

(4) If the train passes over a division of the travelling path, it loses the right to occupy the division.

For a train travelling path, block points are introduced to define the points where a train on the path is blocked to allow other trains to travel on the path. If a block point is set on a train travelling path which a train requests to occupy, the train is given the right to occupy a distance from the head of the train to the block point and the right is used as the train control condition.

2.2 Setting of travelling path and block points for train interval control

A travelling path is defined as a set of sections. Then, block points are introduced to define the points where a train on the path is blocked. If a block point is set on a train travelling path which a train requests to occupy, the train is given the right to occupy a distance from the head of the train to the block point and the right is used for the train control.

For example, the block points are set as follows.

(1) Block point 1: End of a train travelling in front on the travelling path (moving point)

(2) Block point 2: Position of a point on the travelling path (fixed point)

(3) Block point 3: Position related to the travelling path occupied by an oncoming train (fixed point)

2.3 Interlocking function

Unlike the previous interlocking function which has individual circuit logics based on interlocking circuit data of each station, the interlocking function developed under the

concept of securing a travelling path has a shared program as a logic to secure the safety of travelling paths. A conceptual diagram is shown in Fig. 2-1.

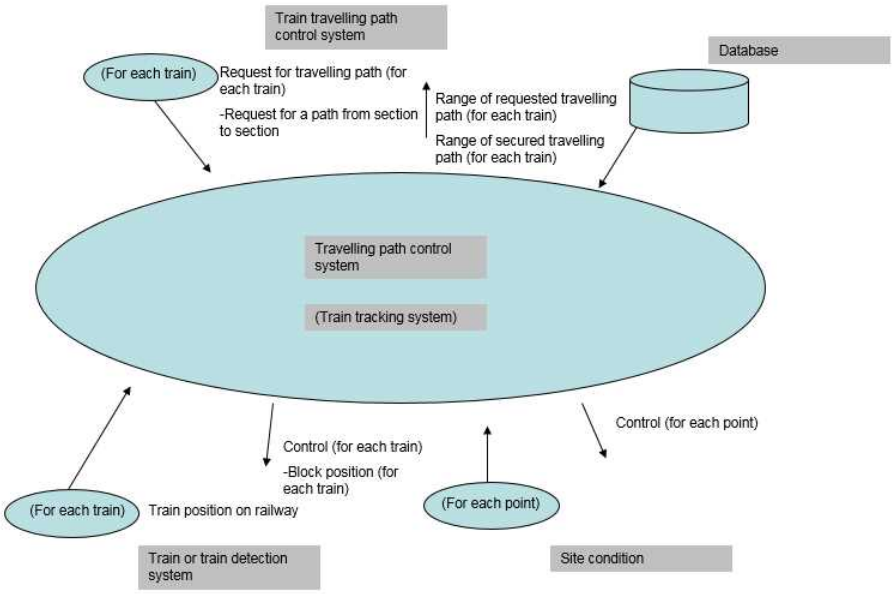


Fig. 2-1 Logic of safety securing

The interlocking function works in the following steps.
(1) When a travelling path (which expresses a path from a starting point to a destination and is defined as a set of sections) is requested (under the control of each train), a travelling path status table for the travelling path is created and an interlocking processing is performed according to the table created. (The table is deleted when the path request is cancelled.)
(2) In the travelling path status table, a series of sections based on a travelling path data table are described and control status of points based on request-acceptance status of each train for a given request and the railway form data table are registered.
In addition, on the basis of the block point data table and the block point positions of each train, an allowed area of each section of the travelling path is registered.
(3) A point is controlled in accordance with its control status and the allowed area of the travelling path is updated on the basis of the indicated status of the point.
(4) Train interval control is made by transmitting the nearest block point position to the trains based on the allowed area of the travelling path.
(5) The current position information of a train is updated according to the train travelling. As the block point position of each train is updated, block point positions of the trains on the path and the released area of the travelling path are set in the travelling path status table.

2.4 Idea about each locking

Since the present system concentrates the logics to the processing unit, the locking conditions for the interlocking can be made as follows.
After the integration of the logics, the rout becomes a travelling path and the functions of route locking (which prevents relevant points from switching until the train or carriage

passes over all points in a route so that other routes that could block the route would not be formed, when a train or carriage enters the route by following an aspect of a signal that directs proceeding or clearance indication of a shunting indicator,), sectional route locking (which divides the route-locked sections and successively unlocks the sections over which a train or carriage passes to improve the efficiency of the train operation and station work), detector locking for signal lever (which is an interlocking between a signal and track circuit to lock the signal to a normal state when a train or carriage exists in the track circuit of the signal on the route), and detector locking (which does not allow a train or carriage to switch a point if the train or carriage exists in the track circuit where the point is installed) are satisfied by the travelling blocking logic that controls a single blocking and single train on the basis of the right given to the train to occupy the path (blocking).

In addition, the functions of approach locking (When a signal is made indicate a sign of proceeding and then a train enters the approach locking section of the signal or when a signal is made indicate a sign of proceeding while a train is entering the approach locking section of the signal, the approaching locking locks points in a route to prevent them from switching for a certain period of time after the train proceeds to the protection area of the signal or after a stop signal is made.), stick locking (which locks points in a route to prevent them from switching in the following cases: During the time period after a signal or shunting indicator is made indicate a sign of proceeding until a train or carriage enters the protection area and during a specified time period after a signal is made indicate a sign of stopping), and time locking (which keeps locking for a certain period of time even when levers of a signal and point are changed from the reserve to normal position) are satisfied after the integration of the logics, since control is made on the basis of train position information by a closed loop between the central station and trains.

Check locking (installed between levers in different signal cabins) is not necessary because of the centralized control. Circuit processing for indicating locking (which checks the consistency between the status of the signals and points and that of the lever and prevents dangerous control if inconsistency is found) is not necessary since on-site conditions of the point control and signal control are compared.

Therefore, the locking logic that the previous type of interlocking system used in the interlocking circuit for each station is not necessary.

3 Failure analysis of software and STAMP

3.1 Analysis of software failures

Many faults occur due to failure of software, although there is no appropriate method to analyze influence of the software failure on the system.

Even FMEA（Fault Tree Analysis） and FTA（Failure Mode and Effect Analysis） contain some shortcomings, although they are often used as a method of failure analysis.

Fundamentally, FMEA has no means to define software failures and assess their impact. Loops, wrong branches and other failures may appear in many different locations, and besides, it is not possible to uniquely define how software behaves in the event of such a failure. Today, a common method of performing FMEA is to focus on the functionality of modules and predict their possible malfunctions. However, this is only a methodology that has been devised as a means of using FMEA instead of paying attention to software bugs. Likewise, FTA, which starts an analysis with a malfunction mode of a system toward deeper levels, can only end with clarifying malfunctions of functional modules, instead of finding out software bugs.

As a solution to overcome such limitations, an accident model called STAMP that focuses on interactions among modules and controls has been advocated by Nancy Leveson. STAMP is spotlighted for its effectiveness in analyzing safety of software-intensive systems.

3.2 Assessment by means of STAMP

STAMP is characterized by the ease of identifying causes of accidents attributed to the design of an entire system such as system mechanism, technologies, human errors and miscommunication among projects, all of which have been difficult to discover by means of conventional accident assessment models (FTA, FMEA etc.). Hazard analyses are performed to identify the causes of accidents (hazards) prior to the occurrence of the accidents and STPA is used as a tool for the hazard analyses. The hazard analysis process using STPA consists of the following four steps.

(1) Preliminary Step 1: Identification of accidents, hazards and safety constraints

In this first preliminary step, accidents, hazards and safety constraints are prepared. This intends to predefine events which systems should prevent and such predefined events are in turn used as input to STPA Step 1.

- Accident: a system accident causing a loss
- Hazard: a system state leading to an accident
- Safety constraint: a rule necessary to maintain the safety of a system

(2) Preliminary Step 2: Establishment of a control structure

A control structure is a diagram depicting the interrelation among functions that control a system. It represents the flow of orders for controls and feedback exchanged among components using arrows.

(3) STPA Step 1: Identification of unsafe control actions (UCAs)

In this step, UCAs that may lead to a hazard are identified and categorized into the following four types:

- Not Provided: Control actions necessary for safety are not provided.
- Incorrectly Provided: Unsafe control actions that may lead to a hazard are provided.
- Provided Too Early, Too Late, or Out of Sequence: Control actions are provided too late or too early, or not provided in a predetermined sequence.
- Stopped Too Soon or Applied Too Long: Control actions stop too soon or are applied too long.

(4) STPA Step2: Identification of hazard causal factors (HCFs)

In the last step of STPA, causal factors of UCAs identified during STPA Step 1 and expected accident scenarios are identified. Causal factors are potential flaws that may appear in a control loop, which are classified according to the following 11 guidewords:

- Control Input or External Information Wrong or Missing
- Inadequate Control Algorithm (Flaws in Creation, Process Changes, Incorrect Modification or Adaptation)
- Process Model Inconsistent, Incomplete or Incorrect
- Component Failures, Changes Over Time
- Inadequate or Missing Feedback, Feedback Delays
- Incorrect or no Information Provided, Measurement Inaccuracies, Feedback Delays
- Delayed Operation
- Inappropriate, Ineffective or Missing Control Action
- Process Input Missing or Wrong
- Unidentified or Out-of-Range Disturbance
- Process Output Contributes to System Hazard

4 Safety assessment

4.1 Assessment result by STAMP/STPA

As mentioned above, the interlocking system controls a travelling path of a train at a station with points.

A conceptual diagram is given in Fig. 4-1.

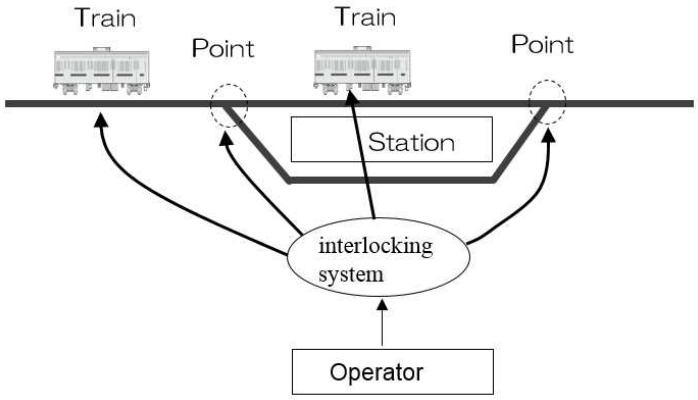


Fig. 4-1 Conceptual diagram of electronic interlocking system

An accident due to the travelling path control is defined as follows on the basis of the conceptual diagram.

(A1) Collision of trains

(A2) Derailing of a train

(A3) Contact of trains

As a result of the analysis of the interlocking system using STAMP/STPA for these accidents, HCF was identified for UCA, although details are omitted here. Possible measures and specific actions for the measures are summarized.

Some of the identified HCFs were categorized as the ones that should be handled by a method other than the interlocking system. Those include the HFCs which need to detect trains securely, such as (1) “A train travelling over a switch cannot be detected” or “A train cannot be recognized correctly,” and the HFC such as (2) “Train collision could occur if a train start travelling when a travelling permission is issued.”

Next, the HFCs of (3) “Switching restraint is not given due to inappropriate control algorithm” and “Switching control is output due to inappropriate control algorithm” require detector locking with an electric locking method. (4) “Travelling permission is immediately cancelled in a situation where a train cannot stop due to inappropriate control algorithm” requires approach locking or stick locking with an electric locking method. (5) “Disapproval of travelling is output but the output of the travelling aspect remained” requires indicating locking with an electric locking method. These were categorized as those associated with the locking conditions of the interlocking.

4.2 Analysis of existing electronic interlocking and present locking

Specific actions for the measures, listed below, are safety function requirements from the interlocking system.

(1) The input circuit for the switching direction of a point shall be constantly checked to make sure of its normality and be made unswitchable if an abnormality is found.

(2) The input circuit for the current position information of a train shall be constantly checked to make sure of its normality and be switched to choose presence of a train.

(3) Locking shall maintain if a signal does not indicate a stop aspect.

(4) A switch shall be made unswitchable while a train exists over it.

(5) Travelling permission control shall be monitored and, if an abnormality is found, it is switched to choose safety side.

(6) A travelling permission shall be cancelled with time for the train to stop in the allowed area.

(7) Status of a point shall be constantly monitored.

(8) Switching restraint control shall be monitored and, if an abnormality is found, it is switched to choose safety side.

(9) Switching control shall be monitored and, if an abnormality is found, it is switched to choose safety side.

It was clarified that, in the previous type of interlocking system, the safety function requirements depended on data of each station, while they depended on the software in the present system. Therefore, in a case where the interlocking function is realized by using circuit data of each station as done in the previous interlocking system, the present interlocking system does not need to verify the safety of the individual data of each station if the safety of the S/W is checked once.

There should be no particular problem in the software if the development method and in-company checking system for the software, which have been proved successful, are continued and if international standards such as IEC 62279 are referenced.

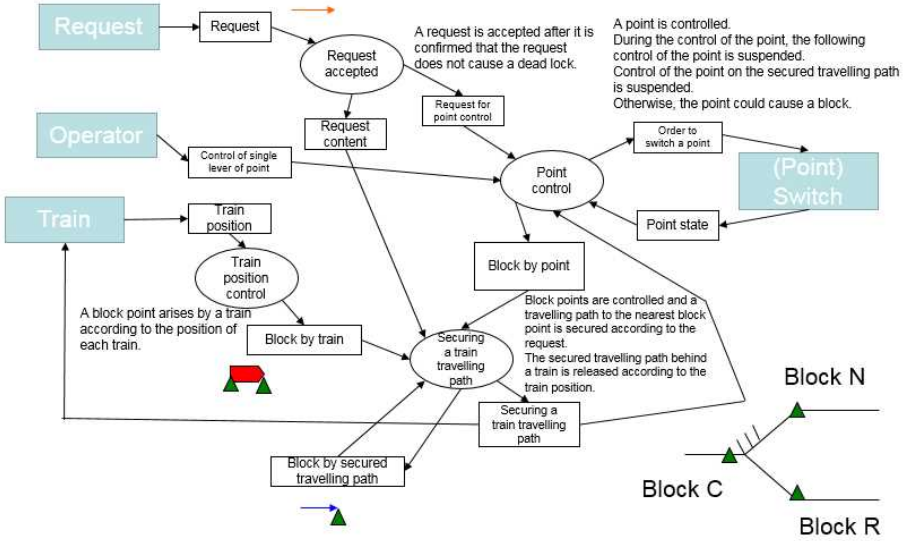


Fig. 4-2 Control flow of interlocking system

5 Conclusions

In this paper, we explained a method to realize an interlocking function based on the concept of "securing a train travelling path" and evaluated the safety of the interlocking system using STAMP/STPA. From the result of the evaluation, we showed the difference from the existing interlocking system and clarified that the interlocking could be realized even without circuit logic of individual stations.

The above method to realize the interlocking function of trains can also be applied to street cars (trams). In the case where interlocking with traffic signals is necessary, similar control can be realized if red signals are considered as hindering points. However, for the realization, it would be necessary to organize the timings of traffic signals and block points release for a train to travel forward and turn right and left. Then, specifications should be made for the organized timings. It would also be needed to make conditions to re-secure the secured train travelling path if the traffic signal condition changes because, for example, the path was secured once at the request of a train but the train could not start travelling due to too many people getting on and off the train.

The authors would like to express gratitude to cooperators from Kyosan who provided us with much advice on our study.

References

Akira ASANO, Tetsuya TAKATA, Hideo NAKAMURA, 2015, Integrated train control system, STECH2015

Yoshihisa Saitou, Akira Asano, Hideo Nakamura, Sei Takahashi, 2016, A Proposal for the Design of integrated Train Control Systems Capable of Improving Reliability and Safety, Railways2016.

Information-Technology Promotion Agency, 2016, First STAMP/STPA,1sted., Apr.2016.

Railway Bureau, Ministry of Land, Infrastructure, Transport and Tourism Technical Regulatory Standards on Japanese Railways.

IEC62278:2002. Railway applications - Specification and demonstration of reliability, availability, maintainability and safety (RAMS).

User-centered development of a train driving simulator for education and training

Birgitta Thorslund[1], Tomas Rosberg, Anders Lindström, Björn Peters
The Swedish National Road and Transport Research Institute (VTI)
581 95 Linköping
[1]Telephone: Int +46 13 20 41 55
E-mail: {birgitta.thorslund|tomas.rosberg|anders.lindstrom}@vti.se

Abstract
A user-centered, agile approach was used to develop a high-standard train simulator for applications in train driver education. Thus, a user group of train operators and train driver educators was formed to share experience and development cost. Joint prioritisation by the user group was used in combination with agile development to iteratively develop new versions of the train simulator, responding to the user group's demands. The user group has grown from 2 organisations in 2015 to 10 organisations in 2019, each of which now use the train simulator in education and training. This approach has been beneficial not only in terms of quality, cost and time. But also, as a response to a highly competitive and fragmented market. Early on in the process, the user group requested process results from applied research. This is well in line with VTI's objectives, and three PhD projects have already started and several other projects on driver behaviour, railway signalling systems and capacity have been initiated.

Keywords
Train driving simulator, User-centered development, Agile development, Simulation-based training, Railway safety systems, Railway signalling systems

1. Introduction

Railways provide one of the safest modes of travel and transport available today. Maintaining a high level of traffic safety in the railway system depends directly on the possibility to educate and train staff to understand and make proper use of the technical and procedural knowledge associated with the train protection system (e.g. ATC, STM, ETCS, ERTMS). These systems are implemented both as technical solutions in signalling systems, in rules and regulations, as well-documented organisational roles that help assign and maintain responsibilities, and as established best practices for handling exceptional situations. The demands on train drivers for competence in these areas are high, as well as on personal qualities such as being responsive, yet meticulous and stress-resistant.

In Sweden, the educational paradigm for train drivers has traditionally been predominantly theoretical, for two main reasons: The first reason is the safety aspect, which prohibits practicing on track in real trains until the train driver candidate has achieved a relatively high level of theoretical proficiency, and even then, only under strict supervision. The second reason is lack of educators and train driver supervisors, which is a strongly limiting factor in train driver education and training (Svenska Byggbranschens Utvecklingsfond, 2018). And even if on-track training with skilled supervisors was – hypothetically an option, many critical situations occur quite rarely in real life. Furthermore,

on-track training provides less controllable training with respect to training curriculum.

The factors mentioned above makes on-track training and experience build-up a relatively slow, inefficient and hard to control approach. At the same time, many of these critical situations require accurate, timely and often immediate action, and their handling therefore needs to be well-rehearsed. Furthermore, the working situation for train drivers is special, since a long period of very limited cognitive load, requiring endurance, can suddenly be replaced by an extremely stressful situation. Such events put high demands on attention, responsiveness, and problem-solving capabilities, and yet must not lead to a loss of focus on safety and procedural correctness. Therefore, a combination of theory and targeted practice of specific situations could potentially improve the effectiveness, quality and efficiency of train driver education. Handling of critical situations could be trained before they occur and save both lives and financial resources.

Driving simulators offer ample opportunity for this type of efficient and relevant, targeted training. They can be used to provide training scenarios featuring both the interactivity, complexity, and simultaneous use of different domains required not only to train novice drivers, but also to train or refresh the skills of older drivers (Lees, Cosman, Lee, Rizzo, & Fricke, 2010; Casutt, Theill, Martin, Keller, & Jäncke, 2014; Pollatsek, Vlakveld, Kappe, Pradhan, & Fisher, 2011). Train driver educators in Sweden have recently started to recognise the benefits of using train simulators in education. The pedagogical advantages with simulator-based training are many, including the possibility to layer theory and practice (Hedman, 2017). In aviation, pilot training in simulators has been common practice for many decades, and simulator-based training is also common in maritime and naval education. Although there are several commercially available train simulators on the international scene, they have not to date been considered a viable option for Swedish actors. This is because they do not include Swedish railway sections, and another reason is that they are based on proprietary, closed software, making joint development and research approaches more difficult. Furthermore, the cost of these commercially available solutions is typically not within the budget frame, neither of national Swedish train driver educators, nor of train operators who typically would rather be in need of a larger quantity of low-cost simulators. National Swedish educators and train operators share the view that in order to allow for as many drivers as possible to be able to use the train simulator for practice, it is important to keep the costs down.

The Swedish National Road and Transport Research Institute (VTI) has more than 40 years' experience in using simulators and is a leading authority in conducting simulator experiments and developing simulator methodology and technology (Thorslund, 2013) primarily targeting road traffic. For some reason, it has been more difficult to get funding for research on rail traffic. Since 2015, a scheme for development of train simulators, based on the same proprietary simulator software and with existing Swedish tracks represented, has been successfully developed, based on user involvement and needs, with a main application in train driver education. During this process, agile methods (Agile Alliance, 2018) and user-centered systems design (UCSD) (Norman, 1986) have been employed, and this contribution describes this development process as a case study and summarises the benefits of joint collaborative development of train simulators in this fashion, which is unique.

2. Objective

There are two general aims of this work. The first aim is to explore a method to continue the development of VTI's train simulators, using UCSD. This should enable cost-effective

and efficient simulator-based driver training on actual Swedish tracks, as a complement to the traditional approach of real-world practice combined with theoretical studies. Specifically, the initial users (train education academies and operators) of the train simulators had expressed the need for rather specific use cases, in the form of implemented driving scenarios on actual Swedish tracks, for individual training. Such scenarios require substantial amounts of software development and would consequently be quite difficult for each user to fund individually. To the concepts of UCSD and agile development was therefore added a third component to the development scheme, namely the idea to share both costs and knowledge between users. The second aim is to create a platform for performing research on driving behaviour from a traffic safety, efficiency and capacity perspective.

The research questions to be answered are:

1. How well does this scheme for joint, agile, user-centered development work with regard to
    a. developing train simulators, and,
    b. carrying out research on driving behaviour?
2. What are the main beneficial outcomes or results from this scheme, and who benefits from these results?

3. Method

The project described here is unique in that it brings together competitors on a de-regulated market in a user-centered collaborative scheme, with the goal of jointly developing a realistic, validated educative simulator tool, to improve traffic safety for all parties involved, and ultimately for the benefit of society. To this end, a user forum was created by inviting train operators and train staf educators, operating in Sweden. They were enrolled in a formally arranged user group, called "TUFFA" ("Tågsimulatorutvecklingsforum för användare" i.e. "Train simulator user development forum", in translation). To join, a pre-requisite was that each member acquires a train simulator from VTI, at a nominal cost of approximately € 15,000 – 25,000, dependent on physical configuration. This is necessary to be able to share knowledge, experience and requests for development. An annual membership fee of € 10,000 was also compulsory. Regular meetings were arranged to generate and prioritise short lists, governing the software development to be carried out by VTI for the joint fees obtained. All development results and updates were shared among the members of TUFFA by digital distribution (via an FTP server), keeping individual costs down.

Two methods are used to answer the research questions. Question 1a) is answered by observing how the TUFFA initiative has proceeded and what joint prioritisations and results the UCSB approach has achieved. Question 1b) is answered by making an inventory of the research projects initiated as a result of the TUFFA initiative. Question 2 is answered through a survey among the TUFFA user group.

3.1 UCSB approach

The user group meet twice per year at each other's sites. These are very similar as can be seen in Figure 1. The first companies to join the user group were train driver academies and they invested in passenger train simulators, like the one at VTI. Freight train operators waited for the freight train model to be established, which was performed during 2016 (Andersson, Lidström, Peters, Rosberg, & Thorslund, 2017). During the user meetings, knowledge is shared, and decisions are made on which developments that should be made.

Examples of shared knowledge is special scenarios and ideas on how to structure test protocols.

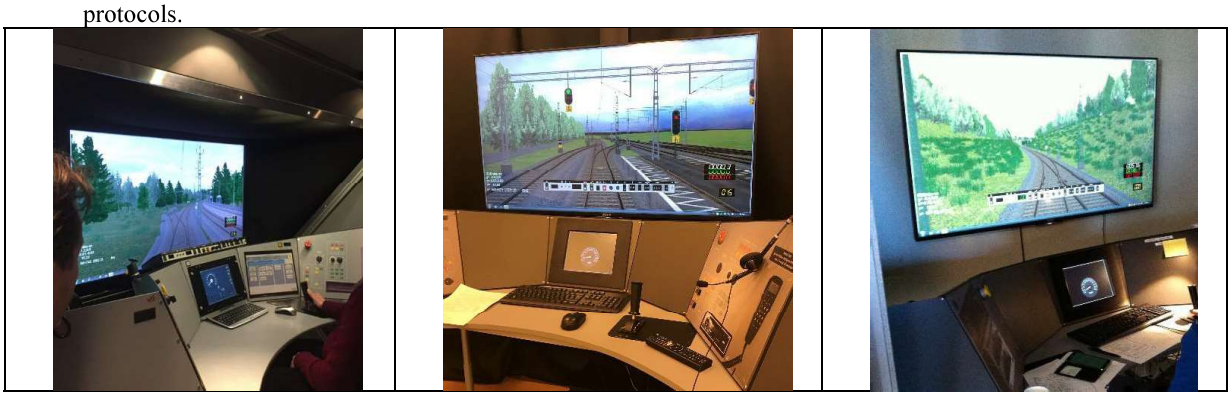


Figure 1 User meeting at 3 different sites. From the left; VTI, Nässjöakademin, and TCC.

The multi user centred design in the user group is described in Figure 2. The main developing loop is driven by the TUFFA member group and has a 6-month cycle time from setting requirements and needs to delivery to train simulation code base. However, parallel with this activity there can be several UCSD processes driven by special customer needs. These activities have in general a shorter cycle time.

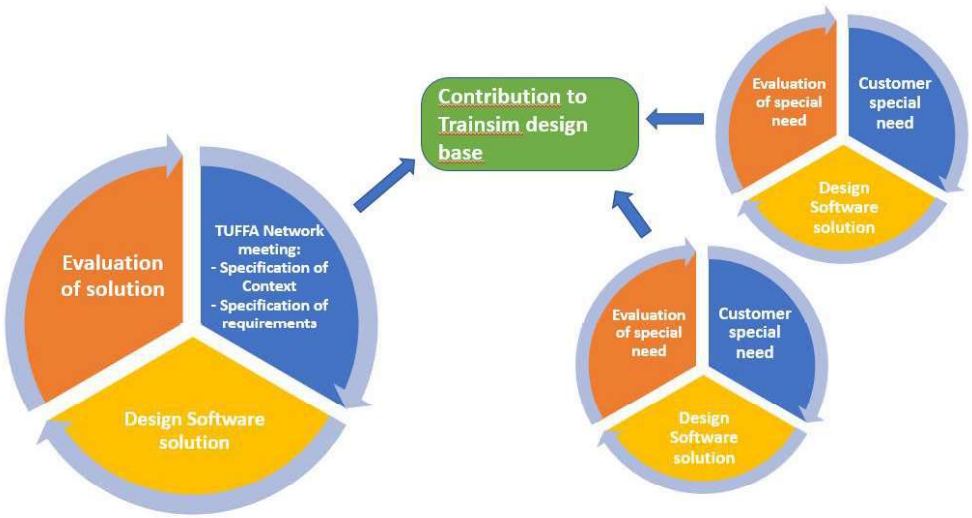


Figure 2: Multi user design development in the TUFFA train simulation network.

The train simulation software has been developed based on the 12 principles behind the Agile manifesto (Agile Alliance, 2018). For the present type of project, an agile way of working was judged to be particularly suitable. In contrast to the traditional plan-based or "waterfall" way of managing projects, the agile methodology is more suitable when the

requirements are not all well-defined beforehand, and when creativity and innovation and maximizing the value of the resulting software is top priority. A prerequisite, though, is that work must be possible to organize into iterative, short development cycles, each resulting in step-wise deliverables, that can be successively evaluated and further developed. In development of VTIs train simulators, the close contact between software developers and train educators, continuous dialogue and mutual feedback has been essential for developing a successful product and ensuring its validation. Important for VTI has been to meet the ideas, changes and deliverables from the members. Requirements have been set from VTI on the customer specific projects. New functions shall not prevent other members to use the software in the way it was intended. On the opposite, these customer specific contributions should benefit all members in the group. This has also proved a successful way to promote collaboration between competitors.

Table 1 Twelve principles of the Agile manifesto (Agile Alliance, 2018):

12 principles of the Agile manifesto:
1. Our highest priority is to satisfy the customer through early and continuous delivery of valuable software.
2. Welcome changing requirements, even late in development. Agile processes harness change for the customer's competitive advantage.
3. Deliver working software frequently, from a couple of weeks to a couple of months, with a preference to the shorter timescale.
4.Business people and developers must work together daily throughout the project.
5. Build projects around motivated individuals. Give them the environment and support they need and trust them to get the job done.
6. The most efficient and effective method of conveying information to and within a development team is face-to-face conversation.
7. Working software is the primary measure of progress.
8. Agile processes promote sustainable development. The sponsors, developers, and users should be able to maintain a constant pace indefinitely.
9. Continuous attention to technical excellence and good design enhances agility.
10. Simplicity--the art of maximizing the amount of work not done – is essential.
11. The best architectures, requirements, and designs emerge from self-organizing teams.
12. At regular intervals, the team reflects on how to become more effective, then tunes and adjusts its behavior accordingly.

3.2 Research initiatives

With VTI being a research institute in the field of transportation, the main interest is to perform transport related research. Together with the user group, relevant research projects have been formulated and applications for funding have been made also in collaboration with other partners. Several areas of interest were identified within training, capacity and energy efficiency:

- Evaluation of the in-train practical training, concerning what is trained, since fortunately difficult situations do not appear very often. However, when they appear it is important that the driver knows what to do.
- The shift to ERTMS and the effects on driving behaviour, capacity and energy efficiency.
- Exploring the possibilities to use train driving simulators to create realistic co-driver scenarios for train management students.
- Creating an interface between the train simulator software at VTI and RailML to enable drivability studies of infrastructure design and speed limits.
- Connecting existing tools for train simulations to include driving behaviour in train scheduling and capacity planning.

3.3 User survey

At a user meeting in Gothenburg, the members were asked to fill in a short survey with questions covering their experiences from the simulator and requests for the future. The questionnaire is translated into English and included in Appendix 1. A short description of the questions including both free text and Likert-type scales (Likert, 1932) follows:

The respondents represented either train operators, train educators or both and stated this as background information. They were asked to describe what they use the simulator for. The facilitation following the opportunity of using the simulator in the education was rated on a scale (from 0 = not at all to 4 = very much). The most important application was described as well as the feature thought to be most important to develop. The time of which the respondents have worked with the simulator was stated and the number of people educated or trained by it. The respondents were also asked if they save time or money with the simulator and in that case how much. If they can train specific situations, other than without the simulator, and in that case which, was also asked. Finally, the respondents rated on a scale (from 0 = no good at all to 4 = very good), how they feel about the concept of sharing resources and being part of developing the simulators.

4. Results

4.1 UCSB approach

The TUFFA user group, formed in 2015, initially consisted of two train driver educators, who primarily focused on passenger train simulators. Freight train operators joined the group once a freight train simulator model was developed in 2016. As of October, 2018, the group consists of 9 members, and a tenth is going to join in 2019. Bi-annual meetings, hosted by alternating members of the group, were used for short-listing and prioritising the desired development, which was then implemented by VTI, and the resulting software and models made available to all members, following these 6-month cycles. As a result, 4 different types of simulators are now available, all based on actual Swedish rail data, each being subjectively regarded to be of high international standard, after being validated in use by a large number of train drivers and students. The development cost for this is estimated to have been in the order of 1/10 of a hypothetical (non-existent), alternative commercial solution. Based on requests from the TUFFA user group, the common simulator software has been made available in several different physical configurations, ranging from a generic laptop PC version equipped with one or two control levers (accelerator and brake), to

authentically mimicking real driving environments a “Regina” passenger train, and one featuring a Traxx™ control panel for freight trains. Figure 3 and Figure 4 show examples of the developments requested by the users during the first two years.

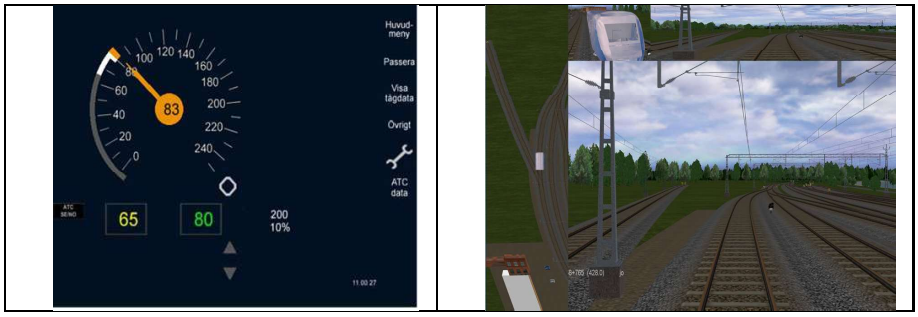


Figure 3: Example of development accomplished during the first year, STM and scenario replay (feedback)

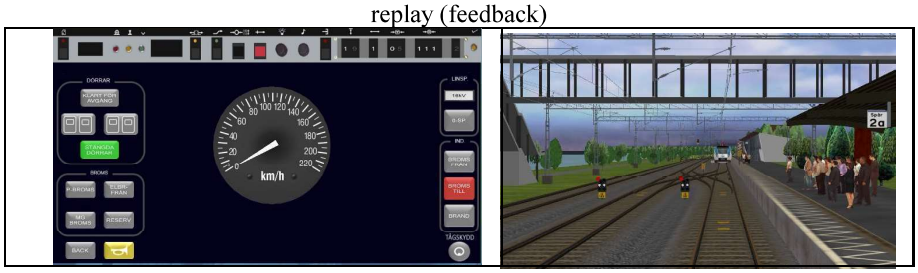


Figure 4: Example of development accomplished during the second year, DMI screen from a customer project and a shunting scenario in Jönköping.

4.2 Research initiatives

As a direct result of the requests by TUFFA users, four research projects have been formulated (and are either nationally funded or about to become funded), based on problem formulations emanating directly from the group’s needs. Three PhD projects have recently started at VTI, all of which are based on ideas generated within the TUFFA group. One thesis focus on driving behaviour studies with the ERTMS signalling system, and capacity modelling. The second PhD student will study driving models in relation to ERTMS, and the third thesis deals with pedagogical aspects of simulator-based train driver education.

4.3 User survey

The members of TUFFA all describe how collecting and sharing experiences within the group has been of great value. A survey was conducted with 8 users (3 operators, 4 educators, and 1 who is both). The survey showed the following results:

- The simulator is reportedly used for
  - introduction and training of special scenarios
  - train control systems (ATC/STM/ERTMS)
  - to transfer theoretical knowledge into practical skills
- The simulator facilitates the training or education significantly

- The most important areas are
    - training before the first practice
    - repetition of difficult cases
    - training of rarely occurring situations
    - training of special traffic safety scenarios

One respondent, a train driver educator, explicitly stated that they gain approximately 25% in time. Two train operators mentioned an (unquantified) economic gain, while another reported the gain to be in terms of quality. All 8 respondents report that they now train situations, using the simulator, that were not possible before, including "preparations before practical training, stop signals, special scenarios, ERTMS driving", and "help vehicle". This also makes the learner drivers better prepared for the on-track training.The TUFFA collaboration concept is top rated (4 out of 4) by all respondents but one, who rates it as 3 out of 4.

5. Discussion

The aims of the reported study were to explore a method to continue the development of the train simulators, using UCSD, and to create a platform for performing research on driving behaviour, from a traffic safety, efficiency and capacity perspective. The research questions are discussed one by one and followed by a method discussion.

Three years of collaboration has led to several improvements and large development steps for the simulator, in line with the user requests. This is evidence for a suitable and fruitful approach concerning development in this context. Furthermore, several research projects have received funding and are initiated together with the users, reveals that the users, although competitors in some way, have the same problems and interest in solving these together. There are many important research topics to be investigated and research is also something that the members request. They wish to learn and understand more about driver behaviour, attention and energy efficient driving for example. Fortunately, the research requested is in line with previous studies of driver behaviour and driver training conducted at VTI (Abadir Guirgis, Peters, & Lidström, 2013; Abadir Guirgis & Peters, 2015). With many members from both train driver educators and operators we will be able to design several interesting studies and collect big data sets.

When it comes to the major beneficial outcomes and for whom, we believe that there are many. With small means we have developed first class train simulators for, training, education and research. With these tools, operators can efficiently educate and train their staff before the shift to ERTMS, which will certainly have an impact on safety, capacity and efficiency. In the train driver education, quality can be improved by the possibility to create situations that are important to be skilled in, however hard to practice. Operators, passengers, and many more will benefit from safe and skilled drivers. Interestingly, most of the users do not believe that they will save neither time nor money by using the train simulators. This is in line with our idea to introduce a complementing tool and not a replacement. We also believe that the quality may be improved by this complement, which was also suggested by the users.

With a wider area of use, enabled by for example developing a Train Management system, possibilities are created to investigate the complete system of train driver, train management and traffic planning. An interface to RailML, will also open for possibilities

drivability studies which can be used for infrastructure planning. These are examples of good beneficial outcomes for the Swedish road administration.

Consistent with Norman, the needs of the interface have dominated the design in the newest portable version of the train simulator developed in collaboration with one of the operators (Norman, 1986). The development has been an iterative process and the involvement of the users has also been very significant (Karat, 1997). Gulliksen and colleagues (Gulliksen, et al., 2003), suggested principles, activity lists and tools for applying UCSD. This collaboration project has used several of the twelve principles of the Agile manifesto (Agile Alliance, 2018). The motive was purely that it seemed like the most reasonable way forward for a stepwise development in which both users and developers expanded the knowledge needed to reach a success goal. But it also required that customers understood the need and benefits of investments of resources in the product development and the need for collaboration despite being competitors. This is a unique approach which turned out applicable to a complicated context.

As mentioned in the introduction, although there are international commercial actors providing train simulators, these solutions do not meet the needs of the users within TUFFA and our research interests. The resulting open source software and the agile development approach used within TUFFA, on the other hand, leads to low costs, custom-tailored solutions for the users, and to the build-up of a sustainable and long-lasting research platform for several types of research on pedagogy, method, safety, capacity, energy-efficiency etc. for us as researchers.

6. Conclusions

In less than three years, a high-standard train simulator which is used for driver training has been developed to become a widely used educational tool, based on user needs. This was done through a user-centered collaborative scheme in the “TUFFA” user group, arranged and organized by VTI, for the benefit of railway traffic safety, but also with the co-purpose of creating a platform for further research into human behaviour. By sharing both experiences and costs for development, competing operators and education actors on the Swedish market are now equipped with train simulators tailored to their collective needs, and which they have put to immediate use in their training and education programmes. They all report how this has been beneficial in terms of quality, cost or time. Also, from a research point of view, this scheme has been very successful, resulting in several funded projects, and in three PhD projects so far. This is a major step forward since the branch needs useful tools to prepare for the shift to ERMTS, and at least in Sweden this would be very difficult to accomplish without this unique collaboration.

References

Abadir Guirgis, G., & Peters, B. (2015). *Simulatorbaserad utbildning i ERTMS-Utvärdering av utbildning och träning för lokförare i VTIs tågsimulator.* Linköping: VTI.

Abadir Guirgis, G., Peters, B., & Lidström, M. (2013). *Lokförarutbilding i Sverige-Simulatoranvändning och ERTMS.* Linköping: VTI.

Agile Alliance. (2018). 12 Principles Behind the Agile Manifesto. Hämtat från www.agilealliance.org

Andersson, A., Lidström, M., Peters, B., Rosberg, T., & Thorslund, B. (2017). *Framtagning av loktågsmodell för VTI:s tågsimulator/Development of a freight train model for the VTI train simulator.* Linköping: VTI.

Casutt, G., Theill, N., Martin, M., Keller, M., & Jäncke, L. (2014). The drive-wise project: driving simulator training increases real driving performance in healthy older drivers. *Front Aging Neurosci, 6*(85), 1663-4365.

Hedman, L.-Å. (2017). *Use of Train simulator in Train driver education.* Nässjö: Nässjöakademin.

Lees, M. N., Cosman, J. D., Lee, J. D., Rizzo, M., & Fricke, N. (2010). Translating cognitive neuroscience to the driver's operational environment: a neuroergonomic approach. *Am J Psychol, 123*(4), 391-411.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 22(140)*, 1-55.

Norman, D. (1986). Cognitive Engineering. i D. Norman, & S. Draper, *User Centered Systems Design.* Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Pollatsek, A., Vlakveld, W., Kappe, B., Pradhan, A. K., & Fisher, D. L. (2011). Driving Simulators as Training and Evaluation Tools: Novice Drivers. i D. L. Fisher, M. Rizzo, J. Caird, & J. D. Lee, *Handbook of Driving Simulation for Engineering, Medicine and Psychology.*

Svenska Byggbranschens Utvecklingsfond. (2018). *Kompetensanalys järnväg i Sverige till 2025.*

Thorslund, B. (2013). Cognitive workload and driving behavior in persons with hearing loss. *Transportation Research Part F: Traffic Psychology and Behaviour, 21*, 113-121.

Appendix 1 Survey on user experiences of the train simulator

1. Do you represent a train operator? ☐			or a train academy? ☐	
2. For what do you use the simulator?				
3. To what extent does the possibility to use the simulator facilitate the education or training?				
Not at all				Very much
☐	☐	☐	☐	☐
4. What is the most important use you think?				
5. What is most important to develop next?				
6. For how long have you been working with train simulators?				
7. How many have you trained with the help of the simulator up to now?	in basic education?		in further education?	
8. Do you save time using the simulator?	Yes ☐		No ☐	
9. If yes, estimate how much (estimate in%)				
10. Do you make a profit by using the simulator?	Yes ☐		No ☐	
11. If yes, estimate how much (estimate in%)				
12. Can you train moments in the simulator that you could not do earlier?	Yes ☐		No ☐	
13. If so, which ones?				
14. What do you think of the TUFFA concept to share costs?				
Not good at all				Very good
☐	☐	☐	☐	☐

Proactive Dispatching of Railway Operation

Markus Tideman, M.Sc. [1], Prof. Dr.-Ing. Ullrich Martin, Dr.-Ing. Weiting Zhao
Institute of Railway and Transportation Engineering (IEV) at the University of Stuttgart
Pfaffenwaldring 7, 70569 Stuttgart, Germany
[1] E-mail: markus.tideman@ievvwi.uni-stuttgart.de, Phone: +49 (0) 711 685 66 540

Abstract
Railway networks are often operated close to their full capacity due to limited infrastructure expansion and increasing traffic demand. Hence, basic timetables are fairly vulnerable to random operational disturbances. In consequence of this, the service level for passengers decreases through a combination of delay propagation and delay accumulation. To solve this problem, a possibility widely used in research is to add extensive recovery and buffer times. Nevertheless, the resulting robust basic timetables would lead to a deterioration of the operating capacity, especially in congested areas. Another approach to reduce the impact of operational disturbances on railway operation is to use conventional dispatching algorithms. Unfortunately, most of them ignore further potential disturbances during the dispatching process, which is why the generated dispatching solution might even worsen train's punctuality.

In this context, at the Institute of Railway and Transportation Engineering (IEV) at the University of Stuttgart a proactive dispatching algorithm has been developed, that generates dispatching solutions under consideration of random disturbances in dynamic circumstances. The algorithm is divided into two main processes. First, the block sections are classified depending on their specific operational risk index by simulating numerous timetables with random disturbances generated in a Monte Carlo scheme and the related negative impacts in the studied railway network are calculated. Second, near-optimal dispatching solutions are automatically generated based on Tabu Search algorithm. This is achieved within a rolling time horizon framework, taking risk-oriented random disturbances in each block section into account.

Keywords
Disturbance management, proactive dispatching, punctuality, capacity research, vulnerability of block sections

1 Introduction and State of the Art

Railway networks are often operated close to their full capacity due to limited infrastructure expansion and increasing traffic demand. Hence, basic timetables are fairly vulnerable to random operational disturbances. In consequence of those endogenous and exogenous disturbances, the service level for passengers decreases through a combination of delay propagation and delay accumulation. To solve this problem, a possibility widely used in research is to add extensive recovery and buffer times (Anderson et al. (2013), Huisman et al. (2007), Kroon et al. (2008), Lindfeldt (2015)). Nevertheless, there are two aspects that restrict the application of this strategy. On the one hand, the resulting robust basic timetables would lead to a deterioration of the operating capacity and to larger travel times, too. On the other hand, there is especially in congested areas no or only severely limited leeway for

additional time reserves. Another approach to reduce the impact of operational disturbances on railway operation is to use conventional dispatching algorithms (Bidot et al. (2006), Cheng (1998), D'Ariano (2008), Espinosa-Aranda and García-Ródenas (2012), Quaglietta et al. (2013)). Unfortunately, most of them ignore further potential disturbances during the dispatching process, which is why the generated rescheduled timetable might even worsen train's punctuality as a result of non-implementable dispatching solutions (Zhao et al. (2017)).

Extensive research activities have been conducted by focusing on the strategies stated above. Regarding this, reference is made to Zhao (2017) for a more detailed literature evaluation. In this context, at the Institute of Railway and Transportation Engineering at the University of Stuttgart an innovative dispatching algorithm has been developed based on several research activities. For instance, Cui et al. (2017a) and Liang et al. (2017) both investigated systematically the influence of dispatching on the relationship between capacity and operation quality. In Cui et al. (2017b), a method not only to prevent, but also to avoid deadlocks in synchronous simulation for railway planning and operations has been developed. Furthermore, Martin et al. (2015) studied, which influence selected disposition parameters have on the result of operational capacity researches.

The proposed dispatching algorithm bases on an operational risk analysis for each block section of the studied railway network, which make it, inter alia, stand out clearly from approaches that uses rolling time horizon frameworks (e.g. Zhan et al. (2016)). Hereby, especially railway infrastructure managers will be capable of generating robust dispatching solutions during operation stage as well as optimising the basic timetable during planning stage. For the former, the algorithm forecasts on the basis of the risk analysis the negative impacts caused by stochastic disturbances occurring in each block section and takes the severe ones more seriously during the generation of dispatching solutions in advance (Tideman et al. (2018)). For the latter, the proposed proactive dispatching approach points out potential for improvement not only of the operating program but also of the existing infrastructure. Based on a classification of the block sections according to their operational risk index, users are even able to prioritise construction measures.

2 Functionality

The above-mentioned advantages of the proposed proactive dispatching algorithm are achieved by a combination of two main processes. The first is to determine the previously alluded operational risk index of each block section of the investigated railway network in offline mode. The second process has the function to automatically generate appropriate dispatching solutions in dynamic circumstances under consideration of the operational risk classification as well as further random disturbances in online mode during railway operation.

2.1 Operational Risk Analysis

To analyse the operational risk of the network's block sections as shown in Figure 1, numerous timetable variants are simulated with the aid of RailSys® software. Due to the use of RailSys® software, a high user comfort can be ensured. As a basic requirement, the infrastructure has to be divided into appropriate sections, such as, for example, block sections.

Starting the procedure of the first algorithm process, a block section is chosen as the so-

called target block section, for which a set of disturbed timetables are generated. Those timetable variants are obtained by artificially imposing random disturbances generated in a Monte Carlo scheme and only occurring on the target block section based on an appropriate disturbance distribution. For instance, the negative exponential or the Erlang distribution could be adapted depending on the specific case of application (Zhao et al. (2017)). After running the simulation for each so-called disturbance scenario within RailSys® software,

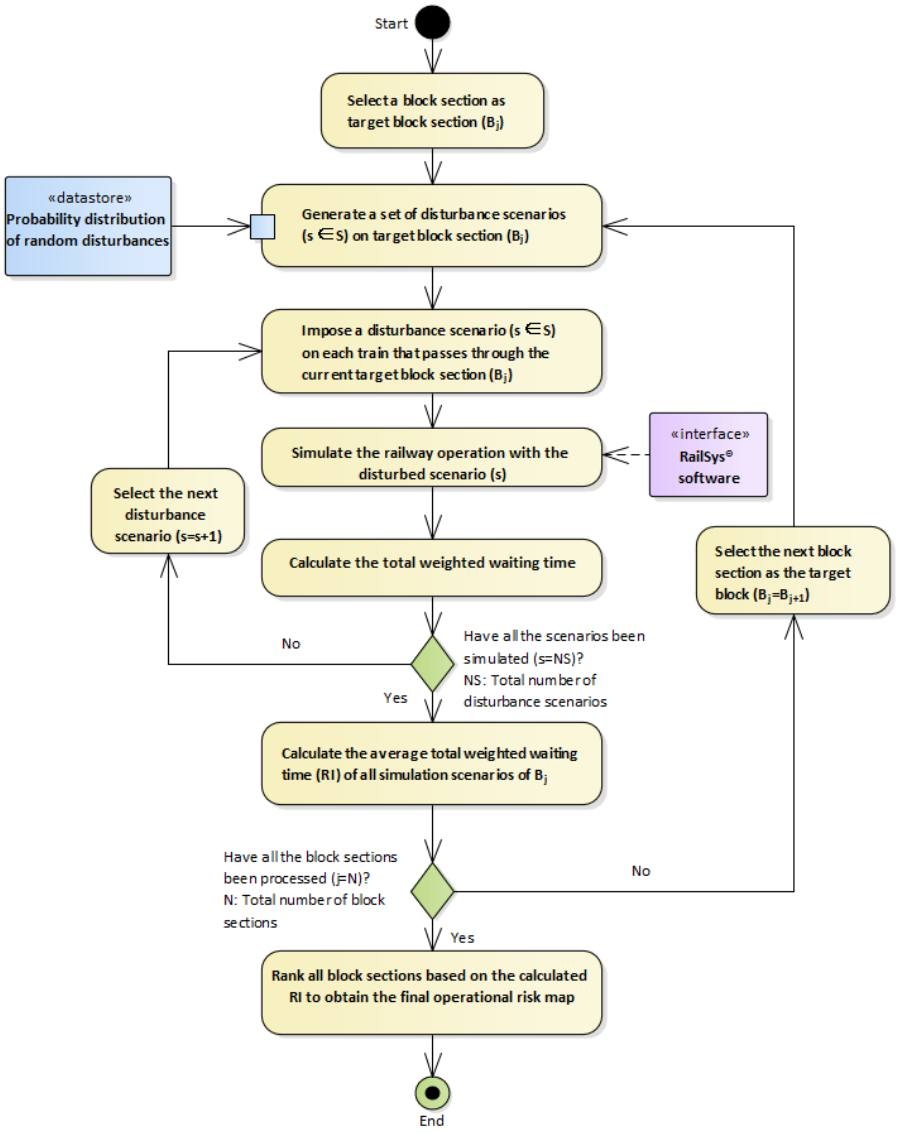


Figure 1: Workflow of the operational risk analysis

the related mean value of the target function, e. g. total weighted waiting time in the whole studied network, can be calculated. Repeating this for every block section, they can be classified according to the obtained mean values, which represent the operational risk indexes. Finally, to facilitate the further activities regarding the second algorithm process, an operational risk map can be drawn. As long as no significant changes both on infrastructure and operating program take place, the classification of the block section can be maintained.

2.2 Generation and Implementation of Dispatching Solutions

Once the first process of the algorithm has been performed, the second process can be executed as shown in Figure 2.

Initially in Task 1, the investigation period is divided into multiple time periods named "dispatching horizon", which are partially overlapping and spaced at fixed time intervals named "dispatching interval". This stage division is depicted in Figure 3 and represents the foundation of a rolling time horizon framework. With the beginning of every new dispatching interval the risk-oriented conflict detection is performed within the prediction horizon of the corresponding stage and with the aid of RailSys® software (Task 2). Hereby, the blocking times of the operating trains are artificially disturbed according to the respective risk index of each passed block section, by what the generated rescheduled timetable should be more robust against potential disturbances. If the algorithm doesn't detect overlaps between the blocking times, no further action will be arranged and the basic timetable will be maintained until the beginning of the next dispatching interval as a time-driven procedure (Task 4). In case of detecting one or more blocking time conflicts, a near-

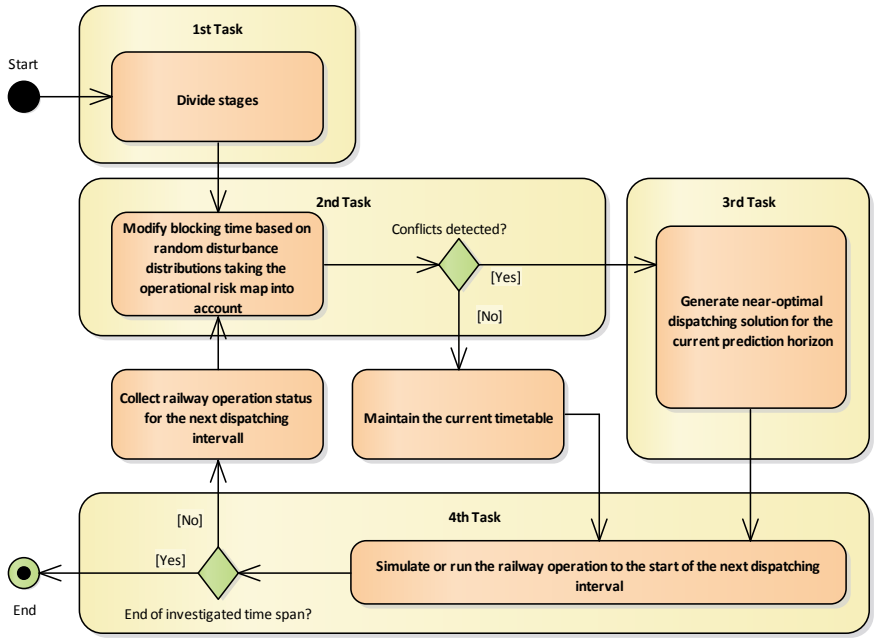


Figure 2: Simplified procedure of the dispatching solutions generation

optimal dispatching solution for the current prediction horizon will be automatically generated (Task 3). Regarding this, two cases must be distinguished. If the value of the target function, e.g. total weighted waiting time, falls below a user-defined limit value, the algorithm will solve the detected blocking time conflicts by retiming the train runs. By contrast, the conflicts will be solved by reordering of the train runs based on Tabu Search algorithm, whenever the value of the target function exceeds the limit value. Subsequently, the resulting rescheduled timetable contains retimed or reordered train movements and will be used for the railway operation unto the beginning of the next dispatching interval (Task 4).

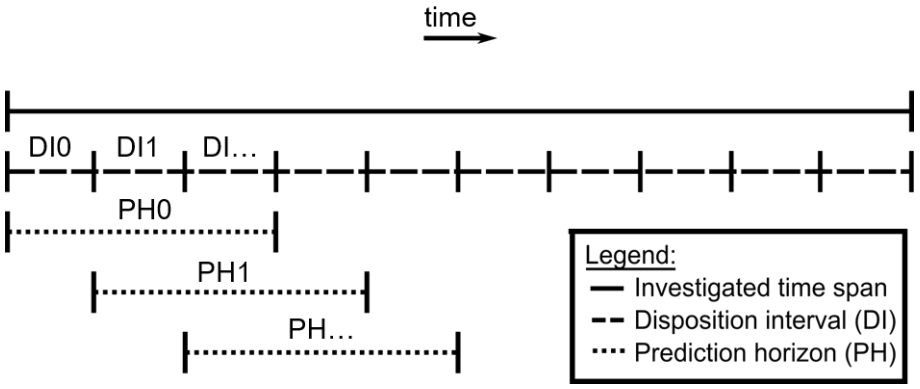


Figure 3: Schematic representation of the rolling time horizon framework

3 Reference Scenario

The development of the proposed proactive dispatching algorithm has taken place based on a realistically reference example, what is standardly used for algorithm development at the Institute of Railway and Transportation Engineering at the University of Stuttgart. The reference model is emulated in RailSys® software, which assists to generate effective and realistic dispatching solutions. This railway network is shown in Figure 4, contains over 43 kilometre track length and includes in total up to 72 long-distance passenger transport, regional passenger transport and freight transport train runs within a time span of six hours.

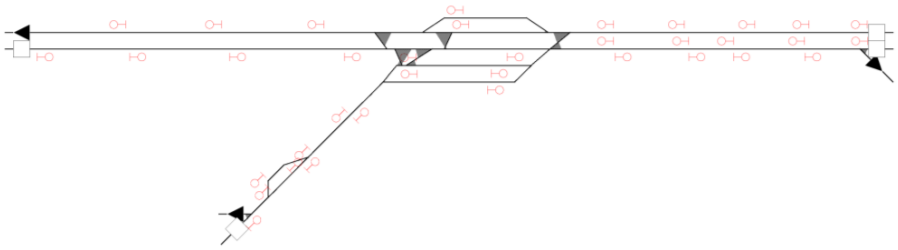


Figure 4: Track layout

Furthermore, the reference model ensures that the proposed dispatching algorithm is capable of handling various types of conflicts such as crossing, following, merging and opposing conflicts.

Technically, the operational risk analysis (first process) as well as the dispatching algorithm (second process) is developed in Microsoft Visual Studio 2015 environment with C#. This code runs the simulations in RailSys® software automatically. Due to the large amount of disturbance scenarios that has to be imposed on every block section separately, the operational risk analysis has to be executed in an offline environment. For the reference scenario this process takes at least four hours (Fujitsu computer, Intel Core i5-4670 CPU @ 3.40 GHz, 8 GB RAM). After achieving the operational risk classification once, the classification will remain its validity in the future as long as there won't occur changes in infrastructure layout or operating program. The code for the dispatching algorithm, which is also written with C#, has to be started manually by the user. With the objective of using this second process in online mode, the calculation of both conflict detection and conflict resolution takes about 20 seconds for each stage.

3.1 Operational Risk Analysis

Regarding the first algorithm process, the railway infrastructure is divided into 35 block sections. Then, a significant amount of disturbed timetables is generated, which consider only within the target block section entry delay, departure time extension, running time extension and/or dwell time extension depending on the characteristics of the target block section, e.g. existing stations. On the supposition that 1 denotes the lowest and 5 the highest risk level and by following the above stated algorithm procedure, it can be obtained that seven block sections belong to each risk level, as it is shown in Figure 5. Based on this, the blocking time of each train passing the respective block section are prolonged for the conflict detection in Task 2 of the second algorithm process.

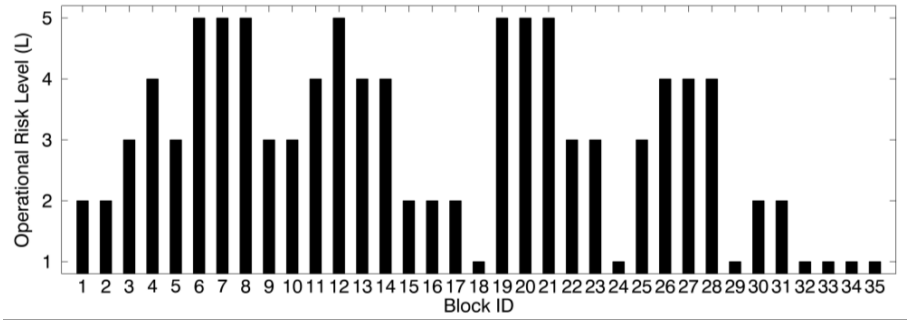


Figure 5: Operational risk levels of all 35 block sections of the reference scenario

3.2 Generation and Implementation of Dispatching Solutions

According to Figure 3, in Task 1 of the second algorithm process the six hours investigation period of the reference example is divided into twelve dispatching horizons with a length of two hours and twelve dispatching intervals (DI-0, DI-1, …, DI-11) with a length of 30 minutes. Based on the results of the operational risk analysis, in Task 2 the scheduled

blocking times of the trains are modified at the beginning of a dispatching interval in accordance to the specific risk level of the block sections the trains are operating in. This is done by adding extra time to the scheduled blocking times with the aid of imposed disturbances. As an example for high-speed trains, in Table 1 the risk-oriented disturbances are listed broken down for the five different operational risk levels.

Table 1: Risk-oriented disturbances imposed on high-speed trains for conflict detection

Operational risk level:	**1**	**2**	**3**	**4**	**5**
Dwell time extension [s]:	3	12.6	23.4	37.2	55.2
Departure time extension [s]:	3	12.6	23.4	37.2	55.2
Entry delay [s]:	15.6	62.4	118.2	186	274.8
Running time extension [s]:	3	12.6	23.4	37.2	55.2

Also in Task 2, the conflict detection for each time stage is performed after the modification of the blocking times has taken place. In doing so for the reference case, it turns out that during seven dispatching intervals no further dispatching actions are necessary and the current timetables can be maintained. For the remaining dispatching intervals the 3rd task has to be executed. Here, the near-optimal dispatching solution for the current prediction horizon is generated. In DI-3 and DI-9 it is sufficient to retime the train runs, whereas for the three remaining dispatching intervals DI-0, DI-4 and DI-8 a new order of the train runs has to be generated by Tabu search algorithm (see Table 2). Irrespective of whether the initial timetable or a generated dispatching timetable are used hereinafter, the 4th task deals as a proxy of the real world railway operation.

Table 2: Dispatching solutions for each dispatching interval (DI)

	Start time	**End time**	**Dispatching solution**
DI-0	0:00	0:30	Reordering
DI-1	0:30	1:00	-
DI-2	1:00	1:30	-
DI-3	1:30	2:00	Retiming
DI-4	2:00	2:30	Reordering
DI-5	2:30	3:00	-
DI-6	3:00	3:30	-
DI-7	3:30	4:00	-
DI-8	4:00	4:30	Reordering
DI-9	4:30	5:00	Retiming
DI-10	5:00	5:30	-
DI-11	5:30	6:00	-

3.3 Comparison with FCFS-Principle

To underline the advantage of the proposed proactive dispatching algorithm and to ensure its effectiveness, the calculated total weighted waiting time of each stage is compared with the total weighted waiting time that results by using first come – first serve (FCFS) rule for the dispatching process. As depicted in Figure 5, it can be seen easily that for both dispatching strategies in sum the curves decrease and that for every stage the proposed

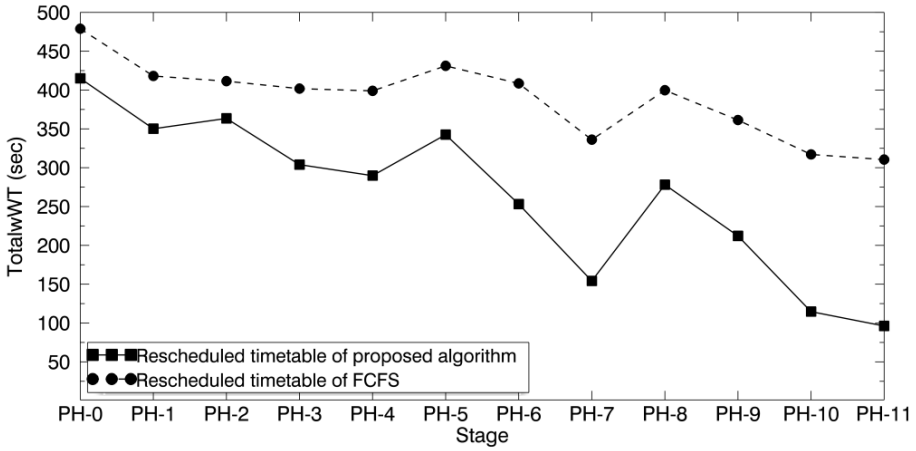


Figure 6: Comparison of the results of the proposed algorithm and FCFS principle

algorithm enables a significantly better operating quality expressed by a higher overall punctuality of the railway network. Regarding this, the two final values differ by ca. 215 seconds.

4 Conclusion

Heavily summarized, the main innovation or rather benefits of the developed algorithm are the automatic calculation of the operational risk of every block section of any investigated railway network in offline mode, the consideration of the operational risk during the automatic generation of dispatching solutions in online mode and the algorithm's sustainable impact on the operation quality due to the implementation of a rolling time horizon framework.

Furthermore, as explained in section 3, the proposed proactive dispatching algorithm is able to solve crossing, following, merging and opposing conflicts. This happens in a very effective manner, mainly because the dispatching algorithm takes the operational risk index for each appropriate infrastructure segment into account. Moreover, the proposed method bases on a rolling time horizon framework, by what the performance of the generated dispatching solutions is evaluated after a certain time span.

Additionally, the reference model enhances the extensibility of the algorithm to large railway networks. Not least because of this, one of the current research activities of the IEV (Martin et al. (2018)) deals, inter alia, with the application of the discussed proactive dispatching approach within a real railway network in Germany for manifesting its practical use.

Also, the presented dispatching approach isn't restricted to the field of railway operation, which is why the IEV also investigates the algorithm's usability in the context of other transportation systems, such as aviation (Tideman and Martin (2018)).

References

Andersson, E.V., Peterson, A., Törnquist, J., 2013: "Quantifying railway timetable robustness in critical points", In: *Journal of Rail Transport Planning & Management*, 3, vol. 3, pp. 95–110.

Bidot, J., Laborie, P., Beck, J.C., Vidal, T., 2006: "Using constraint programming and simulation for execution monitoring and progressive scheduling", In: *Proceedings of the 12th IFAC Symposium on Information Control Problems in Manufacturing*, Saint-Étienne, France, pp. 595–600.

Cheng, Y., 1998: "Hybrid simulation for resolving resource conflicts in train traffic rescheduling", In: *Computers in Industry*, 3, vol. 35, pp. 233–246.

Cui, Y., Martin, U., Liang, J., 2017a: "Decentralized, Autonomous Train Dispatching using Swarm Intelligence in Railway Operations and Control", In: *Proceedings of The 7th International Seminar on Railway Operations Modelling and Analysis (RailLille2017)*, Lille, France.

Cui, Y., Martin, U., Liang, J., 2017b: "Searching feasible resources to reduce false-positive situations for resolving deadlocks with the Banker's algorithm in railway simulation", In: *Journal of Rail Transport Planning & Management*, 7 (1-2), pp. 50–61. DOI: 10.1016/j.jrtpm.2017.05.001.

D'Ariano, A., 2008: *Improving real-time train dispatching: models, algorithms and applications*, Dissertation, TU Delft, Delft, 2008.

Espinosa-Aranda, J.L., García-Ródenas, R., 2012: "A discrete event-based simulation model for real-time traffic management in railways", In: Journal of Intelligent Transportation Systems, 2, vol. 16, pp. 94–107.

Huisman, D., Kroon, L., Maróti, G., 2007: "Railway timetabling from an operations research perspective", In: *Econometric Institute report EI*, vol. 22, Econometric Institute, Rotterdam.

Liang, J., Martin, U., Cui, Y., 2017: "Increasing performance of railway systems by exploitation of the relationship between capacity and operation quality", In: *Journal of Rail Transport Planning & Management*, 7 (3), pp. 127–140. DOI: 10.1016/j.jrtpm.2017.08.002.

Kroon, L., Maróti, G., Helmrich, M.R., Vromans, M., Dekker, R., 2008: "Stochastic improvement of cyclic railway timetables", In: *Transportation Research Part B: Methodological*, 6, vol. 42, pp. 553–570.

Lindfeldt, A., 2015: *Railway capacity analysis. Methods for simulation and evaluation of timetables, delays and infrastructure*, KTH Royal Institute of Technology, School of Architecture and the Built Environment, Department of Transport Science, Stockholm, Sweden.

Martin, U., Liang, J., Cui, Y., 2015: „Einfluss von ausgewählten Dispositionsparametern auf das Ergebnis von Leistungsuntersuchungen", In: *ETR-Eisenbahntechnische Rundschau*, 7+8, vol. 64, pp. 20-23.

Martin, U., Tideman, M., Zhao, W., 2018: *Risk Oriented Dispatching of Railway Operation under the Consideration of Random Disturbances in Dynamic Circumstances (DICORD)*. DFG-project (MA 2326/22-1) – in progress. Institute of Railway and Transportation Engineering, Stuttgart.

Quaglietta, E., Corman, F., Goverde, R.M., 2013: "Stability of railway dispatching solutions under a stochastic and dynamic environment", In: *Journal of Rail Transport Planning & Management*, vol. 3, pp. 137–149.

Tideman, M., Martin, U., 2018: „Proaktive Disposition luftverkehrlicher Prozesse“, In: *Internationales Verkehrswesen*, 4, vol. 70, pp. 60-63.

Tideman, M., Martin, U., Zhao, W., 2018: „Disposition von verkehrlichen Prozessen unter Einbeziehung von zufallsbedingten Unsicherheiten“, In: *ETR-Eisenbahntechnische Rundschau*, 10, vol. 67, pp. 22-25.

Zhan, S., Kroon, L., Zhao, J., Peng, Q., 2016: „A rolling horizon approach to the high speed train rescheduling problem in case of a partial segment blockage“, In: *Transportation Research Part E: Logistics and Transportation Review*, vol. 95, pp. 32-61. DOI: 10.1016/j.tre.2016.07.015.

Zhao, W., 2017: *Hybrid Model for Proactive Dispatching of Railway Operation under the Consideration of Random Disturbances in Dynamic Circumstances*, vol. 22, Books on Demand, Norderstedt. Neues verkehrswissenschaftliches Journal.

Zhao, W., Martin, U., Cui, Y., Liang, J., 2017: “Operational risk analysis of block sections in the railway network”, In: *Journal of Rail Transport Planning & Management*, 7 (4), pp. 245–262. DOI: 10.1016/j.jrtpm.2017.09.003.

A Conflict Prevention Strategy for Large and Complex Networks in Real-Time Railway Traffic Management

Pieter Vansteenwegen [a,1], Sofie Van Thielen [a], Francesco Corman [b]
[a] KU Leuven Mobility Research Centre-CIB, KU Leuven
Celestijnenlaan 300, 3001 Leuven, Belgium
[1] E-mail: pieter.vansteenwegen@kuleuven.be, Phone: +32 (0) 16 32 16 69
[b] Dep. Of Civil, Env. And Geomatic Eng., ETH Zürich
Stefano-Franscini-Platz 5, 8093 Zürich, Switzerland

Abstract
Train timetables are built such that trains can drive without any delay. However, in real-time, unexpected events such as overcrowded platforms or small mechanical defects can cause conflicts, i.e., two trains requiring the same part of the infrastructure at the same time. Currently, such conflicts are typically resolved by experienced dispatchers. However, it is impossible for them to fully anticipate the impact of their actions on the entire network. Conflict detection and prevention tools embedded in a Traffic Management System can help them in making informed decisions. Though some advanced train movement prediction and conflict detection has been developed in the last years, there still exists a need for conflict prevention strategies capable of delivering conflict resolutions on large and complex networks based on retiming, reordering and rerouting some of the trains in real-time.
Our previous work introduced such a conflict prevention strategy that, based on offline calculations, determined which part of the network should be regarded when deciding on a conflict resolution. This work is significantly extended here by considering several new parameters for the Dynamic Impact Zone heuristic. This paper compares results on different sizes of networks, and tackles the challenges for applying the strategy on even larger networks.

Keywords
Conflict Resolution, Real-Time Railway Management, Dispatching, Large Networks

1 Introduction

Transport and mobility are important for inhabitants all over the world. Rail transport is used by many passengers to travel to their work on a daily basis. Clearly, the need for more people and goods mobility has steadily grown worldwide and this trend will continue in the future. Therefore, public transport systems will need to provide a better quality of service, in terms of frequencies, comfort, accessibility and reliability of services, along with transparent information regarding travel times and routing alternatives.

The combination of growth in mobility demand and the difficulties in building new infrastructure presses the need for utilizing the existing infrastructure at the highest possible capacity, at all times. A less costly solution is to improve the quality of the trains to decrease breakdowns and/or to increase the capacity by improving train punctuality.

Though train timetables can account for some delays occurring in real-time, unexpected events such as passenger crowding, bad weather or a small mechanical defect, can make the timetable infeasible. In order to increase the performance of railway services in practice, efficient conflict resolutions are required. These resolutions can be based on the retiming or reordering of trains, or even rerouting them. If the timetable becomes infeasible, dispatchers have to decide on the best resolution. Nowadays, they are often assisted by an advanced Traffic Management System (TMS), including train movement prediction and conflict detection. However, after a conflict is detected, dispatchers often still have to rely on their own experience to decide on the best conflict resolution. Therefore, a Conflict Prevention Module (CPM), which can be easily integrated into a TMS, is required. This module should include a Conflict Prevention Strategy (CPS) capable of resolving detected conflicts.

Our previous work introduced such a CPS that, based on offline calculations, determined which part of the network should be regarded when deciding on a conflict resolution. This work is significantly extended here by considering several new parameters for the Dynamic Impact Zone heuristic. This paper compares results on different sizes of networks, and tackles the challenges for applying the strategy on even larger networks.

Section 2 starts by explaining some important definitions required for the remainder of the paper and Section 3 discusses the related literature. Section 4 explains the simulation framework used for testing the CPS. The CPS and the different novelties are discussed in detail in Section 5. Section 6 shows the experimental results applying the new CPS on larger and more and complex networks. The paper is concluded in Section 7.

2 Definitions

This section describes all relevant definitions used in railway literature and in practice. First, the basic elements that build up a railway network are introduced. Then, it is defined how trains move through a railway network.

2.1 The Railway Network

A railway network can be considered on three different levels: the macroscopic, mesoscopic or microscopic level. The microscopic level includes all details, e.g., switches, tracks, signals. This level is important for the train drivers and dispatchers. The macroscopic level is a reduction of the microscopic level and is often only what passengers experience. The mesoscopic level lies somewhere in between the two previous levels. In this paper it is required that all timings of trains are known in full detail. Therefore, a microscopic level for the network is preferred.

A microscopic network is characterized by the signals present in the infrastructure. Signals give information to trains coming from the direction to which the signal is visible. Every signal indicates either the beginning or the end of a *block section*. The part of the infrastructure between two subsequent, similarly directed signals thus determine a block section, which is typically around 1000 meters long in Belgium.

The network can also be decomposed in zones: *station areas* and *non-station areas*. A station area includes one or more parallel platforms where passengers can embark or disembark if the train has a *stop* at this station. Before and after these platforms, there is a *switch area*, such that many possible combinations between an incoming or outgoing track and a platform can be made. This allows trains to reroute in the station area when their

original platform is not available. The signals at the beginning or end of a switch area, are called the start and end points of the station area.

2.2 A Train Driving Through the Network

Nowadays, in many railway systems around the world, a train drives from signal to signal. This signal gives information about the next block section the train wants to occupy. A railway signal is comparable to ordinary traffic lights and can give a green, double yellow or red light. Green indicates that the next two block sections are free, double yellow indicates that the next block section is free, but the one after that is occupied, and red indicates that the next block section is occupied. In practice, a train has to slow down at double yellow such that it can come to a full stop at a red signal, if necessary. In this way, signals guarantee that trains can be guided safely throughout the network by giving information on the state of the block sections ahead.

According to Hansen and Pachl (2008), a block section is exclusively occupied by one train during a time interval, composed of the actual *occupation time* and safety margins before and after. The time interval during which a train blocks one block section, is called the *blocking time*. By using the blocking time theory, an accurate calculation can be performed to determine the duration of the blocking times. These blocks can then be represented in a time-distance diagram and the result is the so-called *blocking stairway*. A conflict can then be seen as where blocks belonging to different trains overlap.

2.3 Methodological Framework

As indicated by Lamorgese et al. (2017), a large gap still exists between the state-of-the-art traffic management in academic research and the state-of-the-art in practice. This shows that many challenges arise when putting academic research in practice. However, lately, a lot of effort has been put in implementing academic models in practice (see for instance Borndörfer et al., 2017). In order to close the gap between academics and practice, Corman and Quaglietta (2015) introduce a closed loop framework, which is closer to real-life situations than an open or multiple open loop. Open loop rescheduling assumes that control measures are computed and implemented once and for all, thus assuming a perfect knowledge of future traffic states. This implies that predicted and actual traffic states are equal, and that no unexpected events can occur anymore. Open loop approaches have been implemented very often in academic literature (Corman and Meng, 2013; Cacchiani et al., 2014; Pellegrini et al., 2016). An extension is the multiple open loop, where it is assumed that at some points more traffic conditions are known, and the calculations can be reconsidered with this additional information (Corman and Quaglietta, 2015). A closed loop approach calculates dispatching actions every time updated information is available, and adjusts control measures immediately (Caimi et al., 2012; Corman and Quaglietta, 2015). In this setup, new updates of information are taken into account whenever available. The implementation of the Conflict Prevention Strategy (CPS) within a closed loop framework is discussed in detail in Section 4.

3 Literature Review

An overview of some recent papers dealing with the conflict resolution problem is given in Table 1. These approaches perform conflict resolution by using: an Alternative Graph (AG), a Mixed Integer Linear Program (MILP), a Mixed Integer Program (MIP), Constraint Programming (CP), etc. According to the level of detail, we can make a distinction between microscopic, mesoscopic and macroscopic level. The models can be adapted to deal with different objective functions as can be seen in the fourth column. The model is tested on a study area, that can either be a station area (S), a control area (CA), a terminal area of a metro line (TA), a network (N), a line (L) or a railway corridor (RC). Possible control actions can be updating the train timing (retiming RT) or train order (RO) or the routes (RR).

All of these approaches describe the railway problem at a given level of detail, predict future train conflicts, compute control actions to resolve these conflicts, and evaluate using some objective function. Clearly, a variety of methods with different characteristics have been proposed over the last years. We discuss Table 1 in detail in Van Thielen (2019).

For usage in practice, a conflict resolution model should be capable of giving immediate applicable suggestions in a small time frame of merely seconds, even for very large and complex networks. Microscopic conflict resolutions are therefore preferred. Corman and Quaglietta (2015) suggest that the Conflict Resolution Problem should be tested in a closed loop fashion. Though some approaches have been tested in practice (Borndörfer et al., 2017), the size of the networks considered is always very limited.

4 Simulation Framework

This section describes how the real-time railway traffic management is modeled by a closed loop framework. In practice, real-time operations are often affected by external causes, e.g., a mechanical defect. Such external causes possibly lead to a *primary delay*, i.e. a delay that cannot be avoided during planning. Once trains start deviating from their original schedule, other conflicts can arise. A conflict occurs when (at least) two trains require the same part of the infrastructure at the same time. In this case, a dispatcher or a TMS needs to resolve the conflict at once. However, resolving a conflict requires to (locally) reroute or retime/reorder one or multiple trains. One conflict resolution can therefore cause many other conflicts in the (near) future, especially when dealing with a complex network with a high number of trains. The dispatcher or the TMS creates an updated train path plan by resolving the conflict. This new plan can then be executed until a new conflict is detected and needs to be resolved.

The whole of real-time railway traffic management can thus be divided into four major modules, as depicted in Figure 1. Firstly, there is the simulator module, resembling real-life closely by describing the current traffic states. Secondly, the conflict detection module is able to predict future train movements and detect possible conflicts. Thirdly, the conflict prevention module calculates feasible conflict resolutions for every detected conflict. This conflict resolution is then given back to a dispatcher who can still decide to follow the recommendation or not. Alternatively, an automated machine can make the decisions, e.g., always following the recommendations. This is located in the fourth module, the dispatching module.

Work	Mathematical model	Detail	Objective function	Case	Implementation area	Actions	Loop
Bettinelli et al. (2017)	Iterated greedy heuristic	Macro/micro	Combination of delay penalty, dependency breaking penalty, capacity violation penalty and detour penalty	L/S	Real-world instances	RT RO RR	Open loop
Caimi et al. (2012)	AG	Micro	Combination of reliability and punctuality	S	Berne	RT RO RR RS	Closed loop
Chen et al. (2015)	MIP	Micro	Weighted avg delay	RC	Core area of Thameslink route in London	RT RO RR	Open loop
Corman et al. (2012a)	AG	Micro	Min max sec delay	CA	Utrecht-Den Bosch	RT RO RR	Open loop
Corman et al. (2012b)	AG	Micro	Min train delays and missed connections	S	Utrecht	RT RO	Open loop
Corman and Quaglietta (2015)	AG	Micro	Min max sec delay	RC	Utrecht-Den Bosch	RT RO	Open loop, multiple open loop, closed loop
D'Ariano et al. (2008)	AG	Micro	Min max sec delay	CA	Utrecht-Den Bosch	RT RO	Open loop
D'Ariano et al. (2014)	AG/ MILP	Micro	Min max sec delay	L	East Cost Main line	RT RO RR	Open loop
Dolder et al. (2009)	Timed event graph	Micro	None (delay propagation)	N	Intercity network, Germany	RT	Closed loop
Fischetti and Monaci (2017)	AG/ MILP	Micro	Avg final delay	L	railway line nearby London	RT RO (RR)	Open loop
Josyula et al. (2018)	MILP/ Binary tree search	Micro	Min total final delay	L	Karlskrona - Tjörnarp	RT RO RR	Multiple open loop
Kecman et al. (2013)	AG	Macro/ meso	Min max sec delay	RC N	Utrecht-Den Bosch Dutch railway network	RT RO	Open loop
Lamorgese et al. (2016)	MILP	Micro/macro	Min delay cost	L	Trento-Bassano del Grappa, Foligno-Orte, Foligno-Falconara	RT RO RR	Open loop
Mannino and Mascis (2009)	MILP/ AG	Macro	Weighted deviation	TA	Milan (Sesto)	RT RO RR	Closed loop
Meng and Zhou (2011)	IP	Meso	Min total completion time of all involved trains	RC	Academic instances	RT RO RR	Open loop
Pellegrini et al. (2016)	MILP	Micro	Min total sec delay	CA	Mantes-La-Jolie junction, Rouen-Rive-Droite	RT RO RR	Open loop
Quaglietta et al. (2016)	AG	Micro	Min max sec delay	RC	Utrecht-Den Bosch	RT RO	
Rodriguez, J. (2007(@)	CP	Micro	Min max sec delay	CA	Pierrefitte-Gonesse junction	RT RO RR	Open loop
Samà et al. (2015)	AG/ MILP	Micro	Min max sec delay, min sum sec delay, min sum final delay, max punctuality, min sum weighted delays (with threshold), min deviations, min sum arrival times, min travel time trains	CA	Utrecht-Den Bosch	RT RO	Open loop
Samà et al. (2016)	MILP	Micro	Min total sec delay	CA	Rouen-Rive Droite, Lille-Flandres	RT RO RR	Open loop
Törnquist (2007)	MILP	Macro	Min total final delay, min total accumulated delay, min total delay cost	CA	Nörrkoping	RT RO	Open loop
Törnquist and Persson (2007)	MILP	Macro	Min total final delay, min total delay cost	CA	South Traffic District	RT RO	Open loop
Törnquist Krasemann (2012)	MILP	Meso	Min total final delay, min total delay cost	CA	Nörrkoping	RT RO	Open loop
Van Thielen et al. (2017)	Binary search tree	Micro	Min sec delay	CA	Brugge-Gent-Denderleeuw	RT RO RR	Closed loop
Van Thielen et al. (2018)	Data driven heuristic	Micro	Min sec delay	CA	Brugge-Gent-Denderleeuw	RT RO RR	Closed loop

Table 1: An overview of academic models for real-time railway traffic management discussed

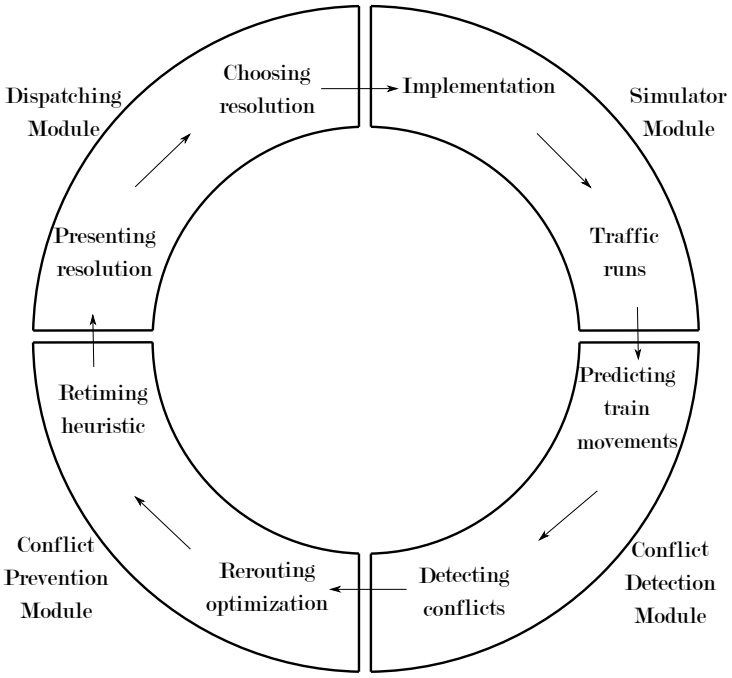


Figure 1: A schematic representation of the framework of a closed loop approach.

These different modules intercommunicate. The simulator module determines all the current states of all trains and all infrastructure parts. This module resembles reality in the sense that it takes into account stochastic dynamics of delays, and progresses synchronously in time. The simulator module starts at a given start hour, for example at 7 a.m. It will evaluate the performance of the conflict prevention module during the *simulation horizon*, which is set to 60 minutes. At the end of the simulation, several evaluation criteria such as the total secondary delay and total passenger delay of all the running trains is calculated. This is the result of many iterations of conflict detection, optimization and implementation of solution measures, in a closed loop fashion. As in real-life, it is assumed that the train information is sensed and immediately adjusted in the simulator module. It gives the necessary information on whereabouts of trains to the conflict detection module. The conflict detection module predicts and detects a certain *prediction horizon* ahead. In this paper, the prediction horizon is limited to 5 minutes because of how challenging and computationally intensive the conflict detection module is. Evidently, a TMS in practice will be capable of detecting conflicts further ahead, giving more time for calculating a good conflict resolution.

If a conflict is detected within the prediction horizon, it is sent to the conflict prevention module, where a suitable resolution is calculated. This resolution can either be based on rerouting in station areas or on retiming/reordering one of the trains. The goal is to find a resolution by minimizing a performance indicator, e.g., the total secondary delay. The objective needs to be determined beforehand by the railway infrastructure manager and/or operators. If the conflict takes place in a station area, the rerouting optimization tries to find a conflict-free solution first. If the rerouting optimization finds a feasible solution lowering the

secondary delay in the station area, and results in alternative routes, then the prediction will be adapted based on the best feasible (or optimal) solution. The advantage of using rerouting is that no retiming/reordering is required and thus no additional delays are imposed. In case that the rerouting does not resolve the conflict or if the conflict is not located in a station area, the conflict needs to be resolved by finding a retiming/reordering solution. Retiming/reordering is optimized through the DIZ heuristic leading to (further) delaying one of the conflicting trains, and/or altering the order of the trains. The computed solution is then implemented by locally rerouting trains and/or changing the time and order of the trains involved in the conflict. The prediction is adapted based on this solution, possibly leading to new conflicts. This process is repeated until no conflicts are found within the prediction horizon.

This conflict resolution induces an updated train plan, including new routes and/or new orders or timings of trains. The presentation and acceptance of this resolution is dependent on the dispatching module, which can either be a dispatcher or an automated machine accepting all proposed resolutions. This dispatching module allows a dispatcher to follow the presented conflict resolution or to implement another resolution based on his/her own experience. The resolution is then sent back to the simulated reality where it can be implemented. The train plans can then be adapted according to the resolution. This will be taken into account in the conflict detection module. Since we are evaluating the performance of the conflict prevention module in this paper, we assume that the dispatching module always accepts the proposed resolutions.

These modules all have a certain computation time, because they are all in real-time. Therefore, a very advanced conflict detection module is required delivering new predictions every two seconds (or less) (Dolder et al., 2009). Also, the conflict prevention module needs to make fast decisions regarding conflict resolutions. Whenever a conflict is detected in real-time, the time required for finding a resolution and the time required for the dispatching module to accept the solution has to be taken into account. Accordingly, changes to the current situation during the calculation should not cause issues when implementing the calculated resolution. Therefore, both the rerouting optimization and the DIZ heuristic start their calculations from the expected situation a control delay after the moment of detecting the conflict. Stated otherwise, no changes to ongoing operations can be executed during the duration of the control delay. Whatever happens within this time interval after detecting the conflict, cannot be changed by the conflict prevention module. Actually, this control delay gives the conflict prevention module this exact time to calculate (and implement) a resolution.

5 Methodology

This section starts by describing our previous approach in Section 5.1. Additional improvements and extensions are then introduced in Section 5.2.

5.1 Previous Approach

The CPS consists of, on the one hand, a rerouting optimization based on a flexible job-shop problem and, on the other hand, a retiming/reordering heuristic. This heuristic examines the progress of all relevant trains for up to two options to resolve a conflict: delaying the first or second train. The first train is the first train arriving at the block section where the

conflict occurs. To be useful in real-time, the progress examination should be limited in time and space in order to limit the computation time. Therefore, based on a trade-off between quality of the solution and the required computation time, a dynamic impact zone in which the progress is evaluated, is created for every conflict.

The conflict detected by TMS and sent to the CPS, is called the *initial conflict*. If this conflict is located in a station area, a rerouting optimization is started to check whether rerouting (some of the) trains reduces the overall delay. The station area is cut out of the network. For this limited network, the optimization based on a flexible job shop problem is started at the detection time plus an additional control delay, and ends when both trains have left the station area. The problem is given to Cplex with a maximum computation time in order to keep this time limited. Afterwards, alternative routes from the best feasible or optimal solution are implemented. For more detailed information on the rerouting optimization, we refer the interested reader to Van Thielen et al. (2018).

If the conflict is not resolved by the rerouting optimization, or if the conflict is not located in a station area, a resolution based on the Dynamic Impact Zone (DIZ) heuristic has to be found. This DIZ heuristic starts by selecting a suitable dynamic impact zone for the conflict. A dynamic impact zone determines which conflicts in the near future are (possibly) affected by the conflict resolution of the initial conflict. Only the trains in these conflicts should be considered during the progress examination such that the computation time remains limited, even for large networks.

The dynamic impact zone starts by considering all potential conflicts during the next half hour. An example network is shown in Figure 2, where every potential conflict is indicated with a figure. The initial conflict is depicted by a square. Conflicts can be divided into groups by considering their relation to the initial conflict. In this way, a conflict is called a *first-order conflict* if one of the trains in this conflict is also in the initial conflict. In case of the example, this means that any conflict involving T_1 or T_2 is a *first-order conflict*, and is depicted as a circle in Figure 2. *Second-order* conflicts are conflicts where at least one of the trains is involved in a first-order conflict, but it is not a first-order conflict itself. In this manner, an nth-order conflict is a conflict of which at least one train is in an $(n-1)$th-order conflict, but it is not a $(n-1)$th-order conflict itself. Second-order conflicts are depicted as diamond shapes, third-order conflicts as triangle shapes in Figure 2.

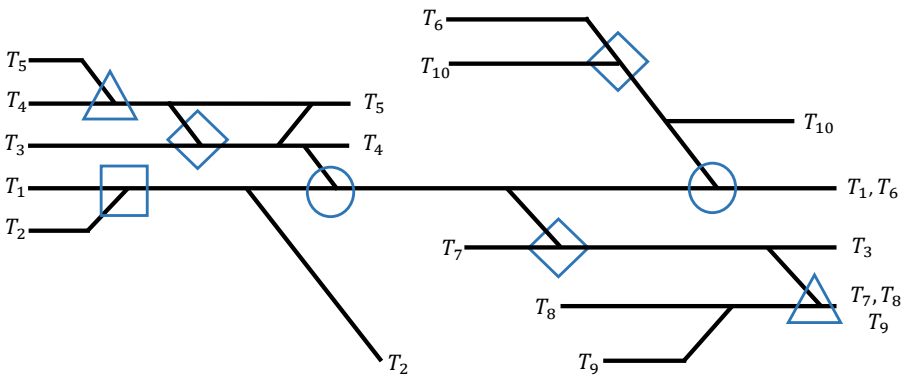


Figure 2: All potential conflicts in an example area.

As shown in Van Thielen et al. (2018), the size of the dynamic impact zone affects the computation time strongly. Therefore, the dynamic impact zone should not be too large to keep the computation time limited, but still large enough to determine a suitable solution. Therefore, offline calculations are carried out to determine which conflicts are *most likely* conflicts. After resolving 350 randomly created delay scenarios using 6 different conflict resolution strategies, the *most likely* conflicts are determined as the conflicts occurring in at least 50 % of all cases. All *most likely* conflicts are indicated by full lines in Figure 3. The dynamic impact zone, indicated with red shapes in Figure 3, is then created by including all first-order conflicts and *most likely* second-order conflicts.

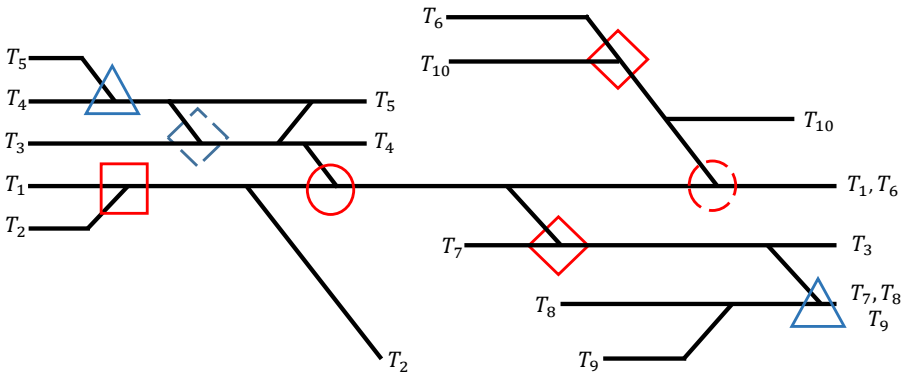


Figure 3: The dynamic impact zone for the initial conflict (indicated by the square) consists of all first-order conflicts (indicated by circles) and most likely second-order conflicts (indicated by full-lined diamonds).

After creating the dynamic impact zone, all possible conflict resolutions are determined. For every conflict resolution, a progress examination of the next half hour is started, including only trains in the dynamic impact zone. During this examination, the first-order and most likely second-order conflicts are considered, as mentioned above, and a resolution needs to be assumed. In our previous approach, it is assumed that these conflicts are solved based on FCFS. Section 5.2 describes how this can be improved. The total secondary delay of every examination is calculated and the solution leading to the lowest secondary delay is chosen.

5.2 Improvements and Extensions

Compared to the research in Van Thielen et al. (2018), several improvements and extensions are included in order to improve the CPS in both quality and computation time and to improve the performance for larger networks.

Resolving New Conflicts

Whenever a conflict is detected during the progress examination, this conflict has to be resolved. Looking at both options and branching further will be too computationally expensive. Therefore, it was first chosen to use FCFS in the progress examination (see Van

Thielen et al., 2018). Now, in order to improve the progress examination, the *immediate impact* of delaying both trains is calculated. A small example illustrates what we call the immediate impact. Figure 4 shows the routes of three trains within a progress examination of the heuristic. The path of T_1 is indicated by the red line, the path of T_2 by the green line and the path of T_3 by the purple line. During the progress examination of the heuristic, a new conflict is detected between T_1 and T_2 on block section **BS-3**. In order to determine which train to delay, the immediate impact in terms of train delay on both trains is determined first. Trains T_1 and T_2 share infrastructure on several subsequent block sections, i.e. **BS-3**, **BS-4**, **BS-5** and **BS-7**. This implies that T_1 and T_2 will be driving behind each other on all these block sections. Moreover, T_1 might be delayed due to a conflict with T_3 on **BS-7** and then further delay T_2, if T_1 drives before T_2. Therefore, in this case, it might be better to let T_2 go first. More generally, if the first train allowed to drive on a common part of the infrastructure is expected to be delayed due to another new conflict later on this common part, then the second train will also be delayed extra. Therefore, it would be better to let the other train go first. This is what we call considering the immediate impact when deciding on how to resolve a new conflict. Specifically, we first determine the set of subsequent block sections with some part of the infrastructure in common belonging to the two trains (T_1 and T_2 in the example) in the new conflict. Then, for every block section in this set (**BS-3**, **BS-4**, **BS-5** and **BS-7** in the example), we look at any other train in the dynamic impact zone that also has common infrastructure (T_3 in the example). If its occupancy based on the delay characteristics at the detection time of the new conflict (partly) coincides with the time interval of one of the two trains in the new conflict, then a delay is imposed on the train in the new conflict (T_1 in the example). This delay is then added to the delay resulting from resolving the new conflict. The option with the least total delay is chosen.

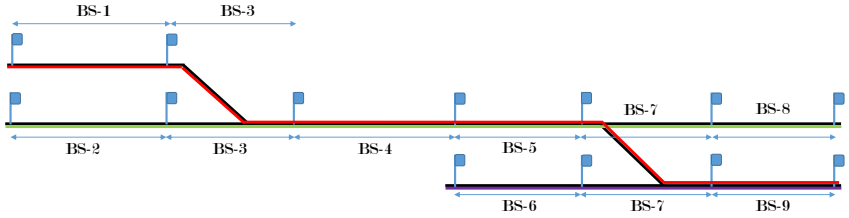


Figure 4: Three train paths on a small network: the red line indicates the path of T_1, the green line the path of T_2 and the purple line the path of T_3.

Updating Potentially Conflicting Trains

The set of all potentially conflicting trains is determined offline beforehand to reduce the computation time online. Two trains have a potential conflict if they want to use the same part of the infrastructure within a time interval of 20 minutes. Dependent on the current situation, at the time the initial conflict is detected, it is possible that these trains are not potentially conflicting anymore. In order to improve and limit the creation of the dynamic impact zone further, trains are updated as potentially conflicting if they share the same part of the infrastructure within a time interval of 10 minutes, according to their current delay. The dynamic impact zone is then created based on this updated set of potentially conflicting trains.

Adding a Maximum Distance from the Initial Conflict
The creation of the dynamic impact zone is limited by imposing a *heuristic horizon* of 30 minutes. If the network is large, this limitation might not be sufficient to control the size of the dynamic impact zone, and thus also the computation time. An additional parameter is therefore included, only searching for first-order conflicts in the dynamic impact zone located within ϵ railway km from the initial conflict.

6 Case Studies

Our proposed CPS is tested on two very large and complex networks including two or three provinces in Belgium. The several adjustments and improvements discussed in Section 5 are tested on both networks. Afterwards, the results of both networks are compared.

6.1 Study Areas

Study Area 1 (SA-1): Provinces of West and East Flanders
This rail network depicted in Figure 5 consists of two provinces in Belgium: West Flanders and East Flanders. The network is approximately 130 km long (De Panne-Puurs) and 60 km wide (Duinbergen-Lauwe). This area contains 130 stations and 11766 block sections. The largest station areas are Brugge, Gent-Sint-Pieters and Oostende. The network is considered in microscopic detail, and is considered with the timetable from 17/03/2017. Between 7 a.m. and 8 a.m., there are at most 240 trains driving in this network. There are 51 rolling stock connections between trains (turnarounds, coupling, de-coupling) taken into account.

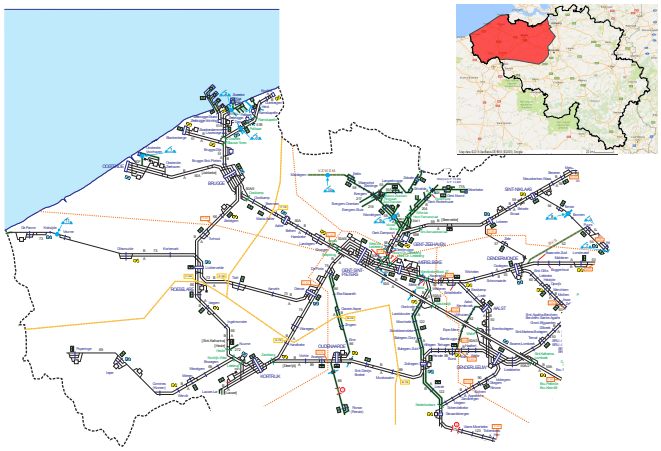


Figure 5: Study area: the provinces of West and East Flanders in Belgium.

Study Area 2 (SA-2): Provinces of Antwerp, West and East Flanders
This rail network includes three provinces of Flanders in Belgium (see Figure 6). This network is approximately 170 km long (De Panne-Zittaart) and 78 km wide (Antwerpen-

Hoeilaart). This area includes 191 station areas and 23917 block sections. The largest station areas are Gent-Sint-Pieters, Brugge, Oostende, Mechelen, Antwerpen-Centraal en Antwerpen-Berchem. Both freight and passenger trains are taken from the microscopic timetable of 17/03/2017. During the time window between 7 and 8 a.m., there are at maximum 353 trains considered and 71 rolling stock connections are provided.

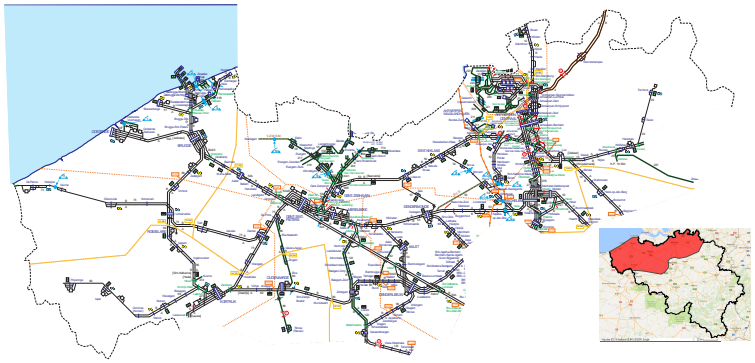


Figure 6: Study area consisting of the three provinces Antwerp, West and East Flanders.

6.2 Results

Table 2 gives an overview of the abbreviations used for the different versions of CPS. Strategy iCPS is the previous strategy from Van Thielen et al. (2018). Strategy iCPS-NC includes the new way of resolving new conflicts. Then, strategy iCPS-UPC includes updating the potential conflicts. These latter two together form the strategy iCPS-IMP. Subsequently, a maximum distance is opposed when creating the dynamic impact zone in iCPS-IMP-10km, iCPS-IMP-50km and iCPS-IMP-100km. The line in bold indicates the best obtained algorithm after performing extensive tests on both networks.

Name	Improvements			Parameters
	Include connections	New conflicts	Update pot conflicts	Max dist conflict
iCPS	✓	-	-	-
iCPS-NC	✓	✓	-	-
iCPS-UPC	✓	-	✓	-
iCPS-IMP	✓	✓	✓	-
iCPS-IMP-10km	✓	✓	✓	10 km
iCPS-IMP-50km	✓	✓	✓	**50 km**
iCPS-IMP-100km	✓	✓	✓	100 km

Table 2: Overview of the different conflict prevention strategies evaluated in this section.

As a standard, the prediction horizon is set to 5 minutes, the heuristic horizon to 30 minutes and the control delay to 60 seconds (see Van Thielen et al. (2018) for more information). The offline calculations are based on 350 runs from a delay scenario with α % of the trains delayed. The value α is randomly taken from the interval $[20\%, 80\%]$. The computation time of the rerouting optimization is limited to 30 seconds.

In order to evaluate the CPS, 20 runs from a delay scenario α are taken, where α is again randomly taken from the interval $[20\%, 80\%]$. In one run, approximately $\alpha\%$ of all trains, which are randomly chosen, are given a random delay from an exponential distribution with an average of 3 minutes and a maximum of 15 minutes. In order to compare the results on both networks, the delay scenarios are first created for the largest network SA-2. The same delay scenarios are then used as input for the network SA-1. The CPS is evaluated based on the total secondary train delay, the average and the maximum computation time. The computation time is the time required to create the dynamic impact zone and perform the progress examinations. All tests are carried out on a Intel(R) Core(TM) i7-3770 CPU 3.40GHz machine.

Table 3 shows the total secondary delay, average and maximum computation time of the different strategies. As expected, the total secondary delay increases when the network is larger. When considering the largest network, more conflicts are detected and resolved. In our simulation, trains cannot reduce their delays during operations, implying that for the largest network the delays can only increase. The improvement compared to FCFS is also somewhat smaller, but still the same order of magnitude (40-50%). The computation time of our CPS remains very similar when enlarging the network, meaning that the dynamic impact zone is well bounded.

Strategy	Train D (in min)		Average computation time (in s)		Maximum computation time (in s)	
	SA-1	SA-2	SA-1	SA-2	SA-1	SA-2
FCFS	660	843	< 0.1	< 0.1	< 0.1	< 0.1
iCPS	369 (- 44 %)	516 (- 38 %)	2.7	3.0	33.7	39.5
iCPS-NC	305 (- 54 %)	478 (- 43 %)	2.7	3.0	33.4	36.8
iCPS-UPC	370 (- 44 %)	513 (- 39 %)	2.3	2.6	32.8	36.8
iCPS-IMP	305 (- 54 %)	468 (- 44 %)	2.3	2.5	33.1	35.4
iCPS-IMP-10km	312 (- 53 %)	475 (- 44 %)	1.8	1.9	34.3	37.5
iCPS-IMP-50km	305 (- 54 %)	467 (- 45 %)	2.2	2.4	32.3	37.5
iCPS-IMP-100km	305 (- 54 %)	467 (- 45 %)	2.3	2.6	33.1	38.4

Table 3: Total secondary delay compared between SA-1 and SA-2.

Clearly, the improvements discussed in Section 5.2 assure that our new Conflict Prevention Strategies perform better both in total secondary delay as in computation time. Combining the new method of resolving new conflicts in the progress examination with updating potential conflicts (iCPS-IMP) attains the same, lower secondary delay as in iCPS-NC, while also attaining the lower computation time as in iCPS-UPC. Opposing a maximum distance from the initial conflict can keep the computation time under control, while also affecting the total secondary delay. Therefore, iCPS-IMP-50km is selected as the best strategy.

6.3 Comparison

By purely extending the network and starting from the same delay scenario, it can be examined which effects extending the network has on the secondary delays.

In order to determine whether the dynamic impact zone is robust enough for extensions to even larger networks, we look closely to the impact zones of conflicts both detected in SA-1 and SA-2 in Table 4. The size of the impact zone is expressed in the number of new conflicts, where a conflict includes two trains and the block section on which the conflict takes place. The average size of the impact zone increases when considering a larger network, leading to the higher computation time, as shown in Table 3.

Strategy	Average size IZ of SA-1	Average size IZ of SA-2	Percentage difference in size IZ
iCPS	279	329	17.9 %
iCPS-NC	274	318	15.8 %
iCPS-UPC	214	256	19.6 %
iCPS-IMP	215	252	17.2 %
iCPS-IMP-10km	80	105	31.3 %
iCPS-IMP-50km	185	216	16.8 %
iCPS-IMP-100km	211	243	15.2 %

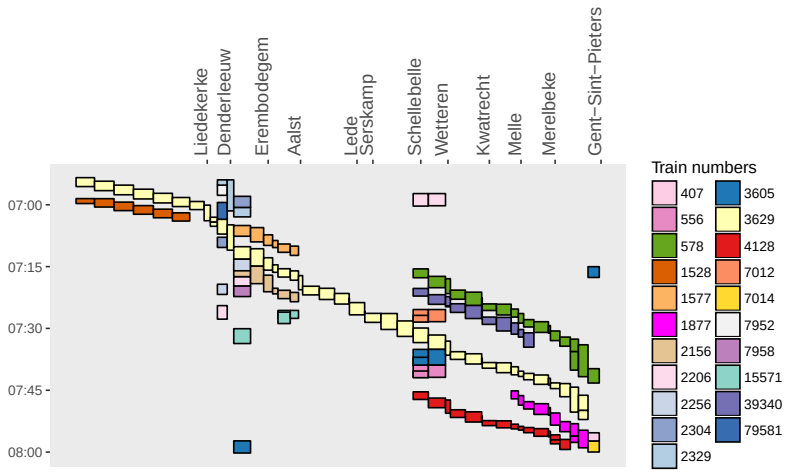
Table 4: Comparison of the size of the impact zone for the same conflicts for both networks. The size is expressed in number of conflicts, which is a couple of trains on one block section.

The strong increase in the size of the impact zone is due to the fact that several conflicts are detected close to the border with the province of Antwerp. When a conflict is located near the border with the province of Antwerp, the dynamic impact zone considered will be larger in SA-2 than in SA-1, since it is artificially bounded in SA-1 by the border of the network considered. Table 5 shows the percentage of initial conflicts located within the maximum distance (10, 50 or 100 km) from the border with the province of Antwerp. This percentage obviously increases when increasing the maximum distance. Consequently, many dynamic impact zones in SA-1 are artifically bounded by 'bumping' into the border of the province of Antwerp. This explains the slightly smaller computation time in SA-1 compared to SA-2. This 'border-effect' could only be avoided by expanding the study area until the entire network of Belgium (and then still international trains cross the borders). Nevertheless, limiting the dynamic impact zone by both the heuristic horizon and the maximum distance will be sufficient in practice to keep the computation time under control.

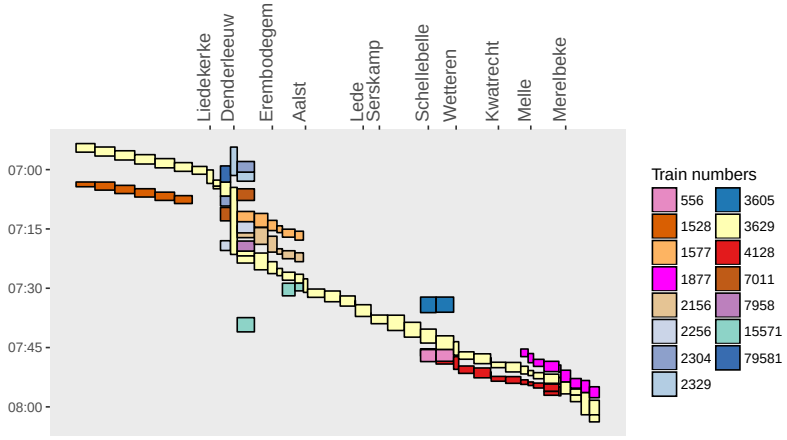
Strategy	Percentage of conflicts
iCPS-IMP-10km	5.2 %
iCPS-IMP-50km	85.9 %
iCPS-IMP-100km	99.7 %

Table 5: Percentage of initial conflicts located closer than 10, 50 or 100 km from the border with the province of Antwerp.

Figure 7 shows the time distance diagram of train 3629 and trains using the same infrastructure within a limited time window. A conflict is detected between trains 2156 and 3629 in both cases, but this conflict is resolved differently. When limiting the network to SA-1, train 2156 is delayed, whereas when the network is extended to SA-2, train 3629 is delayed.



(a) Train path of train 3629 on network SA-1



(b) Train path of train 3629 on network SA-2

Figure 7: Time-distance diagrams of train 3629 resulting from iCPS-50km on SA-1 and SA-2.

7 Conclusion and Future Research

This paper proposes several improvements and extensions to our previous conflict prevention strategy, presented in Van Thielen et al. (2018), making it applicable on larger and complex networks. The extensions allow the computation time to remain limited for real-time purposes. By comparing the same delay scenarios on different sizes of study areas, it is shown that conflicts might be resolved differently. Extending the study area leads to an increase in the total secondary delay, because more conflicts are detected and need to be resolved.

By including these additional improvements and extensions, the total secondary delay decreases, while also reducing the computation time. The basic dispatching strategy FCFS is outperformed by, on average, 40-50 %, while delivering conflict resolutions within 2.4 seconds on average.

Comparing results from both networks is difficult, because when considering the largest network, conflicts in the province of Antwerp are detected and need to be resolved as well. This leads to an increase in the total secondary delay, but also alters the current timetable and route of some trains. Moreover, many dynamic impact zones of initial conflicts considered in the smaller network are artificially bounded by the borders of that smaller network. This slightly reduces the required computation time. Nevertheless, the main conclusions remain that the conflict prevention strategy is significantly improved compared to the previous version and that the computation time can be controlled by limiting the dynamic impact zone by both the heuristic horizon and the maximum distance.

The performance of the conflict prevention strategy is tested using a simulation framework, simulating the real-time situation closely. This assures that the conflict prevention strategy can easily be embedded in a Traffic Management System, such as the one currently implemented at the Belgian railway infrastructure manager Infrabel.

A more detailed analysis would require to consider the entire network considered in practice, the whole of Belgium for instance, in order to mimic the real-time situation as closely as possible. The conflict prevention strategy can easily be applied to such (much) larger networks, but both the Simulator Module and the Conflict Detection Module should be significantly improved before running experiments on larger study areas.

References

Bettinelli, A. and Santini, A. and Vigo, D., 2017. "A real-time conflict solution algorithm for the train rescheduling problem", *Transportation Research Part B*, vol. 106, pp. 237–265.

Borndörfer, R. and Klug, T. and Lamorgese, L. and Mannino, C. and Reuther, M. and Schlechte, T., 2017. "Recent success stories on integrated optimization of railway systems", *Transportation Research Part C*, vol. 74, pp. 196–211.

Cacchiani, V., Huisman, D., Kidd, M., Kroon, L., Toth, P., Veelenturf, L. and Wagenaar, J, An overview of recovery models and algorithms for real-time railway rescheduling. Transportation Research Part B, 63, 15–37 (2014)

Caimi, G. and Fuchsberger, M. and Laumanns, M. and Lüthi, M., 2012. "A model predictive control approach for discrete-time rescheduling in complex central railway station areas", *Computers & Operations Research*, vol. 39, pp. 2578–2593.

Chen, L. and Roberts, C. and Schmid, F., 2015. "Modeling and solving real-time train

rescheduling problems in railway bottleneck sections" *IEEE Transactions on Intelligent Transportation Systems*, vol. 16(4), pp. 1896–1904.

Corman, F. and D'Ariano, A. and Pacciarelli, D. and Pranzo, M., 2012. "Optimal inter-area coordination of train rescheduling decisions", *Transportation Research Part E*, vol. 48, pp. 71–88.

Corman, F. and D'Ariano, A. and Pacciarelli, D. and Pranzo, M., 2012. "Bi-objective conflict detection and resolution in railway traffic management", *Transportation Research Part C*, vol. 20, pp. 79–94.

Corman, F. and Meng, L. A review of online dynamic models and algorithms for railway traffic control. IEEE International Conference on Intelligent Rail Transportation (ICIRT), 2013, Beijing, 128–133 (2013)

Corman, F. and Quaglietta, E., 2015. "Closing the loop in railway traffic control: framework design and impacts on operations", *Transportation Research Part C*, vol. 54, pp. 15–39.

D'Ariano, A. and Corman, F. and Pacciarelli, D. and Pranzo, M, 2008. "Reordering and Local Rerouting Strategies to Manage Train Traffic in Real Time", *Transportation Science*, vol. 42, pp. 405–419.

D'Ariano, A. and Samà, M. and D'Ariano, P. and Pacciarelli, D., 2014. "Evaluating the applicability of advanced techniques for practical real-time train scheduling", *Transportation Research Procedia*, vol. 3, pp. 278–288.

Dolder, U. and Krista, M. and Voelcker, M, 2009. "RCS - rail control system, realtime train run simulation and conflict detection on a netwide scale based on updated train positions", In: *Proceedings of the 3rd International Seminar on Railway Operations Modelling and Analysis, RailZurich (2009)*.

Fischetti ,M. and Monaci, M., 2017. "Using a general-purpose Mixed-Integer Linear Programming solver for the practical solution of real-time train rescheduling", *European Journal of Operational Research*, vol. 263, pp. 258–264.

Hansen, I. and Pachl, J., 2008. "Railway timetable and traffic: analysis, modelling, simulation", Eurail Press.

Josyula, S.P. and Törnquist Krasemann, J. and Lundberg, L., 2018. "A parallel algorithm for train rescheduling", *Transportation Research Part C*, vol. 95, pp. 545–569.

Kecman, P. and Corman, F. and D'Ariano, A. and Goverde, R.M.P., 2013. "Rescheduling models for railway traffic management in large-scale networks", *Public Transport*, vol. 5(1), pp. 95–123.

Lamorgese, L. and Mannino, C., 2015. "An exact decomposition approach for the real-time train dispatching problem", *Operations Research*, vol. 63(1), pp. 48–64.

Lamorgese, L. and Mannino, C. and Piacentini, M., 2016. "Optimal train dispatching by Benders'-like reformulation", *Transportation Science*, vol. 50(3), pp. 910–925.

Lamorgese, L. and Mannino, C. and Pacciarelli, D. and Törnquist Krasemann, J., 2017. Train Dispatching. In: *Handbook of Optimization in the Railway Industry, International Series in Operations Research & Management Science*, chapter 12, pp. 265–283. Springer International Publishing, Switzerland.

Mannino, C. and Mascis, A., 2009. "Optimal real-time traffic control in metro stations", *Operations Research*, vol. 57(4), pp. 1026–1039.

Meng, L. and Zhou, X., 2011. "Robust single-track dispatching model under a dynamic and stochastic environment: a scenario-based rolling horizon solution approach", *Transportation Research Part B*, vol. 45, pp. 1080–1102.

Pellegrini, P., Marlière, G. and Rodriguez, J., 2016. "A detailed analysis of the actual im-

pact of real-time railway traffic management optimization", *Journal of Rail Transport Planning & Management*, vol. 6, pp. 13–31.

Quaglietta, E., Corman, F. and Goverde, R.M.P., 2013. "Stability analysis of railway dispatching plans in a stochastic and dynamic environment", *Journal of Rail Transport Planning & Management*, vol. 3, pp. 137–149.

Quaglietta, E., Pellegrini, P., Goverde, R.M.P., Albrecht, T., Jaekel, B., Marlière, G., Rodriguez, J., Dollevoet, T., Ambrogio, B., Carcasole, D., Giaroli, M. and Nicholson, G., 2016. "The ON-TIME real-time railway traffic management framework: A proof-of-concept using a scalable standardized data communication architecture", *Transportation Research Part C*, vol. 63, pp. 23–50.

Rodriguez, J., 2007. "A Constraint Programming Model for Real-time Train Scheduling at Junctions", Transportation Research Part B: Methodological, 41(2), 231–245.

Samà, M., Meloni, C., D'Ariano, A. and Corman, F., 2015. "A multi-criteria decision support methodology for real-time train scheduling", *Journal of Rail Transport Planning & Management*, vol. 5(3), pp. 146–162.

Samà, M. and Pellegrini, P. and D'Ariano, A. and Rodriguez, J. and Pacciarelli, D., 2016. "Ant colony optimization for the real-time train routing selection problem",*Transportation Research Part B*, vol. 85, pp. 89–108.

Samà, M., D'Ariano, A., Corman, F. and Pacciarelli, D., 2017. "A variable neighbourhood search for fast train scheduling and routing during disturbed railway traffic situations". *Computers & Operations Research*, vol. 78, pp. 480–499.

Törnquist, J., 2007. "Railway traffic disturbance management: an experimental analysis of disturbance complexity, management objectives and limitations in planning horizon", *Transportation Research Part A*, vol. 41(3), pp. 249–266.

Törnquist, J. and Persson, J.A., 2007. "N-tracked railway traffic re-scheduling during disturbances", *Transportation Research Part B*, vol. 41, pp. 342–362.

Törnquist Krasemann, J., 2012. "Design of an effective algorithm for fast response to the re-scheduling of railway traffic during disturbances", *Journal of Transportation Research Part C*, vol. 20, pp. 62–78.

Van Thielen, S. and Corman, F. and Vansteenwegen, P., 2017. "An efficient heuristic for train rescheduling and local rerouting", *Proceedings of 7th International Seminar on Railway Operations Modelling and Analysis (IAROR), RailLille. Lille, France, 4-7 April 2017*, pp. 585–606.

Van Thielen, S., Corman, F. and Vansteenwegen, P., 2018. "Considering a dynamic impact zone for real-time railway traffic management", *Transportation Research Part B*, vol. 111, pp. 39–59.

Van Thielen, S., 2019. "Conflict prevention strategies for real-time railway traffic management", *PhD Thesis. KU Leuven*.

Train Rescheduling Incorporating Coupling Strategy in High-speed Railway under Complete Segment Blockage

Dian Wang [a,b], Jun Zhao [a,b,1], Liuyang Lu [c], Qiyuan Peng [a,b]
[a] School of Transportation and Logistics, Southwest Jiaotong University, Chengdu, Sichuan 611756, China
[b] National United Engineering Laboratory of Integrated and Intelligent Transportation, Southwest Jiaotong University, Chengdu, Sichuan 611756, China
[c] Hangzhou Station, China Railway Shanghai Group Co., Ltd, Hangzhou, Zhejiang 310017, China
[1] Corresponding author, E-mail: junzhao@swjtu.edu.cn

Abstract
This paper investigates the real-time train rescheduling problem in a high-speed railway line under a complete segment blockage by exploring the effectiveness of incorporating train coupling strategy on the train timetable rescheduling. The problem lies on determining the actual arrival and departure time as well as the platform track assignment of trains at stations after a complete segment blockage caused by disruptions, where trains satisfying strict coupling rules could be coupled with others to avoid being cancelled. A mixed integer linear programming model is formulated to minimize the total deviation of trains' arrival and departure time to that in the planned timetable, and to maintain the reasonability of the reordering and coupling decisions. In the model, both the acceleration and deceleration time of trains when departing from and arriving at stations are explicitly considered, while the platform track of trains at passed stations is jointly optimized. A rolling horizon algorithm is designed to effectively solve large-scale problem instances since the rescheduling of timetables is usually determined in stages in practice. Test instances constructed based on the Wuhan-Guangzhou High-Speed Railway in China are utilized to test the effectiveness and efficiency of the proposed approaches. Computational results demonstrate that the train coupling strategy is likely to reduce the total deviation and to relief the propagation of delays. Meanwhile, the rolling horizon algorithm can provide practically acceptable rescheduled timetables quickly. Thus, the train coupling strategy is promising in the field of train timetable rescheduling to cope with large-scale disruptions.

Keywords
Train timetable rescheduling, train coupling strategy, complete segment blockage, mixed integer linear programming, rolling horizon algorithm

1 Introduction

The high-speed railway system is operating based on the preplanned conflict-free timetables and resource utilization schedules if there is no perturbation including disturbance and disruption influencing the railway system. The term "disturbance" is usually utilized for relative small perturbation where only the timetables need to be slightly modified, and the term

"disruption" for relatively large external incidents leading to modifications of not only the timetables but also the duties of rolling stocks or crews (Cacchiani et al., 2014). In real-time operations, however, unexpected perturbations are unavoidable and result in the subsequent infeasibility of preplanned timetables and resource utilization schedules. Passengers experience the negative influences caused by perturbations as train delays, broken connections and even train cancelations. Obviously, it is of great significance and necessity to reschedule train timetables and resources to recover from disturbed or disrupted situations as quickly as possible and to maintain the service level of railway system.

Research in the field of train rescheduling is promising from a practical point of view. However it is also a challenging work especially for the high-speed railway line with dense traffics and higher operating speed. Currently in practice, the rescheduling of train timetables and if necessary rolling stocks and crews, are mainly manually implemented by involved dispatchers based on their experiences and craftsmanship. The practical feasibility and quality of the resulting manually rescheduled plans are not certainly assured. Fortunately on the contrary, the real-time train rescheduling has attracted widely attentions in the academic community recently. Many researchers are devoting themselves to apply their advanced recovery approaches implemented in user-friendly intelligent decision support systems to improve the service and reliability of railway systems.

1.1 Related Works

Recently in high-speed railway system, the most common measures considered in practice and related academic researches to recover from a disturbed or disruption situation to a feasible one is the train timetable rescheduling, which is mainly further composed of retiming, reordering and rerouting, as well as cancelling trains if a large external incidence occurs. To reduce the negative influences caused by unpredicted perturbations, the rescheduling measures should be discreetly adopted to design high quality practically feasible rescheduled timetables. Up to now, a mass of mathematical models and algorithms have been developed to support dispatchers to make reasonable decisions. According to Cacchiani et al. (2014), existing approaches can be classified by the scale of the perturbations including disturbances and disruptions, and the level of detail considered in the railway system known as macroscopic and microscopic perspectives. In macroscopic approaches, the stations and the tracks between adjacent stations (i.e. segments) are treated as nodes and arcs, respectively, and the details of block sections and signals at stations and along segments are not taken into account. However, these aspects are all considered in detail in microscopic researches. In this paper, We focus on the real-time train timetable rescheduling under a complete segment blockage from a macroscopic aspect, where a complete blockage is denoted by Louwerse and Huisman (2014) as the situation in which all tracks of a segment are blocked and no trains can be operated on this segment. Thus, we mainly restrict ourselves to typical previous studies on real-time train timetable rescheduling under disrupted situations from a macroscopic perspective. Interested readers can refer to Cacchiani et al. (2014), Corman and Meng (2015) and Fang et al. (2015) for detailed reviews on traffic management/rescheduling of railway system, and to Törnquist and Persson (2007) and Krasemann (2012) for detailed methodologies dealing with disturbed situations.

Louwerse and Huisman (2014) focused on adjusting the timetable of a passenger railway operator in case of partial or complete blockages. An event-activity network was utilized to formulate their integer programming formulations, while the effectiveness of their

models was tested based on periodic timetables collected from the Netherlands Railways. Zhan et al. (2015) and Zhan et al. (2016) studied similar problems of which the objective was minimizing the number of canceled trains and the total weighted delay (or deviations composed of earliness and tardiness). A two-stage algorithm and a rolling horizon approach were designed respectively to solve realistic instances constructed based on the non-periodic timetables in China. The capacity of infrastructures and rolling stocks as well as rerouting of trains were further considered by Veelenturf et al. (2016). As observed, cancelling trains is an important strategy adopted in existing studies to reschedule train timetables under disruptions. Besides, in these studies only the trains which have not already left their origin station when the disruption occurs are allowed to be cancelled. However, it is challenging to reschedule these trains not allowed to be cancelled, especially when the capacity of stations expressed by the number of platform tracks at stations is relative few, as trains need to dwell on a certain platform track at a reasonable station to wait for the recovery of the disruption.

Except for the common rescheduling measures (i.e. retiming, reordering, rerouting, and cancelling trains if necessary) adopted in practice, there are also other specific strategies in previous works which are designed to reduce the negative influences caused by the disruption or even the cancelation of trains, such as the stop-skipping strategy in Altazin et al. (2017) and short-turning strategy in Ghaemi et al. (2018). Altazin et al. (2017) investigated the train rescheduling problem through stop-skipping in dense railway systems and formulated their problem as an integer linear programming, where some stops of train services can be skipped such that the propagation of delays might be reduced. Ghaemi et al. (2018) formulated a macroscopic integer linear short-turning model in case of simultaneous complete blockages, such that the penalized cancellations and delay of planned trains services can be minimized. In addition to the operator-oriented works mentioned above, passenger-oriented timetable rescheduling is also attractive. Sato et al. (2013) formulated an MIP model to minimize the further inconveniences to passengers caused by the disruption so as to exactly consider the loss of time and satisfaction of passengers.

This paper tries to optimize the real-time train timetable rescheduling incorporating train coupling strategy in a high-speed railway line in case of a complete segment blockage. Under the train coupling strategy, two trains which strictly satisfy specific rules are allowed to be coupled on a platform track at a certain station once a large perturbation occurs, such that these two trains can form one train and run subsequent stations and segments along their planned route together. Obviously, the number of trains can be reduced while not cancelling any train by utilizing the train coupling strategy. Note that the coupling/combining of passenger trains has attracted attentions in early works focusing on the circulation of rolling stocks, such as Fioole et al. (2006) and Peeters and Kroon (2008). In these works, the rolling stocks can be added/combined or removed/splited from trains according to the predefined timetable and passenger demand for the efficient utilization of train units. These problems as tactical decisions arises in an early phase of the railway planning process. However, to the best of our knowledge, there is no previous work investigating the operational train rescheduling incorporating train coupling in the real-time setting.

1.2 Contributions

The contributions of this paper are mainly threefold. Firstly, as far as we know, our paper might be the first one trying to explore the practicability and effectiveness of train coupling strategy to avoid cancelling trains in train timetable rescheduling under a disruption

of complete segment blockage, such that the negative influences caused by the cancellation of trains can be reduced as much as possible. Secondly, different with many existing macroscopic train rescheduling works (Cacchiani et al., 2014; Zhan et al., 2015, 2016), in this paper a station is represented by many platform tracks rather than a single node and the occupation of platform tracks at stations are determined, due to that the capacity of stations is represented more finely. Finally, several operational requirements are further considered in our approaches. The additional acceleration and deceleration time of trains when stopping at stations and the platform track assignment of trains at nonstop passed stations are all exactly incorporated to reflect better the actual situations of high-speed railway systems.

1.3 Outline of Paper

The rest of this paper is organised as follows. Firstly, a detailed problem description is presented in Section 2. In Section 3, a mixed integer linear programming model is established by taking into account many operational and safety requirements. Next, a rolling horizon algorithm is designed in Section 4 to effectively solve large-scale problems. Then, in Section 5 computational tests on instances constructed from Wuhan-Guangzhou High-Speed Railway in China are implemented to test the effectiveness and efficiency of the proposed approaches. Comparison of rescheduling strategies is also conducted in this section. Finally, we conclude our main research works in Section 6.

2 Problem Description and Assumptions

2.1 Problem Description

This paper investigates the real-time train timetable rescheduling incorporating train coupling strategy in a high-speed railway line under a complete segment blockage from the macroscopic prospective, where a station is treated as several platform tracks instead of a single node to model the capacity of stations, as illustrated by Figure 1. We mainly focus on the Chinese situation where trains are running on separated double parallel tracks in a high-speed railway line. When a complete segment blockage caused by disruptions occurs, trains bounding for the disrupted segment in both the downstream and the upstream directions have to wait on the platform tracks at reasonable stations until the disrupted situation is recovered. The consequent negative influences to the operators and passengers should be controlled which is usually achieved by the strategies of retiming, reordering, rerouting and canceling trains to minimize the total deviation of trains' arrival and departure time to that in the planned timetable. Large negative influences are usually inevitable when trains have to be cancelled due to the limited capacity of stations and segments.

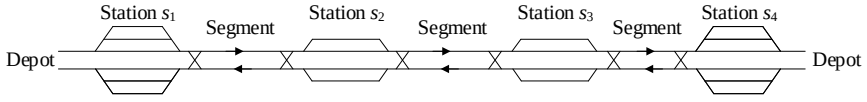


Figure 1: Illustration of a high-speed railway line

The purpose of this paper lies on exploring the effects of train coupling strategy on the train timetable rescheduling such that the cancellation of trains and its subsequent negative

influences might be reduced. Under the train coupling strategy, two trains strictly satisfying specific coupling rules can be coupled together at a certain station to form one train so as to reduce the number of trains needed to be arranged at subsequent segments and stations along the line. We consider the coupling rules in the high-speed railway in China. To be specific, only the trains served by the same type of rolling stock with 8-carriage are able to be coupled with each other. Meanwhile, if two trains are about to couple at a station, they should pass through the same subsequent stations and terminate at the same destination station. Besides, there are mainly two coupling modes of trains based on practical situations. The first one is the shunting mode in which the former train firstly arrives and stops on a platform track at a station. When the latter train arrives at the same station, it firstly stops on another platform track and then couples with the former train through shunting operations. The second one is the receiving mode. There, at the coupling station, the former train arrives and stops on a platform track. Next when the latter train arrives, it firstly bounds for the same platform track and stops behind the former train, and then it couples with the former train with a lower speed. Obviously, the second mode can increase the utilization efficiency of platform tracks. Thus we formulate our train rescheduling approaches based on the second train coupling mode. Moreover, to ensure the practicability of the rescheduled timetables, the detailed occupation of platform tracks of trains at each passed station should also be exactly determined, as the safety requirements at stations and on segments expressed by different headway between trains have to be strictly fulfilled.

The railway line shown in Figure 1 is used to describe our problem. This line has 4 stations denoted as s_1–s_4 along the downstream direction. As trains run independently in the two directions of the line, w.l.o.g. we only consider the train rescheduling in the downstream direction, and trains are not allowed to utilize the tracks that normally are used in the opposite direction. Along the downstream direction, at stations s_1 and s_4 there are 3 platform tracks denoted as k_1–k_3 based on their distance to the main track (i.e. k_1), while only 2 platform tracks are set in intermediate stations s_2 and s_3. There are in total 5 trains numbered as i_1–i_5 running and terminating at station s_4 in the line. The planned timetable of these trains is displayed by the blue lines in Figure 2(a). Suppose that a disruption occurs in segment (s_3, s_4) at time t_1 leading to a complete blockage to this segment, which is predicted to be recovered at time t_2 and expressed by the light gray rectangle, these planned trains will be affected by the disruption and should be rescheduled. Feasible rescheduled timetables without and with the train coupling strategy are illustrated by Figures 2(b) and 2(c), respectively, where red lines indicate that the related trains are affected by the disruption at associated stations and segments, and magenta lines represent that the related trains couple with others at a certain station and pass through the subsequent segments together. Meanwhile, the rescheduled platform track assignment at parts of stations under coupling strategy is shown in Figure 2(d), where dark gray rectangles illustrate the platform track occupations of corresponding trains at associated stations.

As observed from Figure 2, when the disruption occurs, trains i_1–i_3 are directly affected by the disruption and each of them should dwell on a platform track at a certain station to wait for the recovery of the disruption. In the given rescheduled timetable using coupling, these trains are arranged to stop at station s_3 such that the planned timetable of these trains before station s_3 can be strictly fulfilled. Meanwhile, due to the lack of platform tracks, trains i_1 and i_2 couple with each other at station s_3 on platform track k_1 and pass through the subsequent segment (s_3, s_4) together. At this point, train i_3 can arrive at station s_3 and stop at platform track k_2 until the disruption is finished. Even though train i_4 is not

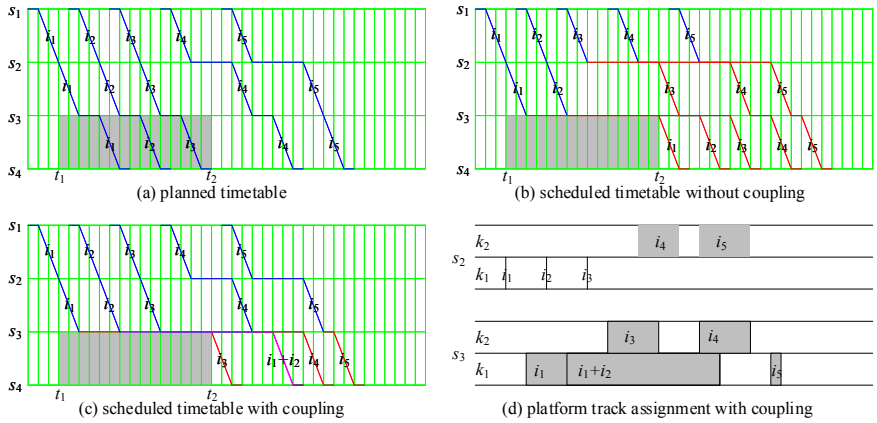


Figure 2: Representation of rescheduled timetable and platform track assignment

directly affected by the disruption, to maintain the enough headway between the departures of coupled train $i_1 + i_2$ and train i_4 at station s_3, train i_4 is postponed to depart from the station. Similarly, train i_5 is delayed at station s_3 to maintain the departure headway between train i_4. Note that trains should occupy the associated main track at each nonstop passed station (e.g. trains i_1-i_3 at station s_2) according to the practical requirement in China. By comparing Figures 2(b) and 2(c), the influence of the disruption to the planned timetable can be obviously reduced at a certain degree.

From the perspective of railway operators, the purpose of train rescheduling after disruptions is to maintain the stability of planned train timetables and to reduce the inconveniences to passengers as much as possible. The total deviation of timetables is widely used as the objective function for train rescheduling (e.g. Zhan et al. (2015)) in China. As a result, there are likely different rescheduled timetables with the same objective function value caused by the reordering and coupling of trains. For example, in Figure 2(b) trains i_1 and i_2 cross segment (s_3, s_4) sequentially based on their planned order. However, it is also feasible by swapping the order of these trains while not increasing the total deviation. At the same time, in Figure 2(c) trains i_1 and i_2 are coupled at station s_3, while coupling train i_1 and i_3 could also obtain the best objective function if the dwell time is enough for the associated coupling of trains i_1 and i_3. Obviously, swapping trains i_1 and i_2 in Figure 2(b) and coupling trains i_1 and i_3 instead of trains i_1 and i_2 might be strange and not be attractive for practical application, they should be prevented as much as possible while not deteriorating the total arrival and departure deviation.

Thus, the real-time train rescheduling problem considered in this paper is defined as follows. Given the layout of the studied high-speed railway line, the capacity of stations and segments, the planned timetable, and the location, start time and predicted duration of the segment blockage, our problem lies on determining the actual arrival time, departure time and platform track of trains at passed stations along their predetermined route, as well as the coupling decisions of trains, such that the weighted sum of the total deviation of trains' arrival and departure time to that in the planned timetable and the strange reordering and coupling decisions is minimized, and specific operational and safety requirements are respected.

2.2 Assumptions

We focus on incorporating the coupling strategy to improve the quality of train rescheduling in a high-speed railway line under a complete segment blockage from the macroscopic perspective. To facilitate the formulation of our model, the following assumptions are made.

- We only consider one side of the stations along the railway line. In other words, trains are not allowed to utilize tracks that normally are used in the opposite direction.
- When disruption occurs, the trains that locate at the blocked segment cannot pass through the segment and they should return to the behind station incident to the segment to wait until the disruption is recovered.
- The earliness and tardiness of arrival time are both allowed, while the earliness of departure time will never occur for the consideration of the boarding of passengers.
- The cancelling of trains is not considered as the train coupling strategy is adopted.
- At most two trains which strictly satisfy the coupling rules can be coupled together at a certain station due to the length of platform tracks at stations. The coupled trains will not be decoupled until their destination station is reached.

3 Model Formulation

3.1 Notation

We formulate our problem as a mixed integer linear programming model. The sets, indices and parameters to be used in the formulation of the model are explained in Table 1, and Table 2 expresses the decision variables.

3.2 Objective

As introduced, from the perspective of railway operators, it is necessary to minimize the total deviation of trains' arrival and departure time to that in the planned timetable so as to maintain the stability of timetable as far as possible after disruptions. At the same time, the unattractive reordering and coupling should be eliminated as much as possible. This objective function is expressed as follows.

$$
\begin{aligned}
&\min \quad U = U_1 + U_2 + U_3 \\
&U_1 = \sum_{i \in T} \sum_{m \in A_i} y_{im} + \sum_{i \in T} \sum_{m \in A_i} (f_{im} - d_{im}) \\
&U_2 = \sum_{i \in T} \sum_{m \in A_i} \sum_{j \in C_{im}} \gamma_{ijm} \cdot x_{ijm} \\
&U_3 = \sum_{i \in T} \sum_{j \in T} \sum_{(m,n) \in B_i \cap B_j} \pi_{ijmn} \cdot \lambda_{ijmn} \cdot u_{ijmn}
\end{aligned}
\tag{1}
$$

The first part of U_1 is the total deviation of arrival time including the tardiness and earliness of arrival time simultaneously, and the second part is that of departure time which only

Table 1: Definition of sets, indices and parameters

Notation	Description
T	Set of trains, $T = \{1, 2, \cdots, \|T\|\}$, $\|T\|$ is the number of trains running in the studied line.
i, j	Index of trains, $i = 1, 2, \cdots, \|T\|$, $j = 1, 2, \cdots, \|T\|$.
S	Set of stations which are indexed along the downstream direction, $S = \{1, 2, \cdots, \|S\|\}$ where $\|S\|$ is the number of stations in the studied line.
m, n, s	Index of stations, $m = 1, 2, \cdots, \|S\|$, $n = 1, 2, \cdots, \|S\|$, $s = 1, 2, \cdots, \|S\|$.
E	Set of segments, $E = \{(m, n) \| m, n \in S\}$.
(m, n)	Index of segments which represents the segment between adjacent stations m and n.
A_i, B_i	Set of stations and segments contained in the predetermined route of train i, respectively.
K_m	Set of platform tracks at station m indexed incrementally by their distance to the main track.
k	Index of platform tracks, where the index of the main track at each station equals to 1.
θ_{im}	Order of train $i \in T$ to leave station $m \in A_i$ based on the planned timetable. Note that θ_{im} is not always equal to i as the overtaking of trains usually exists.
β	Integer constant introduced to assure the attraction of the coupling decision. It requires that a train can only couple with its previous and latter β trains satisfying the coupling rules at a passed station.
C_{im}	Set of trains which can be coupled with train i at station $m \in A_i$. It is generated in advance based on the predefined route of trains and coupling rules as well as β to ensure the reasonability of the rescheduled timetable.
N_{ij}	Set of segments where train i and train j can be coupled together to pass through, $N_{ij} \subseteq B_i \cap B_j$. If these two trains do not satisfy the coupling rules, $N_{ij} = \emptyset$.
t_1, t_2	Start time and predicted end time of the disruption, respectively.
(e_1, e_2)	Disrupted segment, where e_1 and e_2 are its behind and front incident station, respectively.
a_{im}, d_{im}	Scheduled arrival and departure time of train i at station $m \in A_i$, respectively.
r^1_{imn}, r^2_{imn}	Minimum and maximum running time of train i on segment $(m, n) \in B_i$, respectively.
q_1, q_2	Additional acceleration and deceleration time of trains once stopping at stations, respectively.
π_{ijmn}	0-1 constant, 0 if train $i \in T$ enters segment $(m, n) \in B_i \cap B_j$ before train j enters the segment based on the planned timetable, 1 otherwise.
b_{im}	Minimum dwell time of train i at station $m \in A_i$ for the boarding and alighting of passengers.
g_m	Duration time to couple two trains which strictly satisfy the coupling rules at station m.
δ_{ij}	The first station at which trains i and j can be coupled together. If these two trains do not satisfy the coupling rules, $\delta_{ij} = \emptyset$.
h_1	Departure headway of two consecutive trains to depart from the same station.
h_2	Arrival headway of two consecutive trains to arrive at the same station.
h_3	Departure-arrival headway of two consecutive trains not being coupled together.
h_4	Arrival-departure headway of two consecutive trains not being coupled together.

Table 2: Definition of decision variables

Notation	Description
x_{ijm}	Binary variable, 1 if train i is coupled with train $j \in C_{im}$ at station $m \in A_i$, 0 otherwise.
y_{im}	Nonnegative integer variable, represents the arrival time deviation of train i at station $m \in A_i$ compared to that in planned timetable.
c_{im}	Nonnegative integer variable, represents the actual arrival time of train i at station $m \in A_i$.
f_{im}	Nonnegative integer variable, represents the actual departure time of train i at station $m \in A_i$.
w_{im}	Binary variable, 1 if train i stops at station m in the rescheduled timetable, 0 otherwise.
u_{ijmn}	Binary variable, 1 if the actual time of train i to enter segment $(m, n) \in B_i \cap B_j$ is earlier than that of train j, 0 otherwise.
p_{ijm}	Binary variable, 1 if the actual departure time of train i from station $m \in A_i \cap A_j$ is earlier than the actual arrival time of train j at the station, 0 otherwise.
v_{imk}	Binary variable, 1 if train i occupies platform track $k \in K_m$ at station m, 0 otherwise.
z_{ijmn}	Binary variable, 1 if trains i and j couple together to cross segment $(m, n) \in N_{ij}$, 0 otherwise.

contains tardiness. U_2 is introduced to penalize the unattractive train coupling decisions, where γ_{ijm} is a small constant. As coupling consecutive trains seems to be much more

attractive for practical application, we set γ_{ijm} to $|\theta_{im} - \theta_{jm}|$. Similarly, U_3 is utilized to penalize the unattractive reordering of trains, where λ_{ijmn} is a small constant which is also set to $|\theta_{im} - \theta_{jm}|, \forall(m,n) \in B_i \cap B_j$.

3.3 Constraints

Train running constraints
Specific train running requirements should be strictly satisfied to maintain the feasibility of rescheduled timetables and the safety of trains. Constraints (2) mean that the actual running time of trains on a segment should be no less than the minimal time and be no greater than the maximum time to maintain the practical feasibility, where the additional acceleration and deceleration time are exactly considered. Note that the range of running time of a train whether being coupled with others or not on a segment makes no difference as each train has the tractive force. Indeed the actual running time of each train on a segment is also flexible within the range in this paper. Constraints (3) and (4) calculate the deviation of arrival time to that in planned timetable, where the former is dedicated for the tardiness and the latter for the earliness. Constraints (5) require that trains cannot depart from any passed station ahead of planned time. Trains are prevented from entering the disrupted segment during the disruption by constraints (6) to ensure the safety of trains. Besides, these constraints can also maintain that the trains locating at the disrupted segment once the disruption occurs should return to the behind station incident to the disrupted segment.

$$r^1_{imn} + q_1 \cdot w_{im} + q_2 \cdot w_{in} \leq c_{in} - f_{im} \leq r^2_{imn} \quad \forall i \in T, \forall(m,n) \in B_i \tag{2}$$

$$y_{im} \geq c_{im} - a_{im} \quad \forall i \in T, \forall m \in A_i \tag{3}$$

$$y_{im} \geq a_{im} - c_{im} \quad \forall i \in T, \forall m \in A_i \tag{4}$$

$$f_{im} - d_{im} \geq 0 \quad \forall i \in T, \forall m \in A_i \tag{5}$$

$$f_{ie_1} \geq t_2 \quad \text{if } (d_{ie_1}, a_{ie_2}) \cap [t_1, t_2] \neq \emptyset \quad \forall i \in T | (e_1, e_2) \in B_i \tag{6}$$

Train dwelling constraints
Specific train dwelling requirements should be fulfilled to enable the normal boarding and alighting of passengers and the coupling of trains. Constraints (7) ensure that the dwell time of trains at stations should be valued enough for the boarding and alighting of passengers and the coupling of trains if necessary. Constraints (8) are designed to determine whether a train needs to stop at a station after the disruption, where M_1 is a large positive constant and its value could be the length of the studied timetable. Together with constraints (7), no station at which a train is about to stop in the planned timetable will be skipped.

$$b_{im} + g_m \cdot \sum\nolimits_{j \in C_{im}} x_{ijm} \leq f_{im} - c_{im} \quad \forall i \in T, \forall m \in A_i \tag{7}$$

$$w_{im} \leq f_{im} - c_{im} \leq M_1 \cdot w_{im} \quad \forall i \in T, \forall m \in A_i \tag{8}$$

Train coupling constraints
Any two trains if being coupled together should satisfy not only the strict coupling rules but also specific operational requirements. Constraints (9) mean that each train can be coupled with at most one another train at only a certain station for the consideration of

operations. Constraints (10) represent that if trains i and j are coupled together on segment $(m,n) \in N_{ij}$, then they should also be coupled to pass through the immediate subsequent segment $(n,s) \in N_{ij}$ since coupled trains are not allowed to be decoupled until they reach their destination station. Constraints (11) and (12) are introduced to express the relationship between variables z_{ijmn} and x_{ijm} based on their definition, which imply that trains only might be coupled at a station and coupled train cannot decoupled until arrives at destination station. Constraints (13) and (14) assure that the actual departure and arrival time of two trains coupled at a certain station should be equal at subsequent stations.

$$\sum\nolimits_{m \in A_i} \sum\nolimits_{j \in C_{im}} x_{ijm} \leq 1 \qquad \forall i \in T \tag{9}$$

$$z_{ijns} \geq z_{ijmn} \qquad \forall i,j \in T, \forall (m,n),(n,s) \in N_{ij} \tag{10}$$

$$x_{ijn} = z_{ijns} - z_{ijmn} \qquad \forall i,j \in T, \forall (m,n),(n,s) \in N_{ij} \tag{11}$$

$$x_{ij\delta_{ij}} = z_{ij\delta_{ij}n} \qquad \forall i,j \in T, (\delta_{ij},n) \in N_{ij} \tag{12}$$

$$M_1 \cdot (z_{ijmn} - 1) \leq f_{im} - f_{jm} \leq M_1 \cdot (1 - z_{ijmn}) \quad \forall i,j \in T, \forall (m,n) \in N_{ij} \tag{13}$$

$$M_1 \cdot (z_{ijmn} - 1) \leq c_{in} - c_{jn} \leq M_1 \cdot (1 - z_{ijmn}) \quad \forall i,j \in T, \forall (m,n) \in N_{ij} \tag{14}$$

Train headway constraints

There are series of headway requirements that should be strictly met to avoid the potential route conflicts of trains at stations, including the departure headway h_1, arrival headway h_2, departure-arrival headway h_3 and arrival-departure headway h_4. The headway between two consecutive trains which are not coupled together is illustrated by Figure 3.

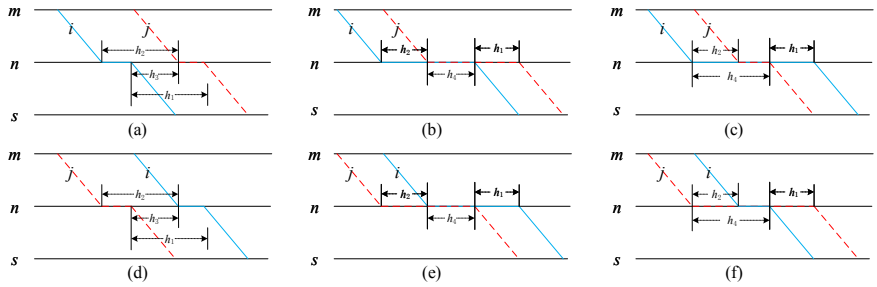


Figure 3: Headway between two consecutive trains

As observed from Figure 3, the arrival and departure headway between two consecutive trains should always be respected, while either the departure-arrival headway (Figures 3(a) and 3(d)) or the arrival-departure headway (Figures 3(b), 3(c), 3(e) and 3(f)) should be strictly satisfied. For example, if the departure-arrival headway between the departure of train i and the arrival of train j at station n is fulfilled shown in Figure 3(a), then the arrival-departure headway between the arrival of train i and the departure of train j at the station can be naturally respected. As a consequence, the train headway constraints are formulated as follows.

$$u_{ijmn} + u_{jimn} = 1 - z_{ijmn} \qquad \forall i,j \in T, \forall (m,n) \in B_i \cap B_j \tag{15}$$

$$f_{im} + h_1 \leq f_{jm} + M_1 \cdot (1 - u_{ijmn}) \quad \forall i, j \in T, \forall (m,n) \in B_i \cap B_j \tag{16}$$

$$c_{in} + h_2 \leq c_{jn} + M_1 \cdot (1 - u_{ijmn}) \quad \forall i, j \in T, \forall (m,n) \in B_i \cap B_j \tag{17}$$

$$f_{im} + h_3 \leq c_{jm} + M_1 \cdot (1 - p_{ijm}) \quad \forall i, j \in T, \forall m \in A_i \cap A_j \tag{18}$$

$$c_{jm} + h_4 \cdot (1 - z_{ijmn}) \leq f_{im} + M_1 \cdot p_{ijm} \ \ \forall i, j \in T, \forall m \in A_i \cap A_j \tag{19}$$

Constraints (15) reflect the relationship between variables u_{ijmn} and z_{ijmn}, which mean that if trains i and j are not coupled together to pass through section $(m,n) \in B_i \cap B_j$, i.e. $z_{ijmn} = 0$, then the time of train i to enter the segment should be earlier than that of train j, or on the contrary. Otherwise, these two trains should enter the segment at the same time and they need not to satisfy the departure headway at station m. Note that these constraints transform to $u_{ijmn} + u_{jimn} = 1$ if $(m,n) \in B_i \cap B_j$ and $(m,n) \notin N_{ij}$. Constraints (16) and (17) maintain the headway between two consecutive trains to depart from a station (i.e. departure headway) and to arrive at a station (i.e. arrival headway), respectively. Obviously, these constraints do not apply for coupled trains on segment (m,n). At the same time, these two constraints can also prevent the overtaking of trains along the segment. The departure-arrival headway of two consecutive trains is guaranteed by constraints (18) which only take effect under the situation that $p_{ijm} = 1$ or $p_{jim} = 1$ illustrated by Figures 3(a) and 3(d), respectively. Constraints (19) are for the arrival-departure headway which should be respected if the actual departure time of train $i(j)$ is not earlier than the arrival time of train $j(i)$ at station m, i.e. $p_{ijm}(p_{jim}) = 0$. Note that $p_{ijm} = 0$ holds if $z_{ijmn} = 1$ according to constraints (13) and (18). Then constraints (19) are transformed to $c_{jm} \leq f_{im}$ which are obviously valid since $c_{jm} \leq f_{im} = f_{jm}$ if $z_{ijmn} = 1$ according to constraints (13).

Station capacity constraints

The capacity of stations is expressed by the headway between two trains to occupy the same platform track since each track can be occupied by only one train or two coupled trains at a time. Meanwhile, a track should have been cleared for a specific time when another train starts to occupy the track. As observed from Figure 3, only under the situations in Figure 3(a) and 3(d), the two consecutive trains which are not coupled together or are about to be coupled at station m can occupy the same platform track at the station. Note that the necessary headway for these trains to occupy the same platform track has ensured by constraints (18). Thus, the station capacity requirements are expressed as follows.

$$\sum\nolimits_{k \in K_m} v_{imk} = 1 \quad \forall i \in T, \forall m \in A_i \tag{20}$$

$$\sum\nolimits_{k \in K_m | k \neq 1} v_{imk} \leq w_{im} \quad \forall i \in T, \forall m \in A_i \tag{21}$$

$$v_{imk} + v_{jmk} \leq 1 + p_{ijm} + p_{jim} + z_{ijmn} \ \ \forall i, j \in T, \forall m \in A_i \cap A_j, \forall k \in K_m \tag{22}$$

$$M_2(z_{ijmn} - 1) \leq \sum\nolimits_{k \in K_m} k \cdot v_{imk} - \sum\nolimits_{k \in K_m} k \cdot v_{jmk} \leq M_2(1 - z_{ijmn}) \quad \forall i, j \in T, \forall (m,n) \in N_{ij} \tag{23}$$

Constraints (20) declare that each train should occupy exact one platform track at each of its passed station. Along with constraints (20), constraints (21) require that the trains not about to stop at a passed station should occupy the associated main track (i.e. $k = 1$) at the station. Constraints (22) and (23) together with the train headway constraints are designed to reflect the station capacity requirements. Constraints (22) mean that if trains i and j

occupy the same platform track k at station $m \in A_i \cap B_j$ (i.e. $v_{imk} = v_{jmk} = 1$), then these two trains should be coupled to pass through the subsequent segment $(m, n) \in N_{ij}$ (i.e. $z_{ijmn} = 1$), or these trains should satisfy the departure-arrival headway illustrated in 3(a) and 3(d) (in other words, $p_{ijm} + p_{jim} = 1$ should hold). Note that constraints (22) will be transformed to $v_{imk} + v_{jmk} \leq 1 + p_{ijm} + p_{jim}$ if $m \in A_i \cap A_j$ and $(m, n) \notin N_{ij}$. Constraints (23) ask that if trains i and j are about to be coupled together to pass through segment $(m, n) \in N_{ij}$ (i.e. $z_{ijmn} = 1$), then they should occupy the same platform track at station m (i.e. $\sum_{k \in K_m} k \cdot v_{imk} = \sum_{k \in K_m} k \cdot v_{jmk}$). where M_2 is a large positive constant and it can be set to the number of platform tracks at station m.

4 Solution Approach

Overall, the real-time train timetable rescheduling incorporating coupling strategy (TRCS) in a high-speed railway line under a complete segment blockage can be formulated as a mixed integer linear programming model to minimize objective (1) under constraints (2)–(23). Obviously, the original problem is NP-hard as it can be easily reduced to the NP-hard problem investigated in Zhan et al. (2015) if trains are not allowed to couple (i.e. to set all x_{ijm} to 0 in advance). Fortunately, our model is a linear programming due to that optimal or high quality feasible solutions for small-scale problems can be obtained quickly by state-of-the-art commercial solvers. Observe that train dispatchers usually reschedule timetables in stages in practice as the duration of the disruption is updated gradually. Thus, a rolling horizon algorithm is customized to effectively solve large-scale problems under the real-time decision requirement of train rescheduling. The effectiveness of rolling horizon algorithm in the field of railway rescheduling has been testified by several previous works such as Zhan et al. (2016) for the train timetable rescheduling and Nielsen et al. (2012) for the rolling stock rescheduling.

In our algorithm, the original problem (TRCS) is decomposed into several small-scale subproblems according to the given horizon length σ and update step size τ. Specifically, the long time span of the original problem is divided into several overlapped shorter stages in each of which a similar subproblem is directly solved by commercial solvers. The procedures of the algorithm are as follows.

Step 1: Initialization. We firstly initialize the stage $l = 0$, the considered train set $T_l = \emptyset$ in stage l, the passed station set $A_i^l = \emptyset$ of train i in the stage. Then, we set the start time of the algorithm denoted as t_{start} to be the earliest planned arrival time of all affected trains at their origin station. Meanwhile, suppose that D_l (which includes the trains of which all the arrival and departure time at all passed stations have been fixed) is composed of the trains certainly not affected by the disruption, i.e. the trains which have crossed the disrupted segment before the occurrence of the blockage and the trains will not pass the disrupted segment according to their predetermined route from the planned timetable. Finally, introduce the best rescheduled timetable $X^* = \{c_{im}^*, f_{im}^*, v_{imk}^*\}$ of the algorithm by setting all of its elements to be 0. Set $l = l + 1$ and go to the next step.

Step 2: Pick out the considered train in stage l. Firstly we calculate the start time t_{start}^l and the end time t_{end}^l of stage l by $t_{\text{start}}^l = t_{\text{start}} + (l-1) \times \tau$ and $t_{\text{end}}^l = t_{\text{start}}^l + \sigma$. Then, we pick out the considered train set T_l in the stage based on the range of $[t_{\text{start}}^l, t_{\text{end}}^l]$. To be specific, $T_l = \{T_{l-1} \cup I_l\} \backslash D_{l-1}$, where I_l includes the trains that are newly about to run at a certain station or segment in stage l (i.e. the trains at least one of their planned arrival and departure time locates within the range).

Step 3: Update the passed station set A_i^l for each train $i \in T_l$. The origin station of train i in stage l is set to either the last station at which its actual arrival time is fixed in stage $l-1$ or its origin station determined by the planned timetable if $i \notin T_{l-1}$. Meanwhile, the destination station of each train in this stage is set to its final destination predefined in the planned timetable to maintain the feasibility of subsequent stages.

Step 4: Solve the subproblem arising from stage l. We firstly fix the actual arrival time and platform track assignment of each train $i \in T_l \backslash I_l$ at its origin station in stage l to those fixed in stage $l-1$. Then, the simpler subproblem (TRCS) in stage l is solved to optimality or until prescribed termination conditions are met. The resulting solution is denoted as X_l. Note that the boundary conditions between consecutive stages including the earliest arrival and departure time of trains, the occupation of platform tracks and the train coupling states should be strictly respected.

Step 5: Fix the rescheduled timetable in stage l. In X_l, if $c_{im} \leq t^l_{\text{start}} + \tau$, then the related c^*_{im} and v^*_{imk} in X^* are fixed to c_{im} and v_{imk} in X_l, respectively. Meanwhile, f^*_{im} is also fixed if $f_{im} \leq t^l_{\text{start}} + \tau$ holds. Stage l is completed. Note that if all trains have already be considered, then fix all associated decision variables based on X_l.

Step 6: Termination condition. Check out whether all of the arrival and departure time as well the track assignment of train i ($\forall i \in T_l$) at all passed station have be fixed in X^*. If so, add this train to D_l. After update the D_l, if $D_l = T$ (i.e. all operations of trains at all passed stations have been fixed), then a rescheduled timetable is obtained and the rolling horizon algorithm is terminated. Otherwise, we set $l = l + 1$, return to Step 2 and the algorithm continues.

We take the planned timetable in Figure 2 as an example to describe the procedures of our algorithm. For simplicity, Figure 4 only gives the obtained rescheduled timetables arising from 3 stages. Besides, we suppose that the value of σ and τ are 10 and 5, respectively. Thus, in stage 1 shown in Figure 4(b), trains $i_1 \sim i_3$ are firstly picked out as they are about to run at one station or segment within the stage (the start and end time of the stage are expressed by the yellow lines). Then, the origin and destination of all these trains are set to s_1 and s_4 respectively since no arrival time is fixed. Next, the underlying simper subproblem (TRCS) is solved and a rescheduled timetable X_1 for trains $i_1 \sim i_3$ is obtained. Finally, the value of parts of variables is fixed if they do not exceed $t^1_{\text{start}} + \tau$ expressed by the black line. To be specific, we fix specified actual arrival time (including $c^*_{i_1 s_1}$, $c^*_{i_1 s_2}$, $c^*_{i_1 s_3}$ and $c^*_{i_2 s_1}$) and actual departure time (including $f_{i_1 s_1}$, $f_{i_1 s_2}$ and $f_{i_2 s_1}$) to that in X_1. Besides, parts of the track assignment decision should also be determined according to X_1, i.e. the occupation of train i_1 at stations $s_1 \sim s_3$ and train i_2 at station s_1. At this point, we check whether the arrival and departure time as well as the platform track of all trains at all passed stations are fixed. If so, the algorithm is terminated. Obviously, the termination condition is not met and we come to stage 2. In this stage, train i_4 is newly picked out and no train can be added to D_2, i.e. $I_2 = \{i_4\}$, $D_2 = \emptyset$, $T_2 = \{s_1, s_2, s_3, s_4\}$. Note that the route of train i_1 becomes to (s_3, s_4) as $c^*_{i_1 s_3}$ is fixed in stage 1, while the route of other trains is still (s_1, s_2, s_3, s_4). The associated subproblem (TRCS) is then solved to obtain a new rescheduled timetable X_2 in Figure 4(c) and parts of variables are fixed based on the time instant expressed by the black line in X_2. These procedures are executed repeatedly until the termination condition is satisfied. Actually, all trains have been considered after stage 3, thus all the unfixed variables in X^* can be fixed based on X_3 and the algorithm terminates.

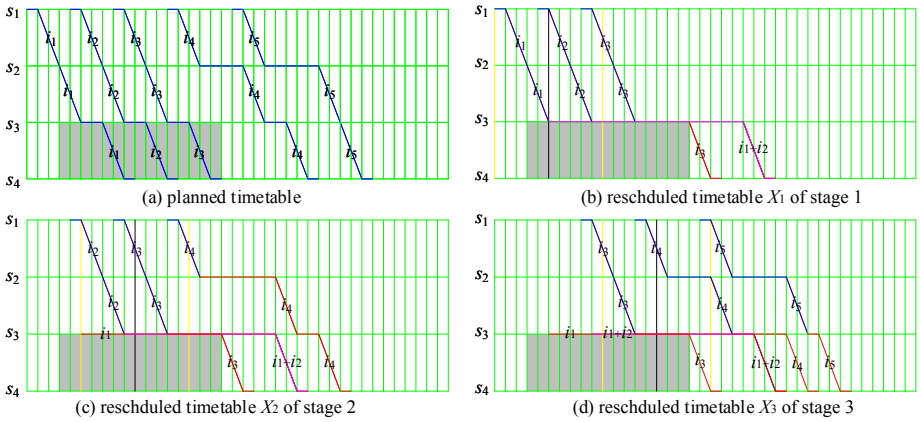


Figure 4: Illustration of the rolling horizon algorithm

5 Computational Tests

We construct realistic instances based on the Wuhan-Guangzhou High-Speed Railway in China to test the effectiveness of the train coupling strategy and the efficiency of our approaches. The train rescheduling model and the rolling horizon algorithm are both coded in MATLAB R2016a, and CPLEX 12.8 is invoked to solve the model, where the parameters of CPLEX are set to their default value.

The computations are executed on a PC with Inter Core i7-7700 3.6 GHz CPU, 16 GB RAM and Windows 10-64 bits operating system. For comparison, the maximum running time of CPLEX is limited to 4 hours. Meanwhile, to satisfy the real-time decision requirement of train rescheduling, the horizon length σ and update step size τ in the algorithm are set as 1 hour and 30 minutes respectively based on our preliminary computational results. The maximum computation time in each stage of the algorithm is limited to 60 seconds to control the total computation time of the algorithm.

5.1 Test Instances and Parameter Setting

The Wuhan-Guangzhou High-Speed Railway line is 1068 km long and it is one of the longest and busiest high-speed railway lines in China. There are 16 stations and 15 segments in total along the downstream direction from Wuhan to Guangzhou of this line at the end of 2016. The location and sketch map of this line are illustrated in Figure 5, where the number in cycles stands for the index of stations, and that in parentheses represents the number of platform tracks at associated stations and the minimum and maximum running time of trains on related segments. For example, the (8,19,24) near station 1 means that there are in total 8 platform tracks in the downstream direction at station 1, while the minimum and maximum running time of trains on segment (1, 2) are 19 and 24 minutes, respectively. Besides, "–" shows that the current station is the end point of the railway line.

The planned timetable utilized in our computational tests is extracted from the actual timetable used from 2015 to 2016 in practice, where only the trains in the downstream direction are adopted. We consider 63 long distance trains that run through the complete

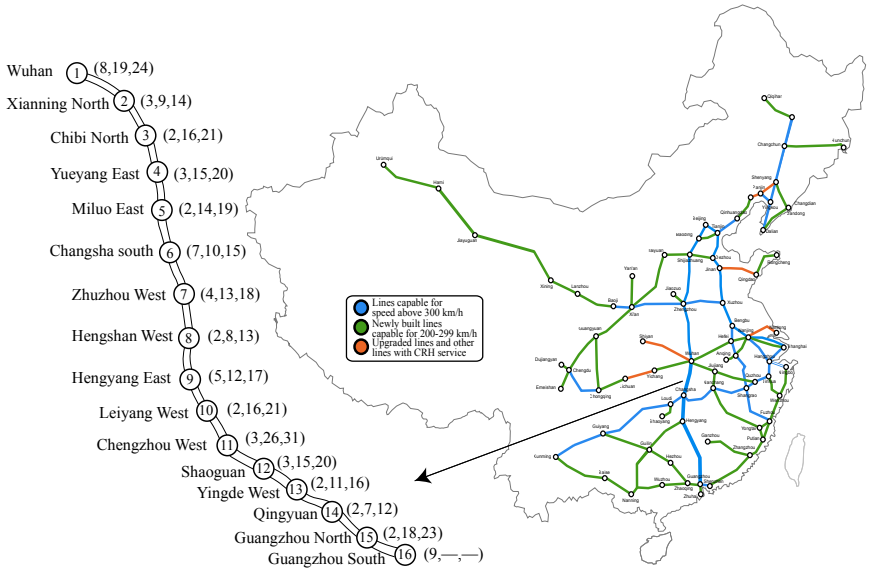


Figure 5: Chinese high-speed railway network and Wuhan-Guangzhou railway line

route from Wuhan Station to Guangzhou South Station, such that the train destination requirement in the coupling rules can be easily satisfied. Besides, the rolling stock type and the formation of some trains are reasonably modified to increase the diversify and applicability of the train coupling in case of complete segment blockages. Specifically, we assume that all trains are served by the same type of 8-carriage rolling stock, such that each two of them can be coupled together at any station passed through by both of the two trains. The considered time span is 6:00-24:00 and the integer time values represent minutes. The associated planned timetable is displayed in Figure 6, where trains are indexed by the sequential order of their planned departure time at their origin station. The planned platform track assignment of trains at stations are not given due to the limitation of space.

To generate representative instances, we firstly construct 3 disruption scenarios according to the location and start time of the disruption: (i) Scenario 1: the disruption occurs at 9:00 and segment (5, 6) is blocked, (ii) Scenario 2: the disruption occurs at 14:00 and segment (9, 10) is blocked, (iii) Scenario 3: the disruption occurs at 19:00 and segment (13, 14) is blocked. We further suppose that the duration of each disruption scenario ranges from 30 minutes to 90 minutes with a fixed increment of 15 minutes. As a result, in total 15 different instances are constructed to test our approaches.

The parameters of the test instances are set as follows. The minimum running time of trains on passed segments and the minimum dwell time of trains at passed stations equal to their predetermined value in the planned timetable. The additional acceleration and deceleration time equal to 2 and 3 minutes, respectively. The maximum running time of each train on each passed segment is set as the minimum value plus 5 minutes. The duration for each station to couple two trains is set as 10 minutes. The arrival, departure, departure-arrival and arrival-departure headway between two consecutive trains not coupled together are set as 3, 3, 2 and 2 minutes, respectively. Finally, we set β to 2 to prevent unreasonable coupling.

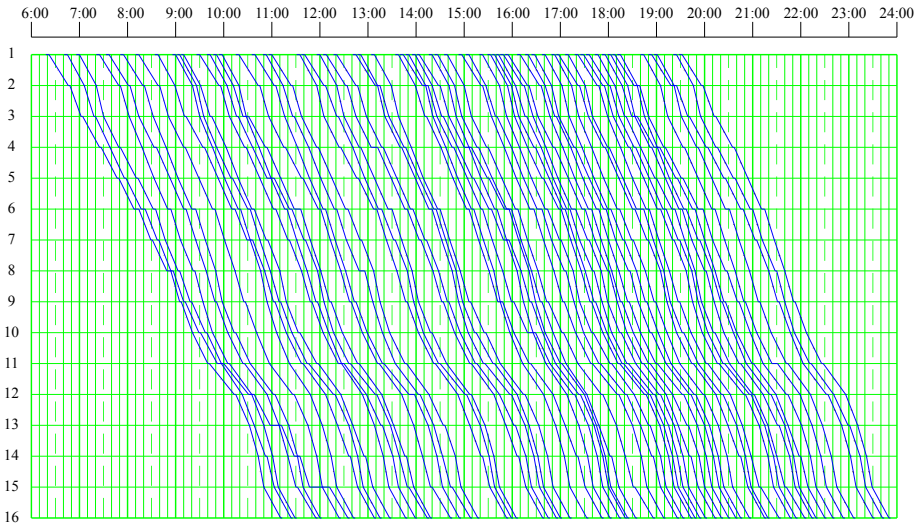


Figure 6: Planned timetable of the test line

5.2 Computational Results

The main results of computational tests are summarized in Table 3, and the meaning of the headers is explained below the Table. Note that the number of variables and constraints in our problem (TRCS) are related to the number of trains, passed stations/segments and platform tracks at stations rather than the disruption instances. Thus, the number of variables and constraints are the same for all test instances. According to CPLEX, in total our model has 354318 constraints and 96940 variables when solving the full problem.

As observed from Table 3, our model (TRCS) can obtain feasible solutions for all instances and in total 4 instances are solved to optimality within the limited time. Note that in the model the convergency rate of the lower bound is much slower than that of the upper bound. Thus the feasible solutions found by CPLEX within 4 hours are likely to be close to the optimal ones. However, the computation time is extremely large especially when the duration of the disruption is long. The average computation time of CPLEX reaches 10885 seconds which obviously does not satisfy the real-time decision requirement. Thus, solving our model directly using commercial solvers is not applicable for large-scale problems due to the real-time requirement of train timetable rescheduling. It is necessary to develop efficient algorithms. Compared to CPLEX, our rolling horizon algorithm can obtain feasible solutions for all instances very quickly. The maximum and average computation time are only 495 and 224 seconds, respectively. The computation time is reasonable for the test instances with such a long time span. Even though the maximum and average relative gaps between the objective value obtained by the algorithm and the best lower bound obtained by CPLEX reach 21.68% and 12.52% respectively, the quality of the solutions found by the algorithm can be improved by 2.13% in average when compared to the solutions obtained by CPLEX within 4 hours. Therefore, our algorithm is capable of solving practical-sized train rescheduling problems incorporating coupling strategy in high-speed railway lines in

Table 3: Summary results of computational tests

Instance[1]	CPLEX									Rolling horizon algorithm								
	OBJ[2]	LB[3]	U_1 /min	U_2	U_3	GAP[4] %	TIME[5] /s	NA[6] /train	NC[7] /train	OBJ	U_1 /min	U_2	U_3	GAP /%	IR[8] /%	TIME /s	NA /train	NC /train
(5,9:00,30)	1776	1776	1726	0	50	0.00	408	3	0	1861	1811	0	50	4.57	-4.92	12	4	0
(5,9:00,45)	3030	3030	2949	0	81	0.00	1802	5	0	3099	3028	2	69	2.23	-2.68	68	5	1
(5,9:00,60)	5049	4287	4929	2	118	15.10	14400	7	1	5122	5069	1	52	16.31	-2.84	249	7	1
(5,9:00,75)	7571	6199	7388	2	181	18.12	14400	8	1	7675	7641	3	31	19.23	-3.42	258	8	2
(5,9:00,90)	10622	8524	10470	3	149	19.75	14400	11	2	10533	10382	5	146	19.07	0.84	495	11	3
(9,14:00,30)	1953	1953	1945	1	7	0.00	1418	5	1	1953	1945	1	7	0.00	0.00	63	5	1
(9,14:00,45)	3392	2965	3362	1	29	12.58	14400	8	1	3395	3364	2	29	12.66	-0.06	278	8	1
(9,14:00,60)	5250	4209	5190	2	58	19.83	14400	8	1	5374	5301	2	71	21.68	-2.14	354	9	1
(9,14:00,75)	6870	6024	6749	4	117	12.32	14400	9	3	6870	6749	4	117	12.32	0.00	302	9	3
(9,14:00,90)	11748	7797	11394	1	353	33.63	14400	22	1	9707	9649	2	56	19.68	15.32	431	12	2
(13,19:00,30)	1080	1080	1074	2	4	0.00	1251	7	2	1088	1086	2	0	0.74	-1.12	64	8	2
(13,19:00,45)	2017	1965	1991	4	22	2.56	14400	10	3	2092	2079	5	8	6.05	-4.42	136	10	3
(13,19:00,60)	3230	2728	3220	6	4	15.53	14400	11	4	3236	3225	6	5	15.69	-0.16	186	11	4
(13,19:00,75)	6137	3817	6053	6	78	37.80	14400	23	4	4727	4687	8	32	19.25	22.57	198	13	5
(13,19:00,90)	7770	5143	7316	8	446	33.81	14400	16	5	6294	6220	8	66	18.29	14.98	262	14	6
Average	5166	4100	5050	3	113	14.74	10885	10.2	1.9	4868	4816	3	49	12.52	2.13	224	8.9	2.3

[1] Instance: the location, start time and duration of the disruption in instances, for example (5, 9:00, 30) means that a disruption occurs at 9:00 and makes segment (5,6) being blocked from 9:00 to 9:30.
[2] OBJ: the objective function value of the best feasible solutions obtained by the CPLEX/algorithm.
[3] LB: the best lower bound obtained by CPLEX within the maximum allowable computation time.
[4] GAP: the relative gap between the objective function value obtained by the CPLEX/algorithm and the related best lower bound.
[5] TIME: the total computation time of the CPLEX/algorithm.
[6] NOA: the total number of trains affected by the disruption.
[7] NOC: the total number of coupled trains.
[8] IR: the relative improvement rate of the objective value obtained by the algorithm to that of CPLEX.

the real-time setting.

It can also be known from Table 3 that the location, start time and duration of disruptions have different influences to the resulting rescheduled timetables. Firstly, the total deviation of arrival and departure time, the total number of affected trains and the associated computation time increase monotonically with the increment of the duration time. Meanwhile, the instances in which the disruption occurs in the segment near to the beginning of the railway line (e.g. Instances 1–5) seem to be easier to solve compared to the instances where the segment in the middle of the line is blocked (e.g. Instances 6–10), since the average computation time are 216 and 285 seconds. The reasons might be explained as follows. Under the former disruptions, many trains are able to be delayed at their actual origin station where much more platform tracks are usually available. Due to that, trains do not need to occupy the somewhat more limited platform tracks at intermediate stations. On the contrary, under the latter disruptions, many trains have already departed from their actual origin station when the disruption occurs. These train have to dwell and wait on a certain platform track at a reasonable intermediate station, making the instances more difficult to solve especially when there are relative few platform tracks at the front station of the disrupted segment. Note that the total deviation under the former disruptions may be worse than that under the latter ones, as trains are likely to be affected at more passed stations in the former cases. Moreover, when the blocked segment is close to the end of the line, the total deviation is likely to be small. However, if the disruption further occurs during the peak hour, much more trains will be affected and more trains could be coupled together to reduce the total deviation due to the restriction of limited station capacity.

5.3 Rescheduled Timetables

We now analyse the detailed rescheduled timetables by adopting the train coupling strategy under a complete segment blockage. For simplicity, we only give the rescheduled timetable of Instance (13, 19:00, 90) as the number of trains affected by the disruption and the number of coupled trains in the instance are both the largest. The rescheduled timetable of the instance is illustrated in Figure 7, where only the trains affected by the disruption are shown. In this Figure, the blue lines mean that the trains run following strictly the planned timetable. The magenta lines represent that the trains couple together at certain stations and pass through associated segments. The red lines indicate that the trains affected by the disruption pass through the associated segments alone. Note that the coupled trains are also affected by the disruption. From this Figure we observe that most affected trains need to dwell at stations 12 and 13 to wait for the recovery of the disruption. Thus, we only provide the rescheduled platform track assignment for the affected trains at these two stations. The rescheduled platform track assignment is depicted in Figure 8, where the left and right margins of the gray rectangles represent the start and end time of the associated trains to occupy the platform tracks, respectively.

Compared with Figure 6, we find in Figure 7 that there are in total 14 trains (i.e. trains 38–51) affected by the disruption, and the total deviation of arrival and departure time reaches 6220 minutes. Other trains are not impacted by the disruption as the buffer time in the planned timetable can relief the propagation of delays. Among the affected trains, train 40 is most heavily influenced and the associated total deviation reaches 763 minutes. Meanwhile, there are in total 6 trains of which the total deviation exceeds 500 minutes, i.e. trains 38–43. Besides, the maximum deviation of a train at a station reaches 196 minutes,

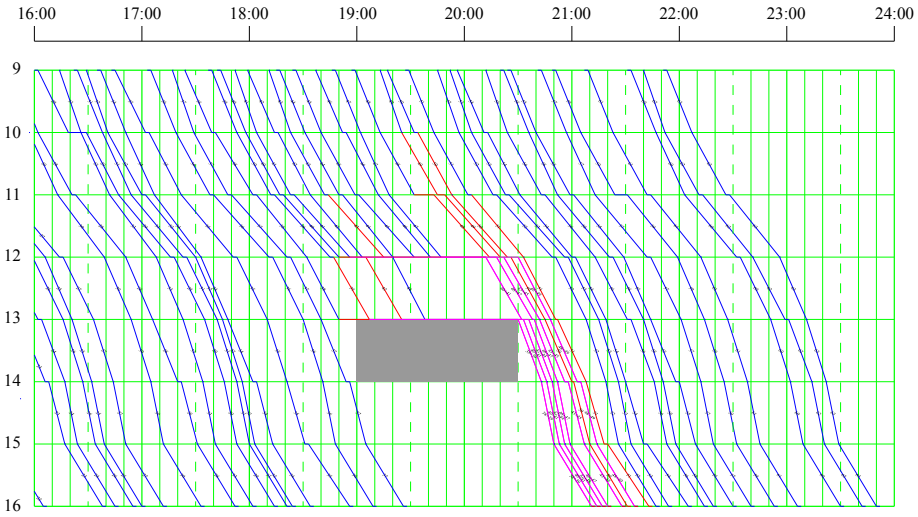


Figure 7: Rescheduled timetable of Instance (13, 19:00, 90)

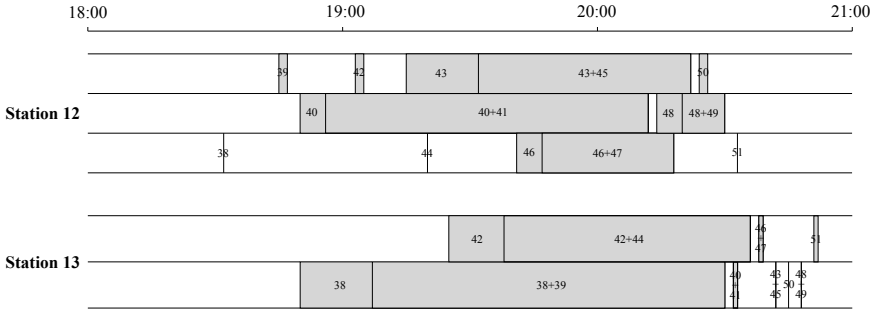


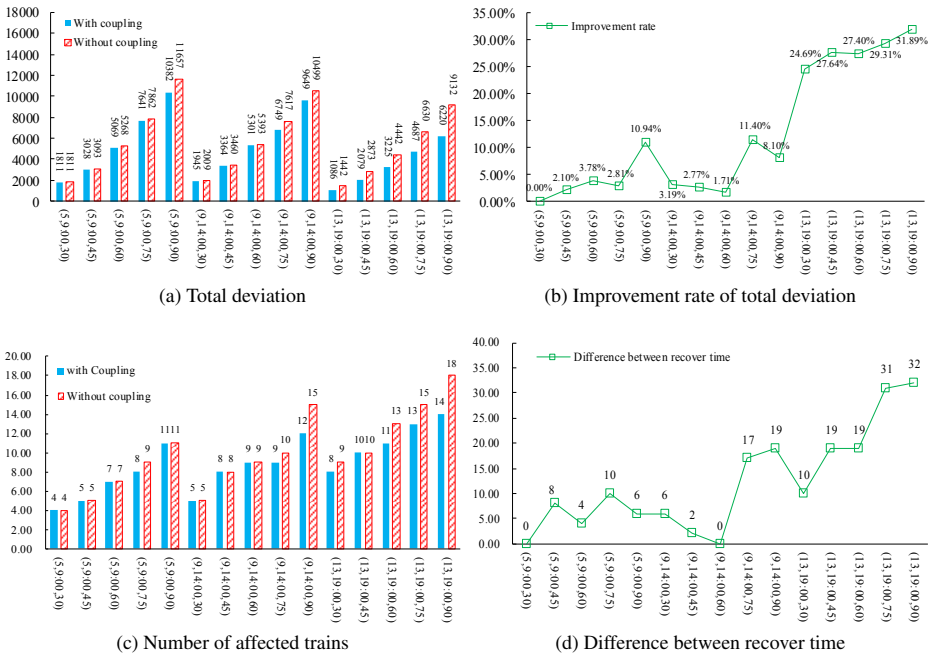
Figure 8: Platform track assignment of Instance (13, 19:00, 90)

leading to a maximum delay of 98 minutes for train 38 at stations 14–16. Regarding to the coupling decision, after the disruption is recovered, trains 40 and 41, trains 43 and 45, trains 46 and 47, and trains 48 and 49 are coupled respectively at station 12, while trains 38 and 39, and trains 42 and 44 are coupled respectively at station 13, such that the limited capacity at stations 12 and 13 and that on segments (12, 13) and (13, 14) are utilized sufficiently to reduce the total deviation. Obviously, most of the coupled trains are composed of consecutive trains except for trains 42+44 and 43+45. The reason might be that train 43 needs to dwell at station 14 based on the planned timetable, thus trains 42 and 43 will be overtaken by trains 44–47 and incur strange train reordering if they are coupled together. Besides, coupled trains 46+47 and 43+45 are swapped after station 12, due to that trains 46 and 47 need not to dwell at stations 14–15 so as to reduce the total arrival and departure deviation. Finally, as shown in Figure 8, every two coupled trains are accommodated on the same platform track at a station, while the departure-arrival headway between two (coupled) trains

occupying the same track is respected strictly. Thus, our algorithm can be used to obtain practically feasible rescheduled timetables and platform track assignments for high-speed railway lines under complete segment blockages.

5.4 Comparison of Rescheduling Strategies

In this section, the rescheduling strategies with coupling and without coupling are tested on all instances to further evaluate the effectiveness of the train coupling strategy. The rescheduling strategy without coupling can be easily realised by fixing all variables x_{ijm} to 0 in advance in the model (TRCS). The two rescheduling strategies are both implemented by our rolling horizon algorithm. The comparison results are illustrated in Figure 9. Figure 9(a) gives the total deviation of all trains at all passed stations under different strategies, while the improvement rate of total deviation by the coupling strategy is shown in Figure 9(b) in which a positive value means that the total deviation with coupling is smaller than that without coupling. The total number of trains affected by the disruption is shown in Figure 9(c). We define that the recover time of timetables equals to the latest departure time of the affected trains at all affected passed stations. The difference between the recover time of timetables without coupling and that with coupling is provided in Figure 9(d) where a positive value means that the recover time with coupling can get earlier than that when coupling is not allowed.



(a) Total deviation

(b) Improvement rate of total deviation

(c) Number of affected trains

(d) Difference between recover time

Figure 9: Comparison results of rescheduling strategies

As seen from Figure 9, compared to the rescheduled timetables where trains are not

allowed to couple, the total deviation and the number of affected trains under the coupling strategy are both reduced. The maximum improvement rate of the total deviation and the maximum decrement of the number of affected trains is 31.89% and 4, respectively. Meanwhile, the recover time of timetables is also likely to be earlier. Besides, the improvement rates are more notable if the disruption lasts for a longer time and occurs at the peak hour with denser traffic volume. Thus, we indicate that the train coupling strategy is promising to reduce the negative influences of large scale disruptions and to relief the propagation of train delays. It could be used as one alternative strategy to reschedule trains in high-speed railway lines in case of complete segment blockages.

6 Conclusions

Real-time train timetable rescheduling under complete segment blockage is of great significance to maintain the operating efficiency and service quality of high-speed railway. Currently, cancelling parts of trains is one of the main strategies to cope with complete segment blockages caused by large-scale disruptions both in academic and in practice, leading to large inevitable negative influences to passengers. Observe that the train coupling strategy gradually begins to be adopted in the daily operations of high-speed railways, this paper aims to explore the effects of this strategy on the real-time train rescheduling, such that the strategy of cancelling trains might be replaced by the better train coupling strategy and the negative influences to passengers can be reduced.

A novel mixed integer linear programming model is firstly formulated to minimize the total deviation of trains' arrival and departure time to that in planned timetable so as to maintain the stability of the timetable as much as possible once a disruption occurs. Meanwhile, strange reordering and coupling decisions are further considered and penalized in the objective function, such that the resulting rescheduled timetables will be more attractive for practical application. Series of operational and safety requirements including the train running and dwelling, train coupling and indispensable headway and station capacity are all considered. The model can be directly solved to find optimal or high quality feasible solutions in short time for small-scale problem instances by state-of-the-art commercial solvers due to its linear feature. To effectively solve large-scale problem instances in real-time setting, a rolling horizon algorithm is developed by utilizing that rescheduled timetables are usually determined in stages in practice. The effects of the proposed approaches are tested on instances generated from the Wuhan-Guangzhou High-Speed Railway in China. Computational results demonstrate that the train coupling strategy is likely to reduce the total deviations and the total number of affected trains. The rolling horizon algorithm can provide high quality rescheduled timetables satisfying the requirement of real-time decisions. Thus, the train coupling strategy is promising in the field of train rescheduling to cope with large-scale segment blockages.

To the best of our knowledge, this paper might be the first one to study the train timetable rescheduling incorporating train coupling strategy in case of a complete segment blockage. We focus on the coupling decisions of trains at stations under practical and safety restrictions, and the decoupling of trains are not taken into account. Thus, the subsequent train coupling rules are strict, making this strategy seeming to be more applicable for dense timetables with a large portion of trains having the same type and route. Therefore, it is valuable to consider the coupling and decoupling of trains simultaneously to extend the application scope of this strategy. Besides, the platform track assignment of trains at stations

is adjusted to assure the practical feasibility of the rescheduled timetables, which may be great different to the planned assignments and increase the operating difficulty of organizing passengers at stations. Thus, it is also significant to consider the stability of the planned platform track assignments in the further study. Finally, we suppose that the end time of the disruption can be predicted in advance and whether the rescheduling of trains should be carried out can be determined in advance. However, it is probably not the case and the uncertainty of the disruption needs to be further considered in the future.

Acknowledgments

We are partially supported by the National Natural Science Foundation of China (No. 61603318, U1834209), and the National Key Research and Development Program of China (No. 2017YFB1200701).

References

Altazin E., Dauzère-Pérès S., Ramond F., et al., 2017. "Rescheduling through stop-skipping in dense railway systems", *Transportation Research Part C: Emerging Technologies*, vol. 79, pp. 73–84.

Cacchiani V., Huisman D., Kidd M., et al., 2014. "An overview of recovery models and algorithms for real-time railway rescheduling", *Transportation Research Part B: Methodological*, vol. 63, pp. 15–37.

Corman F., Meng L., 2015. "A review of online dynamic models and algorithms for railway traffic management", *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 3, pp. 1274–1284.

Fang W., Yang S., Yao X., 2015. "A survey on problem models and solution approaches to rescheduling in railway networks", *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 6, pp. 2997–3016.

Fioole P.J., Kroon L., Maróti, G., et al., 2006. "A rolling stock circulation model for combining and splitting of passenger trains", *European Journal of Operational Research*, vol. 174, no. 2, pp. 1281–1297.

Ghaemi N., Cats O., Goverde R.M.P., 2018. "Macroscopic multiple-station short-turning model in case of complete railway blockages", *Transportation Research Part C: Emerging Technologies*, vol. 89, pp. 113–132.

Krasemann J.T., 2012. "Design of an effective algorithm for fast response to the rescheduling of railway traffic during disturbances", *Transportation Research Part C: Emerging Technologies*, vol. 20, pp. 62–78.

Louwerse I., Huisman D., 2014. "Adjusting a railway timetable in case of partial or complete blockades", *European Journal of Operational Research*, vol. 235, no. 3, pp. 583–593.

Nielsen L.K., Kroon L., Maróti G., 2012. "A rolling horizon approach for disruption management of railway rolling stock", *European Journal of Operational Research*, vol. 220, pp. 496–509.

Peeters M,, Kroon L., 2008. "Circulation of railway rolling stock: a branch-and-price approach", *Computers & Operations Research*, vol. 35, no. 2, pp. 538–556.

Sato K., Tamura K., Tomii N., 2013. "A MIP-based timetable rescheduling formulation and algorithm minimizing further inconvenience to passengers", *Journal of Rail Transport Planning & Management*, vol. 3, no. 3, pp. 38–53.

Törnquist J., Persson J.A., 2007. "*N*-tracked railway traffic re-scheduling during disturbances", *Transportation Research Part B: Methodological*, vol. 41, pp. 342–362.

Veelenturf L.P., Kidd M.P., Cacchiani V., et al., 2016. "A railway timetable rescheduling approach for handling large-scale disruptions", *Transportation Science*, vol. 50, no. 3, pp. 841–862.

Zhan S., Kroon L.G., Veelenturf L.P., et al., 2015. "Real-time high-speed train rescheduling in case of a complete blockage", *Transportation Research Part B: Methodological*, vol. 78, pp. 182–201.

Zhan S., Kroon L.G., Zhao J., et al., 2016. "A rolling horizon approach to the high speed train rescheduling problem in case of a partial segment blockage", *Transportation Research Part E: Logistics and Transportation Review*, vol. 95, pp. 32–61.

AUTONOMOUS FREIGHT TRAINS IN AUSTRALIA

Alexander W Wardrop
Independent Railway Operations Research Consultant
Sydney, Australia
E-mail: awwardrop@gmail.com, Phone: +61 (0) 417 668 092

Abstract
Australia's first autonomous train began running in July 2018. Its running was preceded by extensive trials of both on- and off-train technology. It was not a classic metro train but a 30,000+ tonnes bulk iron ore train, comprising 220-240 wagons, each weighing 130-160 tonnes when laden, and hauled by 2x3280 kW diesel locomotives. This paper discusses the usual rationales for developing autonomous trains and then tests them against the realities of running heavy haul freight trains in remote areas. Any theoretical lack of line capacity is less important than the need for reliable mine-to-port supply chains. Furthermore, mining in remote areas is expensive and increasingly difficult to resource so automation of processes is increasingly attractive to mining companies. The automation of iron ore railway operations beckoned if mining companies could assemble, test and have accepted the various technical building blocks. Pilbara Iron has now completed these steps.

Keywords
Autonomous Trains, Freight Trains, Heavy Haul Railways, Driver Advice Systems

1 Introduction

Australia's first autonomous train began running in July 2018. See Hastie (2018). Its running was preceded by extensive trials of both on- and off-train technology. It was not a classic metro train but a 30,000+ tonnes bulk iron ore train [1], comprising 220-240 wagons, each weighing 130-160 tonnes when laden, and hauled by 2x3280 kW diesel locomotives. The operator was Pilbara Iron, owned by multi-national miner Rio Tinto. Pilbara Iron's railway runs between two ports, Dampier and Cape Lambert, and multiple mines (approximately 13) in Western Australia's Pilbara region. Its principal mainline runs between the Dampier port and the Paraburdoo mine for a distance of roughly 380 kilometres with mines on some branch lines being up to 440 kilometres distant from a port. Figure 1 shows the general Pilbara locality, its ports, mines, railways and roads.

This paper discusses the whys and hows of Australia's first autonomous train running on a freight railway.

2 A General Background to Autonomous Trains

There has to be a rationale for adopting autonomous trains in preference to running manually

operated trains. Typically, a railway might turn to autonomous operation if it needed to increase its throughput beyond what might be possible under manned operation. Because line capacity is a key railway asset, increasing it should increase the numbers of passengers or the amount of freight that could be carried over some reference time period.

However, there will always be limiting factors. Braking distances with respect to maximum permitted speeds generally determine how close trains may run together, either at their free speed or at a restricted speed when closing on preceding trains. On the other hand, any perturbation in the train flow will also reduce the effective train flow. While autonomous operation can eliminate the variability of manual operation, it cannot deal with sources of train flow perturbation that are not related to train driving, such as from station dwell times on passenger railways or junction delays on passenger and freight railways.

Figure 1: Locality Map of the Pilbara Region of Western Australia

Passenger or commodity flows can be increased without increasing train flow if train sizes can be increased. Trains can be lengthened and/or the carrying capacity of carriages or wagons can be increased. However, at some stage, train size must reach its limit.

There are physical and operational constraints that must be respected. For example, passenger trains have to fit inside limiting platform lengths. Growth in the numbers of passengers being transacted through limiting stations will lengthen station dwell times until

train flows are then reduced. The controlling factors are different for freight trains. Increased train weight will eventually exceed the haulage capabilities of their locomotives. Growing train length will eventually debase train braking until increased length is not commercially viable [2]. In any case, infrastructure constraints affect all types of trains.

Nevertheless, passenger railways are more likely to be the beneficiaries of autonomous operation than freight railways because they are more likely to reach their line capacity limits, particularly during urban peak periods, and rarely have the option of increasing train size.

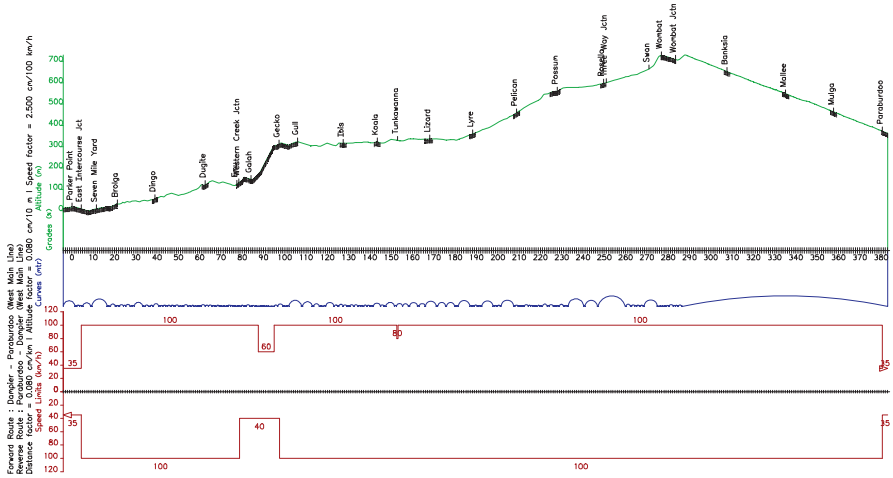


Figure 2: Vertical Profile of the PIRD Mainline between Dampier and Paraburdoo [3]

3 Line Capacity Issues on the Pilbara Iron Rail Division (PIRD)

To address the question of whether line capacity is an issue, I analysed the capacity of PIRD's 2006 mainline network, based on data collected during a 2005 field trip to the Pilbara region. Pilbara Iron then operated two ports: Dampier and Cape Lambert. Originally, Hamersley Iron (HI) ran from Dampier to its Tom Price and Paraburdoo mines while Cliffs Robe River Iron Associates (CRRIA) ran from Cape Lambert to Pannawonica. Eventually, through acquisitions and merger into Pilbara Iron, the two port, railway and mining concerns were connected via a link between Western Creek on the CRRIA mainline and Emu on the HI mainline. Now all mines southeast of the Chichester Range can flexibly dispatch iron ore to either port. By 1978, HI had already duplicated its Chichester Range crossing between Emu and Gull to provide operational flexibility on this difficult stretch of railway with its 2% grade against empty trains, as can be seen in Figure 2. See Hamersley Iron (1978). This gradient also dictates the maximum empty train weight and hence the maximum laden train size.

In 2005, PIRD was steadily extending mainline duplication from Gull to Tunkawanna and onwards to Rosella. PIRD deployed automatic signalling to separate following trains on double track sections and the longer single track sections. The shorter single track sections

were absolute block sections. Originally, HI and CRRIA provided wayside signals to control trains. However by 2005, many of the HI sections were controlled by cab signals with automatic train protection (ATP) [4]. Centralised traffic control was superimposed over the signalling to direct trains into and out of crossing loops and over crossovers on bi-directional double track, although trains would normally take the left hand track on double track, as in the rest of Australia.

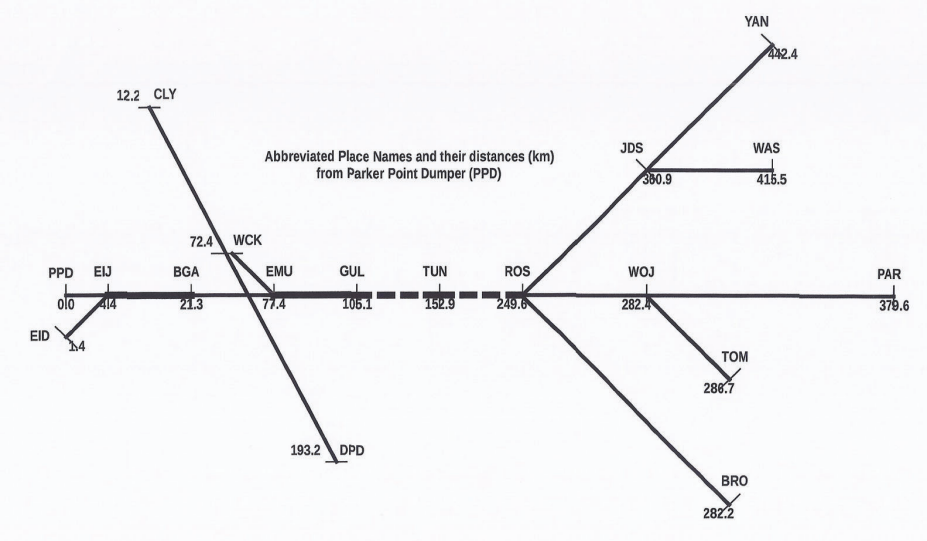


Figure 3: Schematic Network Diagram of PIRD Operations circa 2006

A static capacity analysis of PIRD's operations is presented in Table 1 (refer to Figure 3 for a schematic network diagram). In calculating line capacity, single track and double track sections need to be treated differently. In single track sections, sectional running times for opposing trains plus their respective sectional clearance times are needed to calculate the theoretical capacity for a two-way train flow. In double track sections, the minimum times for trains to clear each signal block joint to its corresponding limit of authority must be calculated to estimate theoretical capacity in each direction of travel. In general, block lengths, like crossing loop siding lengths, are sufficient for a train to stop from a line speed of typically 75 km/h. However, stopping distances are always affected by gradients. The 2% gradient down the Chichester Range is a significant impediment to laden trains, requiring trains to maintain a downhill speed of no greater than 40 km/h (see the bottom speed limit trace in Figure 2).

In 2006, three duplication projects were being considered in the Emu to Rosella section:

- the original Emu to Gull section;
- the then recently completed extension from Gull to Tunkawanna; and
- the intended extension from Tunkawanna to Rosella, the junction for two branch lines.

It can be seen that progressive duplication of the trunk Emu to Rosella section would eventually reduce theoretical line capacity utilisation from nearly 60% to less than 15% with an expected improvement in operational reliability. This is because high utilisation levels are inevitably accompanied by exponentially increasing delays, the more so on single track railways than on double track railways. See Jones & Walker (1973).

Table 1: Static Capacity Analysis of PIRD Operations circa 2006

Line Section	**Len**	**Single Track**				**Double Track**			
(Refer to Figure 3 for locations)	***(km)***	***Dir***	***Cap (t/d)***	***Dmd (t/d)***	***Utl (%)***	***Dir***	***Cap (t/d)***	***Dmd (t/d)***	***Utl (%)***
East Intercourse Jctn (EIJ)-Brolga (BGA)	16.9	1W				OB	465	16	3
		1W				IB	129	13	10
Brolga (BGA)-Emu (EMU)	56.1	2W	61	29	48				
Cape Lambert (CLY)-Emu (EMU)	65.2	2W	46	6	13				
Emu (EMU)-Gull (GUL)	27.7	1W				OB	116	20	17
		1W				IB	117	15	13
Gull (GUL)-Tunkawanna (TUN)	47.8	1/2W	59	34	58	OB	140	20	14
						IB	133	15	11
Tunkawanna (TUN)-Rosella (ROS)	96.7	1/2W	59	32	54	OB	127	18	14
						IB	133	15	11
Rosella (ROS)-Juna Downs (JDS)	111.3	2W	28	15	54				
Rosella (ROS)-Wombat Junction (WOJ)	32.8	2W	49	11	22				
Rosella (ROS)-Brockman (BRO)	32.6	2W	38	6	16				

Notes:
1. Direction of travel (Dir) is either 1-way (1W) outbound from the ports (OB)/inbound (IB) to the ports or 2-way (2W).
2. Theoretical sectional capacity (Cap) has been calculated for a 24-hour period.
3. Actual sectional demand (Dmd) has been obtained from a nominally 25 trains/day timetable.
4. Utilisation (Utl) is the percentage of theoretical capacity consumed by actual demand.

Parallel dynamic capacity analyses, using the SKETCH model, were then undertaken to determine the minimum operational delays that would be visited on an optimal nominal 25 trains per day timetable for the different infrastructure scenarios. For its mathematical basis, see Pudney & Wardrop (2004). Duplication of only the Emu to Gull section would yield a minimum average delay of 13% of bare sectional running times. Extending the duplication from Gull to Tunkawanna would only reduce the minimum average delay to 12%. However, completing the duplication from Tunkawanna to Rosella would drive the minimum average delay down to 7%.

Real railways cannot operate with such theoretical delays. Typically, real operating delays would be double theoretical delays. For example, iron ore railway experience in the Pilbara suggests that real delays would be roughly 20% of bare travel times (see Figure 4 for historical data on the variability of components of HI's train cycle times). Infrastructure changes, such as complete duplication between Emu and Rosella, would thus induce an acceptable level of operational reliability into the mine-to-port supply chain.

Taking into account the above static and dynamic capacity analyses, PIRD operations are clearly not running close to line capacity. Furthermore, any emerging limiting sections are now likely to be branch lines to the mines or to the ports. In both instances, conventional single line section division or duplication would deliver requisite line capacity improvements. Therefore, autonomous train operations are not being pursued on line capacity grounds.

4 Train Size Issues on PIRD

Early Pilbara railway developments from the late 1960s followed contemporary North American practice, including axle loads of the order 32.5 tonnes for locomotives and wagons. Thus, 6-axle locomotives typically weighed up to 196 tonnes in working order and 4-axle ore wagons weighed up to 130 tonnes fully laden. All else being equal, drawgear capacity, locomotive power and the action of direct-release air brakes were the limiting factors. Early trains were hauled by 3x2700 kW locomotives (and were banked by three more locomotives for the first 100 km) and grossed 23,000 tonnes. Since then, much effort has gone into training drivers to avoid breaking drawgear as train weight and length were increased.

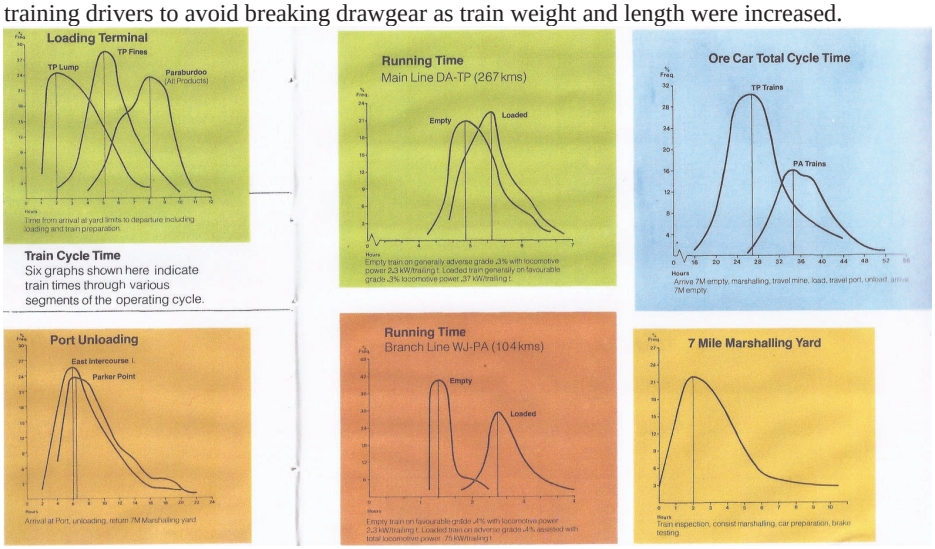


Figure 4: Components of HI's 1978 Train Cycle Times

The original track plant, based on 59 kg/m rail on timber sleepers, proved to be inadequate for trains comprising the above rolling stock, so it was gradually replaced with conventional 68 kg/m rail and closer-spaced timber sleepers. Given the steady increase in annual tonneage, the track plant has since been upgraded to head-hardened 68 kg/m rail and concrete sleepers.

Progressive improvements in train technology have lead to the adoption of roughly 3280 kW locomotives with higher factors of adhesion, so that trains can now gross over 30,000 tonnes. Behind these statistics are improvements in locomotive and train braking. Locomotives now have extended range dynamic brakes and wagons are now fitted with ECP brakes [5]. The adoption of locomotives with ac traction motors offers further scope for operational improvement. However since the major haulage task is predominantly downhill, braking performance, rather than starting tractive effort, is probably more important. Since SD70ACe (EMD) and ES44ACi/ES44DCi (GE) locomotives entered service from the mid 2000s, there have been no significant changes in train size on the original iron ore railways.

However, the recent mining entrant, FMG, directly adopted 40 tonne axle load wagons and built loading and unloading infrastructure to match. This encouraged the established miners to also move towards 40 tonne axle load ore wagons. However, larger wagons still have to fit through existing wagon dumpers [6].

The conclusion that can be drawn is that PIRD's train size has plateaued within locomotive tractive effort and drawgear constraints. The only way to haul more iron ore is to run more trains, needing more drivers.

5 The Pilbara Mining Back Story

The way mining is carried out in remote areas is the key to understanding PIRD's move towards autonomous trains.

The Pilbara region is roughly two hours flying time from Perth, already one of the world's most remote cities. The cost of labour weighs heavily on mining, especially for low value, high quantity commodities, such as iron ore. Originally, the mine workforce lived on site in the Pilbara. However increasingly, resident employees are being converted to fly-in-fly-out (FIFO) employees [7] because companies can save accommodation costs by placing them in short-stay camps.

Large quantities have to be mined and shipped to distant customers. Large quantities require matching infrastructure, such as:

- ocean ports capable of handling ships up to 320,000 DWT;
- sufficient berths to handle such ships for different miners;
- stockyards behind the berths for cargo assembly;
- wagon unloaders and conveyor runs to feed the stockyards;
- railway lines bringing laden trains into each port;
- a range of mines to exploit different ore bodies and to facilitate ore blending to create a consistent iron content; and
- wagon loaders to efficiently place ore in trains.

Consequently, unit costs of production, including the amortisation of this entire infrastructure, have to be driven down. Therefore, mining and transportation costs are being attacked through automation.

Train loading and unloading were early candidates for automation. Train loading can be carried out by hauling trains at a fixed low speed under bins or through load-out tunnels, usually without a driver. Alternatively, empty trains can be indexed [8] under bins, without attached locomotives. All laden trains in the Pilbara are emptied by being indexed through tipplers, without attached locomotives.

Pilbara Iron led the way with the early automation of ore haulage from the mine face to the train loader. This was to reduce costs, introduce electrification of the haulage trucks and to improve mine safety by avoiding haul road accidents. Some miners are even taking automation directly to the mine face. See Lucas (2018).

PIRD has been pursuing mainline railway automation, not specifically to increase throughput, but to provide itself with train scheduling flexibility. The rostering of mainline train drivers was quite rigid to ensure that drivers could be employed productively. Amongst other practices, drivers were rostered to swap in the field between outbound empty trains and

inbound laden trains, so that they could work out-and-back from a home location. If trains could not be despatched on time, for whatever reason, drivers' shifts could be wasted.

The clear object of the automation of mainline train operations was to be able run trains when they were ready and then to run them straight through from port to mine or vice versa, subject only to the usual traffic delays. There would still be a need for manual locomotive driving to position trains for loading and unloading, to fuel and provision locomotives and to move locomotives or wagons to and from maintenance and repair facilities.

6 Requisite Steps for Automating Mainline Train Operations

It is important to understand that PIRD's trains are autonomous, rather than remotely controlled. Thus, they need an on-board intelligence to be able to drive themselves within their limits of authority and within prevailing geographic constraints of gradients, curvatures and speed limits. Their progress would be monitored but they realistically could not be remotely driven, because the problems of rostering drivers would then be transferred from the field to a central office.

There appear to be a number of actions (although not an exhaustive list because PIRD have not disclosed their actual implementation) that had to be undertaken prior to running an automated mainline railway:

- provision of direct communications between a remote control office and each train;
- location of each train to be determined by GPS;
- provision of forward-facing CCTV and proximity sensors on each train with a continuous feed back to the control office;
- setting of each train's route from the control office;
- setting of each train's limits of authority from the control office and/or local automatic signalling and delivering these limits of authority to each train;
- delivering approaching track alignment and speed limit knowledge to each train;
- development of suitable driving control systems;
- control of traction and dynamic (electric) and air braking on each train; and
- provision of precise control of air brakes down the length of each train.

PIRD was an early adopter of centralised traffic control, whereby routes could be remotely set throughout its mainline network. Limits of authority were initially set by wayside signals, but these have been progressively replaced by cab signals, whose aspects can be transmitted to trains via coded track circuits. The control office, as well as the passage of trains, can set home signals (or their equivalent). The passage of trains alone can set automatic signals. Originally, the control office was adjacent to PIRD's mainline near Karratha. Nowadays, it is located in Rio Tinto's Perth offices.

Track circuiting provides positive detection of the presence and absence of trains but cannot precisely locate trains. However, coded track circuits, as adopted by PIRD, can continuously provide trains with their limits of authority, in a wholly distributed manner. GPS can continuously locate trains, but not to a precision that can prevent collision.

Geographic knowledge can be transmitted to trains in a number of ways [9]. In PIRD's case, alignment data is transmitted en route through passive on-track balises. The alignment data is then assembled on board the train so that driving calculations could be made to manage

the train. Initially, PIRD used the alignment data to calculate the ATP braking trajectory that each train would run within, as it was reaching the limit of its current authority. However with autonomous train operation, complete power, hold, coast and/or brake trajectories must be calculated by the on-board driving control systems. See Albrecht et al (2016) for the basis behind optimal train control.

The driving of long and heavy freight trains, such as PIRD's ore trains, not only has to continually set the train's operational mode (ie power, hold, coast or brake), but also has to handle in-train forces. A long-standing problem with iron ore trains, particularly those with direct release air brakes, was the creation of force shock waves running backwards and forwards along the length of a train. If these forces became too intense trains would break apart. PIRD accordingly developed a driving simulator to train its drivers in the appropriate handling of long trains. Thus, long trains had to be kept stretched under traction and braking to avoid running in and out and generating shock waves along the train.

While long freight trains were fitted with direct release air brakes [10], control of in-train forces was an issue that was only handled by drivers with route knowledge, reinforced with simulator training. This was an impediment to automated driving because of the need to continuously estimate the magnitude of in-train forces and then manipulate the air brakes to keep the train stretched. There was also the issue that classic air brakes took a significant time to apply and release because of the speed at which service brake applications propagated down the length of a train.

However, PIRD has now fitted ECP to its ore trains. ECP allows air brakes to be graduated on and off, simultaneously down the length of a train, while keeping the train stretched train. Dynamic braking now supplements air braking when a train's downhill speed needs to be held against a grade, rather than vice versa with direct release air brakes.

The installation of ECP was an important step towards implementing automated train driving. It allowed all directional, traction and braking decisions to be made by driving control systems. When complemented by electronically available alignment data and limits of authority, all the technical building blocks for running autonomous trains were now available.

7 Running Autonomous Trains in Remote Areas

The Pilbara region of Western Australia is semi-arid and sparsely populated. Various iron ore railways pass from the Indian Ocean coast to the highlands where much of the region's iron ore is mined. The railways are unfenced, are typically paralleled by gravel access roads, are crossed by watercourses and are intersected by road/rail level crossings, most of which are only passively protected. Cattle and native animals may venture onto railway reserves in search of feed. People in cars and trucks often drive along the access roads and cross the railways in their travels. There is an ever present, but low, risk of collision (more likely with intruders than other trains), irrespective of whether trains are manually or automatically driven. By their nature, trains can warn intruders but cannot take evasive action.

The biggest non-technical issues for running autonomous trains are the detection and mitigation of accidents when trains break down, collide or derail. PIRD already runs a highly instrumented railway to identify train health issues before they lead to failure. However, such measures cannot detect collisions with people, road vehicles, landslides or washouts. PIRD

will have placed detection devices, such as CCTV and presence detectors, on board its autonomous trains, not so much to avoid collision, as to detect it.

There is still the need to physically respond to incidents that stop autonomous trains. The issue with autonomous train operation will be how long it will take the central office to be aware of an incident. Necessarily, response times to reach failed trains will still be long and the means of recovery will still be variable. However, the commercial pressures of maintaining the mine-to-port supply chain mean PIRD will have to evolve suitable processes from those already in place for manned train operations.

8 Conclusions

PIRD's pursuit of autonomous mainline railway operations has been long and deliberate and should be seen in the context of a general automation of mining, particularly in remote areas for which it is difficult to recruit suitably skilled employees. In itself, the running of autonomous trains will not immediately lead to running more trains. However, the combination of remote train despatching and on-board driving control systems should lead to better timekeeping of individual trains, better control of train flows and more intense use of the railway.

PIRD gradually assembled the technical building blocks of:

- centralised and remote route setting;
- the setting of limits of authority for individual trains;
- implementation of competent ATP;
- conversion of train air brakes to ECP; and
- development of on-board driving control systems and their integration with locomotive control systems.

Actual field-testing and the regulatory authorisation of autonomous train operation took more time. Now autonomous trains are running in revenue service. The exciting future prospect is that what has been applied on remote heavy haul freight trains could also be applied to suburban passenger trains.

References

Albrecht A, Howlett P, Pudney P, Xuan Vu & Peng Zhou 2016, *The key principles of optimal train control – Parts 1 and 2*, Transportation Research Part B 94, Elsevier, pp 482-508 and pp 509-538.

Hamersley Iron 1978, *Hamersley Iron Railways*, Hamersley Iron Pty Ltd

Hastie H 13 July 2018, *Rio Tinto autonomous train in the Pilbara makes first iron ore delivery*, Sydney Morning Herald/Business.

Jones JCM & Walker AE July 1973, *The Application of Models of Single Railway Track Operation to Evaluate Upgrading Alternatives*, Rail International, pp 787-801.

Lucas J 2018, *Aussies have eyes on world's first fully automated underground gold mine in Africa's Mali*, ABC Goldfields (ABC News), accessed 9 August 2018.

Pudney P & Wardrop A 2004, *Generating Train Plans with Problem Space Search*, Computer Assisted Scheduling of Public Transport, San Diego.

End Notes

1. While there have been autonomous trains since the 1920s, the British Post Office railway under central London being a notable example, it is most likely that the first large autonomous freight train ran on the Black Mesa and Lake Powell Railroad in 1973. This train hauled coal between a mine and a power station in a closed operation, using remotely controlled electric locomotives.
2. On 21 June 2001 BHP Iron Ore ran the largest ever freight train, grossing 99734 tonnes and extending 7300 metres, between Yandi and Port Hedland in the Pilbara (See Railway Gazette 1 August 2001). However, train traction and braking was unstable, notwithstanding multiple locomotives being distributed throughout its length.
3. The grading line is shown in green, the heading line is shown in blue, the distance baseline (in kilometres) is shown in black and directional speed limits are shown in red. Crossing loops and duplicated track are highlighted as black bands under the grading line.
4. Block sections were rationalised under cab signalling to reduce the wayside plant. The combination of static balises, to provide track geometry data, and signal aspects via coded track circuits, to give the limits of authority, allowed trains to calculate their braking trajectories over long sections to each limit of authority.
5. Electronic Control of Pneumatic brakes (ECP) is a freight train air brake technology, which permits air brakes to be simultaneously applied or released in a graduated manner down the length of a train, regardless of length.
6. Pilbara practice is to rotate wagons (tipple) to empty them because they are simple gondolas without bottom doors. The tipplers were built to handle wagons of fixed lengths. Typically, pairs of wagons are inverted to dump the ore into under-track hoppers with the ore being carried away by conveyors to stockpiles.
7. Fly-in-fly-out (FIFO) is a remote employment practice whereby employees are flown from their homes, say in Perth, to a mine, say Yandicoogina, and employed for, say, two weeks straight on long shifts before being flown home to rest for one week.
8. An indexer is a lineside mechanism, which accurately steps a train of wagons, one or two wagons at a time, through a loader or unloader.
9. As an example, TTG's Energymiser train driving advice system downloads complete alignment data for a journey, which is then consumed during that journey.
10. Note that on PIRD single-pipe direct release air brakes were fitted, whereby auxiliary brake reservoirs were recharged over the same pipe as the air pressure braking signals were sent, so that recharge and application had to take place sequentially.

Train-set Assignment Optimization with Predictive Maintenance

Meng-Ju Wu [a] and Yung-Cheng (Rex) Lai [a,1]
[a] Department of Civil Engineering, National Taiwan University
Room 313, Civil Engineering Building,
No. 1, Roosevelt Road, Sec. 4, Taipei, 10617 Taiwan
[1] E-mail: yclai@ntu.edu.tw, Phone: +886-2-3366-4243

Abstract
The efficiency of rolling stock utilization is an important objective pursued in practice. Rolling stock assignment plan including the assignment of utilization paths and maintenance tasks. Previous studies have adopted the fixed periodic maintenance (PM) strategy; however, the difference in the reliability of rolling stock is not considered. Maintenance planners have to manually adjust utilization and maintenance tasks on the basis of experience. Consequently, this study proposes an optimization process for assigning rolling stock to utilization paths and maintenance tasks in accordance with the predictive maintenance strategy (PdM) with trainset-specific reliability models. Results of the empirical study demonstrate that the developed process with PdM can assign utilization paths and schedule maintenance tasks to each trainset efficiently and reduce the total cost by over 14% compared with the PM-only strategy. Adopting this process can help planners improve the efficiency and reliability of rolling stock utilization.

Keywords
Train-set assignment, maintenance scheduling, and predictive maintenance

1 Introduction

Train-set is an expensive asset of a railway system (Caprara et al. (2007); Cheng (2010)). Taiwan Railways Administration (TRA), manages and maintains a number of train-sets through train-set assignment, which includes the assignment of utilization paths and schedule of maintenance tasks. In practice, maintenance scheduling is performed with periodic maintenance (PM) strategy. For train-set of the same type, a fixed set of rules is applied to all of them because their quality and performance are supposed to be similar. However, the reliability of each train-set is actually unique and may differ. Previous studies have adopted the fixed PM strategy for the train-set assignment problem (Yun et al. (2012); Li et al. (2016)) but maintenance intervals cannot be flexibly adjusted according to the difference in train-sets. Although a few studies have considered the reliability of train-sets, maintenance thresholds remain fixed without any flexibility (Moghaddam and Usher (2011); Asekun (2014)). To perform effective train-set maintenance scheduling, researchers proposed the predictive maintenance (PdM) strategy, such as wheelset maintenance (Li et al. (2014)). Other studies have adopted PdM in train-set maintenance by assuming a fixed degradation rate (Herr et al. (2017)). These studies resulted in a local optimum rather than the global optimum.

With the rise of big data analysis and artificial intelligence (AI) techniques, a train-set

specific reliability model can now be obtained from reliability and maintenance data over time. We propose an optimization process for assigning train-set to utilization paths and scheduling maintenance tasks in accordance with the PdM strategy with train-set specific reliability models. Using this process can help planners evaluate the trade-off between reliability and cost.

2 Train-set Assignment and Maintenance Problem

The train-set assignment plan of TRA includes the assignment of utilization paths and maintenance schedules in accordance with utilization schedule (demand) and maintenance requirements. A utilization schedule contains a set of utilization paths created based on a timetable. Each utilization path identifies the ideal type and amount of train-set to meet the demand. However, if a particular type of train-set is unavailable, an alternative type of train-set can be used subject to a penalty cost (i.e., replacement cost) due to the difference in seat arrangements.

Table 1 presents the maintenance rules of commuter train-sets at TRA. The rules include four levels, namely, daily maintenance (DM), monthly maintenance (MM), bogie maintenance (BM), and general maintenance (GM). Fixed thresholds by accumulative operating days are adopted by these PM rules. High maintenance levels (BM and GM) are scheduled in advance for each train-set. These levels require longer maintenance times and consider the limited workshop capacity. By contrast, low maintenance levels (DM and MM) must be considered during the assignment at the operational level along with restrictions on maintenance location and capacity. The DM process takes approximately an hour whereas MM requires a day and thus cannot be performed during the connection or an overnight period in a utilization path. The maintenance tasks of high maintenance levels include all maintenance tasks in low maintenance levels; therefore, after one class of maintenance process, all accumulative operating days of the executed maintenance level and the corresponding low maintenance level return to zero. Previous studies have improved the efficiency of train-set usage. Their processes do not consider train-set specific reliability. Hence, this research examines the possibility of PdM strategy in this process and its potential benefit.

Table 1: Maintenance regulations in TRA

Maintenance level	Accumulative operating days	Maintenance location
DM	3 days	Train-set depot
MM	3 months	Train-set depot
BM	3 years	Workshop
GM	6 years	Workshop

3 Methodology

According to literature (Kaczor and Szkoda (2016); Yin et al. (2017)), a two-parameter Weibull distribution is suitable for describing the degradation of a train-set. Figure 1 lists the input, output, and consideration of train-set assignment planning. With the input regarding train-set and maintenance, this process assigns train-set to utilization paths and maintenance tasks by considering the costs, reliability, and efficiency of the utilization.

These objectives can be attained by minimizing the maintenance costs and expected costs of failure. Efficiency of utilization can also be ensured by the minimization of the MM cost because the less frequent MM is, the better the train-set availability is.

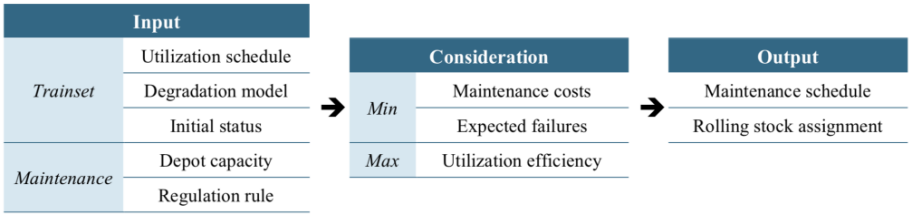


Figure 1: Input and output of train-set assignment planning

To identify and assign appropriate maintenance tasks to each tension section for a year according to reliability evaluation, a mixed-integer programming (MIP) model is formulated by minimizing expected cost of failures, expected cost of operation loss, and expenditure on maintenance. Inspection tasks are ignored in this model because they are not scheduled in the annual maintenance planning process.

I denote the set of all partial utilization paths; I^A, I^B, I^O and I^I are the subset in J that represents DM paths, MM paths, operational paths and starting partial paths; K denote the set of all time intervals and K^I is the subset of all starting time intervals for partial path I^I; S denotes the set of all stages that discretize accumulative days; U denotes the set of types of train-set; V denotes the set of all available train-sets.

C^D, C^M and C^H represent the cost for DM, MM, and train-set replacement; D^A denotes the accumulative operating days upper bound for DM; F denotes the expected cost of failures; $F_{v,s}$ denotes the discretized expected number of failures on train-set v in each stage s; G denotes the MM capacity in the depot; $N_{u,i}$ denotes number of train-sets of type u required in partial path i; P denotes number of discrete stages in each period; P_i denotes operation time of partial path i; Q denotes minimum number of times of MMs need per week; S_i denotes mileages of partial path i; M denotes the relatively large positive number; W denotes the relatively small positive number ensuring that all accumulative values return to zero.

$d^A_{v,k}$ and $d^B_{v,k}$ are non-negative integer indicates the DM and MM accumulative operating days of train-set v at the end of time interval k; $f_{v,k}$ is non-negative integer denotes the expected number of failures on train-set v of time interval k; $q_{u,i,k}$ is binary integer that indicates whether the partial path i in time interval k is operated by type u or not; $x_{i,v,k}$ is binary integer that expresses whether train-set v operates partial path i in time interval k or not; $z_{v,k}$ and $r_{v,k}$ are binary variables that denote whether train-set v executes DM or MM in time interval k or not; $\theta^+_{v,k,s}$ and $\theta^-_{v,k,s}$ are auxiliary binary variables that linearize the nonlinear function.

The MIP model is as follows:

Objective function

Min

$$C^D\sum_{v\in V}\sum_{k\in K} z_{v,k} + C^M\sum_{v\in V}\sum_{k\in K} r_{v,k} + C^H\sum_{u\notin U^I}\sum_{i\in I}\sum_{k\in K} q_{u,i,k}$$
$$+F\sum_{v\in V}\sum_{k\in K} f_{v,k} + W\sum_{v\in V}\sum_{k\in K}(d^A_{v,k} + d^B_{v,k}). \quad (1)$$

Equation (1) minimizes the total cost of train-set assignment, including the DM cost, MM cost, replacement cost (due to undesired train-set), expected cost of failures, and accumulative duration variable of the utilization path that returns to zero when the corresponding maintenance tasks are executed. The reliability of the utilization is governed by the minimization of the expected cost of failures, which is computed as the sum of ticket refund loss and emergency maintenance cost (= expected number of failures × cost of minimum repair).

Assignment Constraints

To satisfy the demand train-set (utilization paths), train-set assignment constraints are presented in the following equations.

Subject to

$$\sum_{u \in U} q_{u,i,k} = 1. \qquad \forall i \in I^I, k \in K^I \qquad (2)$$

$$\sum_{v \in V^u} x_{i,v,k} = N_{u,i} q_{u,i,k}. \qquad \forall u \in U, i \in I^I, k \in K \qquad (3)$$

$$\sum_{i \in I} x_{i,v,k} \leq 1. \qquad \forall v \in V, k \in K^I \qquad (4)$$

$$(n-1) x_{i,v,k} = \sum_{h=1}^{n-1} x_{i+h,v,k+h}. \qquad \forall i \in I^O, v \in V, k \in K \qquad (5)$$

$$x_{i,v,k} = x_{i+1,v,k+1}. \qquad \forall i \in I^O, v \in V, k \in K \qquad (6)$$

Equations (2) and (3) ensure that every starting operational partial path satisfies the required type and amount of train-set. Equation (4) guarantees that each train-set can only be assigned to one path at most. The starting partial paths of utilization paths are considered in Equation (5) due to the multiple-day paths in TRA. Equation (5) ensures that all partial paths of incomplete utilization paths are correctly connected, and it works with Equations (2) and (3) to complete the complicated multiple-day path assignment. For example, when a one-day path with two-time intervals (either morning–evening or evening–morning) is encountered, Equation (5) can be expanded as Equation (6), as shown in Figure 2.

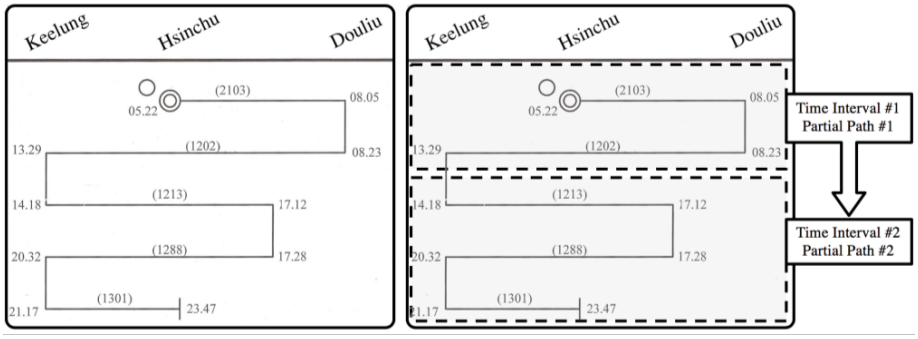


Figure 2: Time interval and partial path
(left, origin; right, divided into time intervals)

Maintenance-related Constraints

Equations (7) and (8) deal with the accumulative times of train-set that should return to zero after executing the maintenance tasks. Equation (9) ensures the DM regulation of train-set in accordance with the PM requirement. Equation (10) ensures that the train-set executing the DM task cannot be assigned to the operational paths, and Equation (11) is for the MM task. Equation (12) ensures that the amount of MM does not exceed the depot capacity. Equation (13) guarantees the minimum number of MM tasks to avoid over-concentrating the train-set executing the MM task or wasting the human resource of the maintenance crew. The MM task requires a full day. Thus, Equation (14) ensures that the MM task starts off the day.

$$d^A_{v,k} \geq d^A_{v,k-1} + \sum_{i \in I} P_i x_{i,v,k} - M r_{v,k} - M z_{v,k}. \qquad \forall v \in V, k \in K \tag{7}$$

$$d^B_{v,k} \geq d^B_{v,k-1} + \sum_{i \in I} P_i x_{i,v,k} - M r_{v,k}. \qquad \forall v \in V, k \in K \tag{8}$$

$$d^A_{v,k} \leq D^A. \qquad \forall v \in V, k \in K \tag{9}$$

$$\sum_{i \in I^A} x_{i,v,k} \geq z_{v,k}. \qquad \forall v \in V, k \in K \tag{10}$$

$$\sum_{i \in I^B} x_{i,v,k} \geq r_{v,k}. \qquad \forall v \in V, k \in K \tag{11}$$

$$\sum_{v \in V} r_{v,k} \leq G. \qquad \forall v \in V, k \in K \tag{12}$$

$$\sum_{v \in V} \sum_{k \in K} r_{v,k} \geq Q. \tag{13}$$

$$r_{v,k} = r_{v,k+1}. \qquad \forall v \in V, k \in K^I \tag{14}$$

Reliability-related Constraints

Train-set specific degradation models are presented in the form of reliability functions. To transform the nonlinear Weibull distribution to linear parameters, the nonlinear relationship is discretized into stages. Equation (15) indicates that for all stages, only one stage can make $\theta^+_{v,k,s}$ equal to 1 only when the MM accumulative days of train-set fall within *s* and*s-1*. Then, $\theta^+_{v,k,s}$ and $\theta^-_{v,k,s}$ are equal to 1 and 0, respectively. Equation (16) ensures that at every stage, only one stage can make $\theta^+_{v,k,s}$ equal to 1. Equation (17) sets up $\theta^-_{v,k,s}$ to control the invalid situations in Equation (15) with a large *M*. Equation (18) obtains the expected number of failures from the different stages. Equations (19) and (20) describe the properties of the variables. Three positive and five binary variables are available in the proposed optimization model.

$$-M\theta^-_{v,k,s} + \frac{s-1}{P}\theta^+_{v,k,s} \leq d^B_{v,k} \leq M\theta^-_{v,k,s} + \frac{s}{P}\theta^+_{v,k,s}. \qquad \forall v \in V, k \in K, s \in S \tag{15}$$

$$\sum_{s \in S} \theta^+_{v,k,s} = 1. \qquad \forall v \in V, k \in K \tag{16}$$

$$\theta^+_{v,k,s} + \theta^-_{v,k,s} = 1. \qquad \forall v \in V, k \in K, s \in S \tag{17}$$

$$f_{v,k} = \sum_{s \in S} \theta^{+}_{v,k,s} F_{v,s}. \qquad \forall v \in V, k \in K \tag{18}$$

Variable Domain

Equations (19) and (20) describe the properties of the variables. Three positive and five binary variables are available in the proposed optimization model.

$$d^{A}_{v,k}, d^{B}_{v,k}, f_{v,k} \geq 0. \tag{19}$$

$$r_{v,k}, x_{i,v,k}, z_{v,k}, \theta^{+}_{v,k,s}, \theta^{-}_{v,k,s} \in \{0,1\}. \tag{20}$$

In practice, a train-set assignment plan is determined daily for the following seven days. Therefore, a rolling horizon process is also developed to implement the proposed train-set assignment optimization model. The lengths of the decision and implementation horizons are two decisions that should be decided for the process. The first decision (the length of the decision horizon) should consider solution quality and computational time. The other decision (implementation horizon) is based on the degree of uncertainty in train-set availability. A short implementation horizon is usually better than a long one due to the increase in flexibility.

4 Case study

This study applies the process in Hsinchu depot of TRA. The MIP model is coded in Python environment with Gurobi solver. Hsinchu depot mainly manages commuter trains. 11 multi-day utilization paths are present in the utilization schedule, and they have to be fulfilled by 6 sets of EMU500 and 40 sets of EMU700 trains. EMU500 can operate as single or double train-sets depending on the utilization path, and EMU700 often operate as a pair of two train-sets. Table 2 shows detailed information on the utilization paths in Hsinchu depot. To demonstrate the benefit of adopting PdM in MM, we set the planning horizon to 180 days. Parameters are obtained or estimated from Railway Reconstruction Bureau and TRA.

Table 2: Utilization paths for EMU500 and EMU700 train-sets at Hsinchu depot

Path No.	Required type	Required quantity	Accumulative operating days	Operating frequency
E5	EMU500	1	1	Every day
E6	EMU700	2	3	Every day
E7	EMU700	2	3	Every day
E8	EMU700	2	2	Every day
E9	EMU700	2	4	Every day
E10	EMU700	2	3	Mon, Tue, Fri, Sat, Sun
E10_1	EMU700	2	2	Wed, Thu
E11	EMU700	2	2	Every day
E12	500/700	2	2	Mon, Tue, Wed, Thu, Sun
E12_1	500/700	2	1	Fri, Sat
E13	EMU500	1	1	Every day

Table 3 presents the results with the "PM-only" strategy (DM and MM are scheduled based on a fixed threshold) and with the "PM + PdM" strategy (DM via a fixed threshold/MM via a PdM strategy). Train-set assignment under the PM + PdM strategy provides a lower total cost than that under the current PM-only strategy. Especially the outcomes in the expected cost of failure are different because the PM + PdM strategy considers the degradation model of each train-set and failure cost as opposed to treating all train-sets with a fixed set of maintenance thresholds. These degradation models provide additional information regarding the reliability of each train-set. As a result, the maintenance cost under the PM + PdM strategy is reduced by 4.59%, a saving from the increase of the MM interval mainly for the EMU700 train-sets due to their better reliability performance. The expected cost of failures from the PM+PdM strategy also outperforms the PM only strategy because reliability and their expected cost of failures were considered in the proposed model.

Table 3: Comparison of assignment result

Model	**PM only**	**PM + PdM**	**% difference**
Number of DMs	1,433	1,431	
Number of MMs before PM	18	12	
Number of MMs at PM	32	2	
Number of MMs after PM	0	32	
Expenditure on maintenance	311,750	298,070	-4.59%
Expected cost of failures	1,152,011	977,283	-17.87%
Total cost	1,463,761	1,275,353	-14.77%

Figure 3 shows the cumulative days before MM for all train-sets on the basis of PM-only and PM + PdM strategies. The accumulative operating days under the PM-only strategy are generally near the MM threshold of 90 days. On the contrary, the accumulative operating days under the PM + PdM strategy vary in accordance with the actual reliability of the train-set. In terms of the EMU700 train-sets, the accumulative operating days before entering the MM under the PM-PdM strategy is about 95 days on average, an extension from the 90-day threshold adopted by the PM strategy. However, the accumulative operating days of the EMU500 train-sets (i.e., EMU542, EMU544, and EMU546) are considerably lower than the maintenance regulation. This is because, the EMU500 train-sets, as the oldest types of existing train-sets for commuter trains, has much lower reliability than that of the EMU700 train-sets. Introducing the PdM strategy provides flexibility in the maintenance schedule by train-set specific reliability models. As a result, an efficient and reliable assignment plan can be determined through the proposed process.

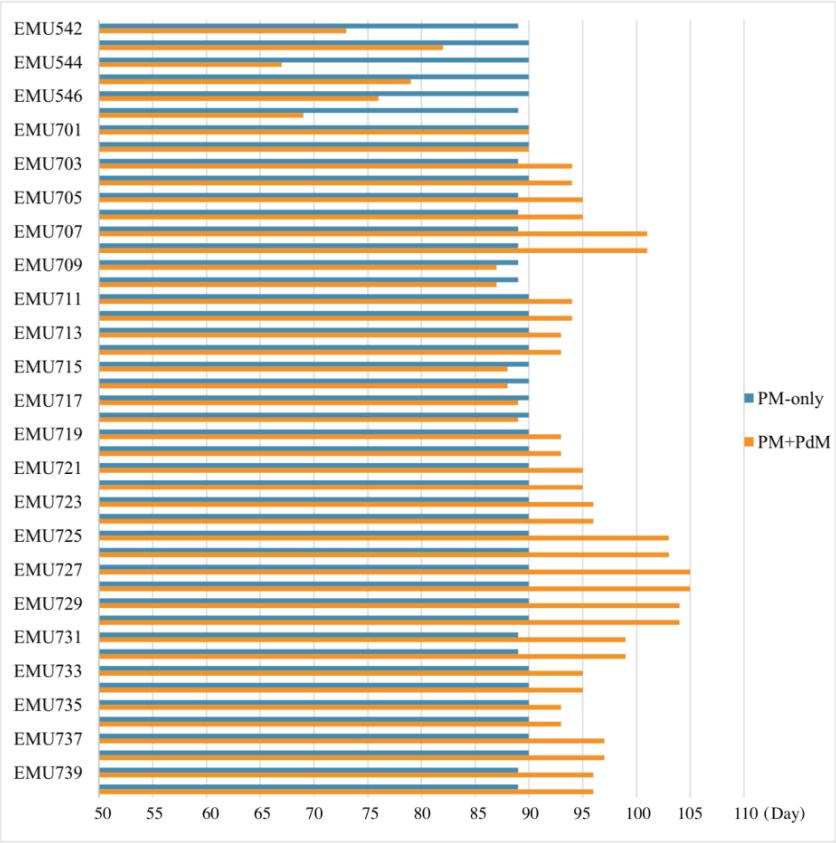


Figure 3: Accumulative days of MM

5 Conclusions

This study proposes an optimization process for assigning train-set to utilization paths and maintenance tasks in accordance with the PdM strategy with train-set specific reliability models. The results of the empirical study demonstrate that the developed process can assign utilization paths and schedule maintenance tasks to each train-set efficiently and reduce the total cost by over 14% compared with the PM-only strategy. Adopting this process can help planners improve the efficiency and reliability of train-set utilization.

References

Asekun, O.O., 2014. "A Decision Support Model to Improve Rolling Stock Maintenance Scheduling Based on Reliability and Cost", *Master thesis of Engineering in Engineering Management in the Faculty of Engineering at Stellenbosch University*.

Caprara, A., Kroon, L., Monaci, M., Peeters, M., and Toth, P., 2007. “Passenger Railway Optimization”, *Handbooks in Operations Research and Management Science*, vol. 14, pp. 129-187.

Cheng, Y.H.,2010. “High-speed rail in Taiwan: New experience and issues for future development”, *Transport Policy*, vol. 17(2), pp. 51-63.

Herr, N., Nicod, J.M., Varnier, C., Zerhouni, N., Cherif, M., and Fnaiech, N., 2017. “Joint optimization of train assignment and predictive maintenance scheduling”, In: *Proceedings of 7th International Conference on Railway Operations Modelling and Analysis*, Lille, France.

Li, H., Parikh, D., He, Q., Qian, B., Li, Z., Fang, D., and Hampapur, A., 2014. “Improving rail network velocity: A machine learning approach to predictive maintenance”, *Transportation Research Part C: Emerging Technologies*, vol. 45, pp. 17–26.

Li, J., Lin, B., Wang, Z., Chen, L., and Wang, J., 2016. “A Pragmatic Optimization Method for Motor Train Set Assignment and Maintenance Scheduling Problem”, *Discrete Dynamics in Nature and Society*, vol. 2016.

Moghaddam, K.S. and Usher, J.S., 2011. “Preventive maintenance and replacement scheduling for repairable and maintainable systems using dynamic programming”, *Computers & Industrial Engineering*, vol. 60, pp. 654–665.

Yun, W.Y., Han, Y.J. and Park, G., 2012. “Optimal Preventive Maintenance Interval and Spare Parts Number in A Rolling Stock System” In: *Proceedings of 2012 International Conference on Quality, Reliability, Risk, Maintenance, and Safety Engineering*, Chengdu, China.

Improvement of maintenance timetable stability based on iteratively assigning event flexibility in FPESP

Raimond Wüst [a,1], Stephan Bütikofer [a], Severin Ess [a], Claudio Gomez [a], Albert Steiner [a],
Marco Laumanns [b], Jacint Szabo [c]
[a] Institute of Data Analysis and Process Design, Zurich University of Applied Sciences ZHAW, PO Box, 8401 Winterthur, Switzerland
[1] E-mail: wura@zhaw.ch, Phone: +41 (0) 58 934 65 81
[b] Bestmile SA, 1007 Lausanne, Switzerland
[c] IBM Research, now at Google Switzerland, Brandschenkestrasse 110, Zurich, Switzerland

Abstract
In the operational management of railway networks, the fast adaptation of timetable scenarios is an important requirement, in which operational disruptions or time windows with temporary unavailability of infrastructure, for instance during maintenance time windows, are taken into consideration. In those situations, easy and fast reconfiguration and recalculation of timetable data is of central importance. This local and temporal rescheduling results in shifted departure and arrival times and sometimes even in modified stop patterns at intermediate stations of train runs. In order to generate reliable timetabling results it is a prerequisite that train-track assignments as well as operational and commercial dependencies are taken into consideration. Therefore, it is crucial for the computer-aided planning process to refer to the right level of detail for the modelling of the track infrastructure and train dynamics. In this article we present a generic model that we call Track-Choice FPESP (TCFPESP), as it implements suitable extensions of the established PESP-model. We show how the service intention (the timetable specification resulting from line planning) together with resource capacity information can be utilized to configure the TCFPESP model.
In addition, we can calculate quantitative performance measures for assessing timetable quality aspects. To achieve this, we make use of the max-plus algebra for evaluating timetable stability. By utilizing delay impact values resulting from max-plus algebraic performance analysis, we are thus able to iteratively distribute event flexibility in such a way that overall stability of the maintenance timetable is improved.
This approach supports the planner to generate integrated periodic timetable solutions in iterative development cycles.

Keywords
Flexible PESP, Mesoscopic railway topology, Service Intention, Timetabling with track assignment, Timetable stability analysis

1 Introduction

1.1 Generating timetable scenarios for short term planning

In the operational management of railway networks, an important requirement is the fast adaptation of timetable scenarios, in which operational disruptions or time windows

with temporary unavailability of infrastructure, as for instance during maintenance time windows ('possessions', see RailNetEurope (2017)), have to be accounted for. In those situations, easy and fast reconfiguration and recalculation of timetable data is of central importance. This local and temporal rescheduling results in shifted departure and arrival times and sometimes even in modified stop patterns at intermediate stations of train runs. Only recently, van Aken et al. (2017a) presented a PESP based macroscopic model for solving train timetable adjustment problems (TTAP) under infrastructure maintenance possessions (2017a). They show, that by applying TTAP, they are able to adjust a given timetable to a specified set of station track and complete open-track possessions by train retiming, reordering, short-turning and cancellation. In van Aken et al. (2017b) they apply several network aggregation techniques to reduce the problem size and thus enable the model to solve large instances within short computation times with instances of the complete Dutch railway network.

However, in order to generate reliable timetabling results it is prerequisite that besides train-track assignments, also operational and commercial dependencies are taken into consideration. Hence, finding the right level of detail for modelling track infrastructure and train dynamics is crucial for supporting the planning process in an optimal way.

In day-to-day business, determining the feasible event times for individual train runs and the corresponding resource-allocation fitting into efficient transport chains resulting from an integrated clockface timetable is time-consuming and is carried out manually. On the other hand, algorithmic approaches for solving this task computationally require models based on microscopic information about track capacity. This capacity information is aggregated to (normative) minimum headway constraints that are used for solving standard periodic timetabling problems. To facilitate this step, several research groups made suggestions, how to combine common timetabling procedures with constraints resulting from mesoscopic infrastructure information. Hansen and Pachl (2008) show how running, dwell and headway times at critical route nodes and platform tracks must be taken into account for train processing and present a deep timetable quality analysis depending on these parameters. De Fabris et al. (2014) calculate arrival and departure times, platform and route assignments in stations and junctions that trains visit along their lines. Bešinović et al. (2016) present a micro–macro framework based on an integrated iterative approach for computing a microscopically conflict-free timetable that uses a macroscopic optimization model with a post-processing stability evaluation. Caimi et al. (2011) extend PESP (see e.g. Serafini and Ukovich (1989) and Liebchen and Möhring (2007)) and propose the flexible periodic event scheduling problem (FPESP), where intervals are generated instead of fixed event times. By applying FPESP, the output does not define a final timetable but an input for finding a feasible timetable on a microscopic level, (Caimi (2009) and Caimi et al. (2009)).

1.2 Service Intention based approach for timetable specification

To improve customer value even under limited operating conditions, such as those encountered during infrastructure maintenance intervals, our modelling approach for creating temporary schedules is also based on an extension of PESP and takes the 'service intention' (SI) as input data. The SI was first described in Wüst et al. (2008), formally specified in Caimi (2009) and integrates commercial timetabling requirements given by the respective demand oriented 'line concept' on one side and technical constraints on the other. The 'line concept' results from a strategical planning process which is executed by the transport carrier. In this process, the available amount, the dynamics and the circulation of rolling stock are taken into account. In Switzerland, the integrated fixed-

interval timetable (IFIT) is created on the basis of SI's. The required system times (minimum travel times between node stations, see for example Herrigel (2015)) are a prerequisite (see e.g. Liebchen and Möhring (2007)).

The maintenance interval planning step (denoted as IP in the sequel) is executed by the infrastructure manager. In this step, the functional requirements of the SI are brought together with this mesoscopic infrastructure data model of a given scenario. Altogether these data can be maintained in a standard timetable editor (see for instance SMA Viriato, 2018). In this way, the SI represents functional timetabling requirements including line data, line frequencies and separations as well as line transfers at specific stations. Hence, it contains explicit information about intended transport chains but is still flexible enough, to allow different ways of operational planning and resource allocation. Like de Fabris et al. (2014), we call this level of abstraction of the available resources 'mesoscopic topology'. We call our FPESP model that we apply to this mesoscopic topology 'Track-Choice FPESP' (TCFPESP). To facilitate the problem of searching feasible solutions for local resource restrictions during maintenance intervals we make the assumption, that the train network outside the maintenance corridor is not affected by the restrictions at the level of mesoscopic topology. This allows us to separate the network into aggregated network partitions outside the IP relevant corridor and the disaggregated network partition at mesoscopic topology level. This network segmentation has also some similarity to the decomposition approach suggested by Lamorgese et al. (2016). They present an iterative dispatching algorithm in which the network is sequentially decomposed into a macroscopic line dispatching (master) and a microscopic station dispatching (slave) algorithm.

To evaluate timetable stability criteria we use a special algebraic approach that is commonly known as max-plus algebra. This approach has been elaborated in mathematical detail by Goverde (2007) who also demonstrates the benefits of this algebraic approach for timetable stability analysis in practical applications. According to this approach, timetable stability is defined in terms of the difference between the timetable period T and λ_0, defined as the maximum cycle mean over all circuits in the event activity network. If $\lambda_0 < T$ the timetable is considered to be stable. We apply this method for evaluating the stability of our resulting timetable and try to improve the timetable based on this performance evaluation in successive re-planning iterations. More specifically, we show how the max-plus-delay impact analysis can help to improve timetable stability by iteratively adjusting local flexibility constraints in the configuration of the TCFPESP model.

As we want to demonstrate the operational benefit that can be obtained by utilizing the max-plus stability analysis for TCFPESP based re-planning, we finally present a case study without fixed time constraints for the planning step. This configuration represents the use case of designing a new timetable rather than the use case for altering an existing timetable, which is the typical constellation when planning temporary timetables for maintenance intervals. However, we think that in this way we can clearly point out the mentioned relationship between the planning step and the performance analysis step.

1.3 Structure of this article

This article is structured as follows: In section 2, we describe the methodology for achieving the research goals. In section 2.1 we summarize the FPESP model which implements the idea of periodic timetabling with event flexibility. Extending this FPESP to our proposed mesoscopic model we present in Section 2.2 our TCFPESP-model. For the iterative configuration of the event flexibility in the TCFPESP we make use of the

delay impact vector that we obtain from max-plus analysis. This is shown in section 2.3. In section 2.4 we describe the TCFPESP heuristic for reducing the overall delay impact. In section 3 (Case Study 'Kerenzerberg') we present the results from applying the methods introduced in section 2 and the coordinated application in a real-world scenario from eastern Switzerland. Finally, in section 4 we conclude with a summary of the findings and an outlook on future work.

2 Methodology

2.1 Periodic Timetabling with Event Flexibility

The classical PESP tries to determine a periodic schedule on the macroscopic level (i.e. without using the tracks at an operation point) within a period *T*. Event $e \in E$ takes place at time $\pi_e \in [0, T)$. The schedule is periodic with time period *T*, hence each event is repeated periodically $\{\dots, \pi_e - T, \pi_e, \pi_e + T, \pi_e + 2T, \dots\}$.

The choices of the event times π_e depend on each other. The dependencies are described by arcs $a = (e, f)$ from a set A and modelled as constraints in the PESP. The constraints always concern the two events *e* and *f* and define the minimum and maximum periodic time difference $l_a$and u_a between them. These bounds are given as parameters in the PESP model. We therefore look for the event times π_e for every $e \in E$ that fulfill all constraints of the form

$$l_a \leq \pi_f - \pi_e + p_a T \leq u_a$$

for all $a = (e, f) \in A$,, where p_a is an integer variable that makes sure, that these constraints are met in a periodic sense.

In order to avoid tedious iterations between the process steps "microscopic capacity planning" and "mesoscopic capacity planning" in case of infeasibility of the micro-level problem, one can improve the chance of finding a feasible solution by enlarging the solution space in the micro-level. This approach has been described in detail in Caimi et al. (2011b). We also implement this event flexibility method by adding some flexibility for the events of the event and activity network (E, A) by introducing lower and upper bounds to the event times of the arrival and departure nodes in Figure 1b. The final choice of the event times in the range between the lower and upper bound shall be independent for each event such that each value of the end of an activity arc should be reachable from each time value at beginning of that activity arc (see Figure 1a).

We are not forced to add this flexibility to all the events, but we can select the nodes where we want to add it, for instance only nodes corresponding to events in a main station area with high traffic density, where it is more difficult to schedule trains on the microscopic level. In general, one can say, that this placement of flexibility is the timetable configuration feature, which has the highest level of influence on improving operational stability. This is where the information provided by the max-plus measures of delay impact (see section 2.3 et seq.) can be utilized in order to achieve timetable stability. For more details regarding the FPESP method, we refer to the article of Caimi et al. (2011b).

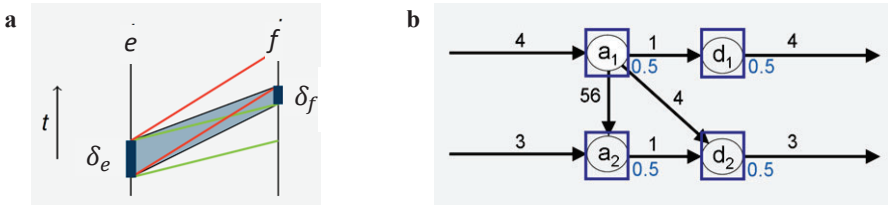


Figure 1: Target oriented placement of time reserves. a) Time frames $[\pi_e, \pi_e + \delta_e]$ in place of time points π_e. By implementing this method, the normal PESP constraints $l_a \leq \pi_f - \pi_e + p_a T \leq u_a$ now become $l_a + \delta_e \leq \pi_f - \pi_e + p_a T \leq u_a - \delta_f$ (see next section). In the example b) this means that instead of planning time points $(\pi_{a_1}, \pi_{d_1}, \pi_{a_2}, \pi_{d_2})$ we plan time frames $[\pi_e, \pi_e + 0.5]$ for $e \in \{a_1, d_1, a_2, d_2\}$.

2.2 Track-choice FPESP.

For our proposed timetabling model, we extend the FPESP method with events at track-level in order to generate event slot timetables on a mesoscopic level. In the TCFPESP model, the mesoscopic infrastructure consisting of sections is summarized as a set I of operation points. Operation points are largely tracks and stations but can also be other critical resources such as junctions (see OP 'Tiefenwinkel' in Figure 2b). As mentioned before, each operation point $i \in I$ is associated to a capacity consisting of a set of tracks T_i. A train run $l \in L$ is described by a sequence of operation points of I.

Based on this mesoscopic model we form an event-activity network (E, A). The set E of events consists of an arrival event arr_{li} and a departure event dep_{li} for each train run $l \in L$ and operation point $i \in l$. The activities $a \in A$ are directed arcs from $E \times E$ and describe the dependencies between the events. For every train run we have arcs between arrival and departure events at the same operation points (dwell times or trip times) and arcs between departure and arrival events of successive operation points (running time between operation points). Further arcs include connections between train runs, headways and turnaround operations. Headway arcs $a \in A_H$ are especially important for explaining the 'track-choice FPESP model' below. Headways are used to model safety distances between trains running in the same and in opposite directions. For the sake of simplicity in the formal description of the TCFPESP we consider only headways related to one operation point, i.e. we omit headways for train runs in opposite directions over several successive operation points. The TCFPESP-model can be easily extended to include general headways. They are included in our implementation of the timetable model.

We extended the classical PESP resp. FPESP model by using the number of tracks T_i at each operation point $i \in I$. The track-choice FPESP model assigns the arrival event arr_{li} and the departure event dep_{li} of train run l at operation point i uniquely to a track in T_i. We can use these assignments to switch on headway arcs $a \in A_H$ by using the following big-M-approach. In addition to variables π and p from the classical PESP model we need:

(i) Binary variables tc_{et} (track choice) for each event $e \in E$ and track $t \in T_{i(e)}$, where operation point $i(e)$ is associated to event e, i.e. e is equal to arr_{li} or dep_{li} for a train run l.

(ii) Binary variables h_a for every headway edge $a = (e, f) \in A_H$. As mentioned before, headway edges are always between events at the same operation point, therefore $T_{i(e)} = T_{i(f)}$ holds.

(iii) Positive variables δ_e for each event $e \in E$ to model the event flexibility.

The track-choice model is defined by:

$$\min f(\pi, \delta)$$

$$\begin{aligned}
s.t. \quad & l_a + \delta_e \leq \pi_f - \pi_e + p_a T \leq u_a - \delta_f, && \forall\ a = (e,f) \in A \setminus A_H, && (1)\\
& l_a + \delta_e - (1 - h_a)M \leq \pi_f - \pi_e + p_a T && \forall\ a = (e,f) \in A_H, && (2)\\
& \qquad \leq u_a - \delta_f + (1 - h_a)M, && && \\
& \sum_{t \in T_{i(e)}} tc_{et} = 1, && \forall\ e \in E, && (3)\\
& tc_{arr_{li}t} = tc_{dep_{li}t}, && \forall\ l \in L, i \in l,\ t \in T_i, && (4)\\
& h_a \geq tc_{et} + tc_{ft} - 1, && \forall\ a = (e,f) \in A_H, && (5)\\
& && t \in T_{i(e)} && \\
& tc_{et}, h_a \in \{0,1\}, \pi_e \in [0,T), p_a \in \mathbb{Z}, \delta_e \geq 0, && \forall\ e \in E, t \in T_{i(e)}, a \in A, &&
\end{aligned}$$

where M is a big enough natural number.

In (1) the normal FPESP constraints are summarized (without headway arcs). In (2) are the headway constraints, which can be switched off with a big-M technique. The assignment of the events to the tracks is done in (3). (4) is used to assign the corresponding arrival and departure events to the same track. In (5) the headway variable is set to 1, if the events take place on the same track, i.e. the headway is required at this operation point.

There are many different objective functions $f(\pi, \delta)$ suggested by Caimi et al. (2011b) for the FPESP model. To generate the traffic plan for our test scenario we use iteratively the TCFPESP with different objective functions (see Wüst et al. (2018b)), namely:

- We minimize all passenger relevant times (i.e. $t \in A_T$ the set of trip arcs, $d \in A_D$ the set of dwell arcs and $c \in A_C$ the set of connections times). The weights w_t, w_d and w_c can be used for prioritizing certain times, e.g. connection times. We will call the model in this case MINTRAVEL, according to Caimi et al. (2011b). The objective function is defined as follows:

$$f_{TT}(\pi) = \sum_{t \in A_T} w_t \pi_t + \sum_{d \in A_D} w_d \pi_d + \sum_{c \in A_C} w_c \pi_c \tag{6}$$

- We maximize the flexibility in a certain range at certain arrival and departure events. The objective function is defined as follows:

$$f_{flex}(\delta) = \sum_{e \in V} w_e\, \delta_e, \tag{7}$$

  where $V \subseteq E$ is the set of all events where flexibility is introduced.

  Furthermore we add two constraints. The passenger travel time f_{TT} has to be smaller than $(1 + \epsilon)$ times the best possible travel time f_{TT}^* from the model MINTRAVEL. The flexibility for all events is bounded by a maximal flexibility δ_{max} for a better distribution of the flexibility to all events. The two constraints are given by

$$f_{TT}(\pi) \leq (1 + \epsilon) f_{TT}^* \quad \text{and} \quad \delta_e \leq \delta_{\max}\ \forall\ e \in E, \tag{8}$$

where f_{TT}^{*} is the optimal value found for f_{TT} in (6).

We will call the model in this case CONTRAVEL according to Caimi et al. (2011b). ϵ is a parameter controlling the quality of the schedule for the passengers' travel times and the weights w_e will be used for individual adjustments in event flexibility in order to maximize timetable stability (see section 2.3 and 2.4).

Both models MINTRAVEL and CONTRAVEL are therefore mixed integer linear problems. By using the models MINTRAVEL and CONTRAVEL iteratively we can generate a traffic plan covering stability and travelling time aspects (see Wüst et al. (2018b)).

2.3 Computation of the Cumulative Delay Impact

The Cumulative Delay Impact (CDI) is a measure to quantify the overall impact that a certain delay κ at a specific event f has on all other events e. Formally the CDI is computed as follows:

$$CDI_f(\mathrm{R}) = \sum_{e \in E \setminus f} \max\left(\kappa - r_{ef}, 0\right)^{\gamma}, \tag{9}$$

where E denotes the set of all events. R represents the recovery matrix of size $| E | \times | E |$ and r_{ef} represents the actual buffer time between events f and e given a periodic timetable π (For the details on the calculation of the recovery matrix R and the buffer times r_{ef} see Goverde (2005, 2007)). The event times are resulting from TCFPESP by taking the lower bounds of the event time intervals calculated.

κ is the parameter that denotes the initial delay (in minutes) applied to node f, for which CDI_f shall be calculated. Finally, $\gamma \geq 1$ is a parameter to increase the impact of positive differences between the delay κ and r_{ef}. In this study γ was always set to 1. Furthermore, CDI_f is strictly monotonically increasing in κ and $CDI_f(\mathrm{R}) = 0$ for $\kappa = 0$, $\gamma \geq 1$. The initial delay κ can of course be set for each event $f \in E$ individually, e.g. when κ is determined with the help of a statistical delay analysis for each event $f \in E$.

2.4 Heuristic for improvement of delay impact

We measure the stability of a periodic timetable π by the sum of all cumulative delay impacts, i.e. we consider $f_{sta}(\pi) = \sum_{f \in E} CDI_f(\mathrm{R})$. Given an acceptable κ (from an operational point of view), we would like to have this measure as small as possible. From the definition of CDI, it follows, that $f_{sta}(\pi)$ is bounded from below by 0.

It would therefore be natural to use $f_{sta}(\pi)$ in the CONTRAVEL model as objective function. Since we don't have a direct solution approach for this case, we propose the following heuristic.

Iteratively we try to use the weights w_f in the function $f_{flex}(\delta)$ to give more flexibility to the events $f \in E$, where $CDI_f(\mathrm{R})$ is > 0. Weight w_f is computed as follows:

$$w_f = \begin{cases} \left(\frac{CDI_f(\mathrm{R})}{\max\limits_{f \in E} CDI_f(\mathrm{R})}\right)^{\alpha} & \text{if } \max\limits_{f \in E} CDI_f(\mathrm{R}) > 0 \text{ and } CDI_f(\mathrm{R}) \geq \theta \cdot CDI_{max} \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

where $CDI_{max} = \max\limits_{f \in E} CDI_f(R)$ is the maximum CDI value observed, θ is a

threshold parameter to determine, which values of $CDI_f(R)$ are a considered for subsequent weightings, $\alpha \geq 1$ is a parameter to over proportionally increase the weights, the larger $CDI_f(\mathrm{R})$ is.

Iteration scheme: Improving delay impact

Input:

- Periodic timetable π computed with the CONTRAVEL model. The weights w_f in the objective function are set to 1 for this initial timetable π.
- Set initial delay κ and parameter α and γ.

Iteration steps:

Step 1: Compute $f_{sta}(\pi)$ by summing up the $CDI_f(\mathrm{R})$ for all $f \in E$.
If $f_{sta}(\pi) = 0$,
Stop iteration and accept timetable π.
Else
Set $\beta = f_{sta}(\pi)$.
End If

Step 2: **While** $f_{sta}(\pi) \leq \beta$
For timetable π set the weights w_f according to equation (10).
Recompute a new timetable π_{new} with the help of the CONTRAVEL model and the new weights w_f.
Compute $f_{sta}(\pi_{new})$.
If $f_{sta}(\pi_{new}) < \beta$,
Set $\beta = f_{sta}(\pi_{new})$ and $\pi = \pi_{new}$.
Else
Set $\beta = (-1)$ (leave while loop).
End if
End While

Step 3: Accept timetable π.

In the iteration scheme above we compute in step 1 the sum of the cumulative delays of the initial timetable π. As mentioned above the timetable π from the CONTRAVEL model corresponds to computed lower bounds of the single events. In step 2 we enter a while loop as long as the adaption of the weights w_f leads to an improvement of the stability measure f_{sta}. The timetable π with the minimal stability measure f_{sta} during the iterations will be accepted at the end.
All timetables during the iterations fulfil the same service intention (see section 1.2), but the resulting timetable is the most robust one with respect to the cumulative delay impact measure (among the constructed timetables during iterations). We illustrate this iteration scheme in our case study in section 3.

3 Case study 'Kerenzerberg'

In order to illustrate the iterative improvement of timetable stability for IP scenarios, we selected a railway corridor in the eastern part of Switzerland. We call the case study 'Kerenzerberg' and the maintenance work is planned on the network section between Flums and Mels. The impact on the schedule is that there is a reduced velocity on that

section during normal operation hours.

3.1 Network segmentation

To avoid putting too much effort into entering information that is not needed and rather focus on the relevant perimeter for the IP timetabling scenario, one has to identify which part of the entire railway network has to be investigated and which part will be assumed to remain as given by the ordinary timetable. In a first step, the relevant lines and services operating on the subnetwork, which will be affected by the construction sites, have to be identified. In a second step, those lines, which are coupled (e.g. by transfers or technical dependencies) to these affected lines have to be found.

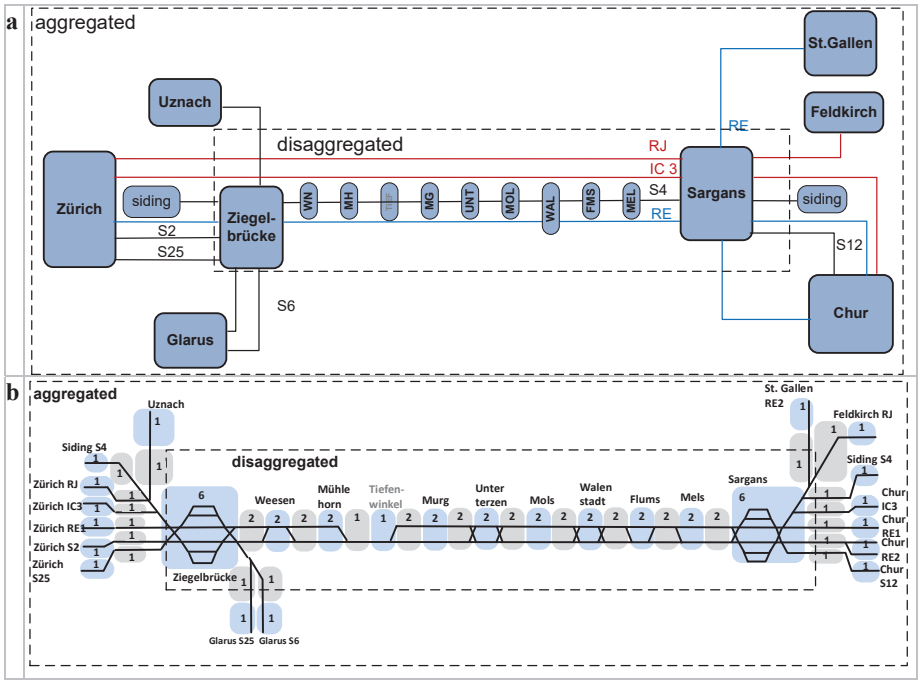


Figure 2: Case study Kerenzerberg a) In order to divide the relevant infrastructure for the IP timetabling scenario into a network partition with the relevant level of detail and a peripheral part with more coarse information, the railway network is divided into subnetworks. A disaggregated subnetwork containing the relevant infrastructure segments at mesoscopic level and an aggregated subnetwork, representing simplified infrastructure on the macroscopic level. b) Shows the track topology for the both, the aggregated and disaggregated network partitions. The grey shaded topology points represent section type operation points, light blue shaded topology points indicate operation points. Numbers indicate the topology point's number of tracks. In order to avoid treating line interactions outside the disaggregated partition, each line has an individual periphery with a section between the final destination point and the boundary operation point. The section that separates the two partitions from each other is configured with aggregated running times and dwell times of the respective line.

One has to identify the sub-network nodes which isolate the relevant infrastructure partitions from the fixed periphery. In this way one obtains a disaggregated subnetwork containing the relevant infrastructure segments and an aggregated subnetwork, representing infrastructure on the macroscopic level (see outer dashed square areas in Figure 2a and b). The disaggregated subnetwork is configured with all mesoscopic details. On this disaggregated subnetwork all train movements are planned in detail. For each line coming from or going beyond the boundary nodes of the disaggregated subnetwork we create a virtual end station node which is connected by a single section to the corresponding boundary node. The section lengths with the appropriate trip times, the turnaround times of the line outside the disaggregated subnetwork together with the run- and dwell times within the disaggregated subnetwork have to sum up to the proper roundtrip time. The mesoscopic track topology of the disaggregated subnetwork is illustrated in Figure 2b).

3.2 Network of the case study Kerenzerberg.

The planned construction or maintenance work for our test scenario 'Kerenzerberg' is located on the network section between Flums and Mels. During the IP interval, trains are running with reduced speed in both directions. We decided to use the corridor Ziegelbrücke-Sargans as the disaggregated partition of the test network. It has to be mentioned, that there is a single-track section between the operation points 'Mühlehorn' and 'Tiefenwinkel'. For this disaggregated network partition, we iteratively generate IP timetable scenarios (see section 3.4). The western part of Ziegelbrücke is aggregated, i.e. we introduced the nodes Uznach, Zürich, Glarus and a siding of Ziegelbrücke and connecting tracks. The aggregated network will be used to maintain vehicle circulation (e.g. turnarounds) aspects of lines and to model connections to peripheric lines (see the description of SI in section 3.3). The eastern part of Sargans also belongs to the aggregated partition. We introduced the nodes St.Gallen, Feldkirch, Chur and a siding of Sargans. In the aggregated network we assume to have enough track capacity to compensate for delays.

3.3 Description of Service Intention

The configuration of the SI is mainly done in the planning system Viriato. Additional information like upper boundaries of time intervals and flexibility of event times as required in the TCFPESP model is maintained in an R-based table editor (see chapter 2.2). The SI-lines represent the lines in the corresponding timetable 2018 with minor adaptations. In order to demonstrate the turnaround operations within our test scenario, we decided that the line S4 makes a turnaround in a siding next to Ziegelbrücke and Sargans, respectively. The other commuter lines (S x) rotate between a final station and a boundary node or between two final stations via a boundary node. Minimal line rotation times and line frequencies are indicated in Table 1.

Table 1: Line rotations and line frequencies

Line ID	Minimum line rotation time (min)	Line frequency (repetitions per hour)
S4	58.8	1
RJ	47.3	0.5

IC 3	43.8	1
RE 1	50.3	1
S12	12	1
S25	13	1
S6	16	1
RE 2	12	1
S 2	11	1

Table 1: Line rotations and frequencies. The minimum line rotation times are computed according to the approach of Liebchen and Möhring (2007). The corresponding turnaround intervals are computed in such a way, that a service with a minimal number of rolling stock is possible. In our case study the line S 4 is operating with one rolling stock. The other lines operate with more than one rolling stock due to longer round-trip times. These bounds are not computed according to Liebchen and Möhring (2007), they are set manually and have reduced line rotation times.

Ziegelbrücke and Sargans are considered as local hubs. At these stations the traffic plan has to account for passenger transfers between lines. Technically, these transfer requirements result in connections constraints in our TCFPESP-model. These line connections are indicated in Table 2. For a detailed definition of the infrastructure and the SI specification including time intervals of running times, dwell times, turnaround times, separation times etc. see Wüst et al. (2018b).

Table 2: Line Connections at Stations

Connection [1, 15] From/To at station	S 25 (ZB-GL)	S 25 (GL-ZB)	S 4 (ZGB-SA)	S 4 (SA-ZGB)	S 12 (SA-CH)	RE 2 (CH-SG)	IC 3 (ZGB-SA)	RE 1 (SA-ZGB)	RE 1 (ZGB-SA)	S 6 (GL-UZ)	S 6 (UZ-GL)
S 4 (ZGB-SA)	ZGB	ZGB			SA						
S 4 (SA-ZGB)		ZGB			SA						
S 25 (GL-ZB)			ZGB								
S 25 (ZB-GL)				ZGB							
IC 3 (ZGB-SA)						SA					
S 12 (CH-SA)							SA				
RE 2 (CH-SG)								SA			
RE 2 (SG-CH)							SA	SA			
RE 1 (ZGB-SA)											ZGB
RE 1 (SA-ZGB)										ZGB	ZGB
S 6 (GL-UZ)								ZBG	ZGB		
S 6 (UZ-GL)								ZGB			

Table 2: Case study Kerenzerberg: Line connections at stations are dependent on the direction of the train runs. The time intervals for connection arcs [lb, ub] is configured identically for all connections: [1 min, 15 min].

3.4 Iterative improvement of timetable stability

Once the configuration of the SI and the mesoscopic infrastructure is complete it is

transformed into the TCFPESP model which was implemented in GAMS by applying the CONTRAVEL model as indicated by equations (7) and (8) with parameter $\epsilon = 0.5$ and $\delta_{max} = 10s$. In case the SI is feasible with respect to the capacity constraints given by the infrastructure, GAMS returns the timetable π with flexibility δ.

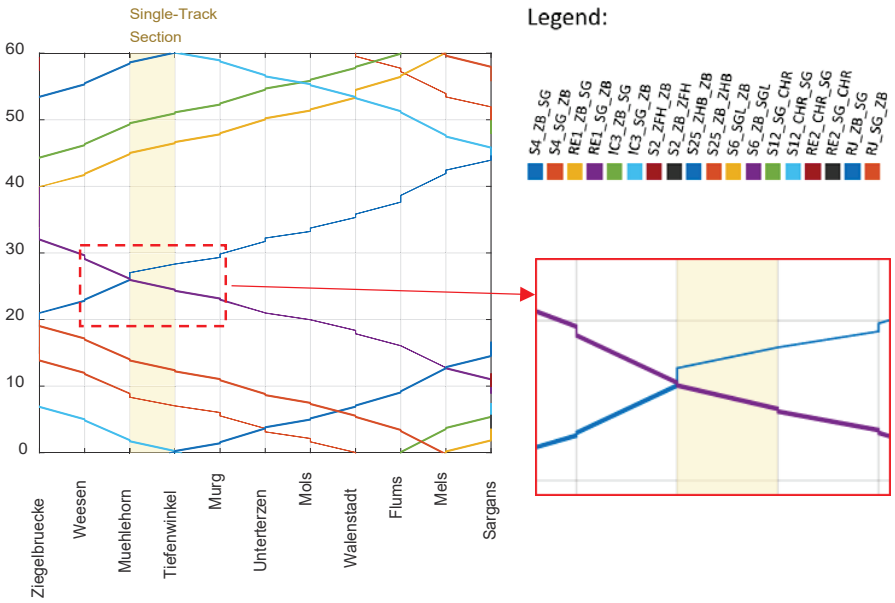


Figure 3: Time-distance diagram: GAMS output for TCFPESP applying step 1 of the iteration scheme of section 2.4. Line names and directions are indicated by colours as shown in the legend. One can also see the narrow but variable width of the capacity time bands indicating a low flexibility of each train run in a range below 10 seconds.

These are plotted as time-distance diagram as shown in Figure 3. This represents the result of the first step in the iteration scheme in section 2.4 . The timetable is the result of a CONTRAVEL-model configuration (see equation (8) in section 2.4.). As can be seen in the diagram, the range of flexibility of the train runs is quite narrow which is due to a small $\delta_{max} = 10s$, but show variable width within a certain range up to δ_{max}.

They are quite homogenously distributed, indicating some, but low flexibility in all timetable events. The stability of this result is assessed by calculating $CDI_f(\mathrm{R})$ for an initial delay κ of 3 minutes. Figure 4 a illustrates the delay impact of each timetable event to all other network events indicated by the corresponding colour (dark colours indicate higher delay impacts) together with the interdependencies (connecting arrows) in the event activity network. In order to demonstrate the influence a target oriented adjustment of the event flexibility, for step 2 of the iteration scheme (for details see section 2.4) we decided to define two rather different settings: (i) with a threshold of $\theta = 0.95$ only a limited number of event nodes was selected for weighting, whereas (ii) with a threshold of $\theta = 0.40$ quite a large number of event nodes was selected for weighting. The weights w_f were subsequently used to calculate a more robust timetable π (see equation 10). This

time the parameter δ_{max} is set to 60 seconds in order to assign more flexibility to the critical events. The weights are shown in red in Figure 4b. The time-distance diagrams of the resulting timetables π with $\theta = 0.40$ and $\theta = 0.95$ are shown in Figure 5a and 5b. In step 2 only one iteration was performed until the timetable was accepted. One can clearly see that here certain timetable events have much more flexibility than others. If we sum up the delay impacts of all events of the two scenarios $\theta = 0.40$ and $\theta = 0.95$. respectively, we obtain an $f_{sta}(\pi)$ −value reduced to 87.0% and 79.3%, respectively, compared to the one of step 1 (see Figure 6d).

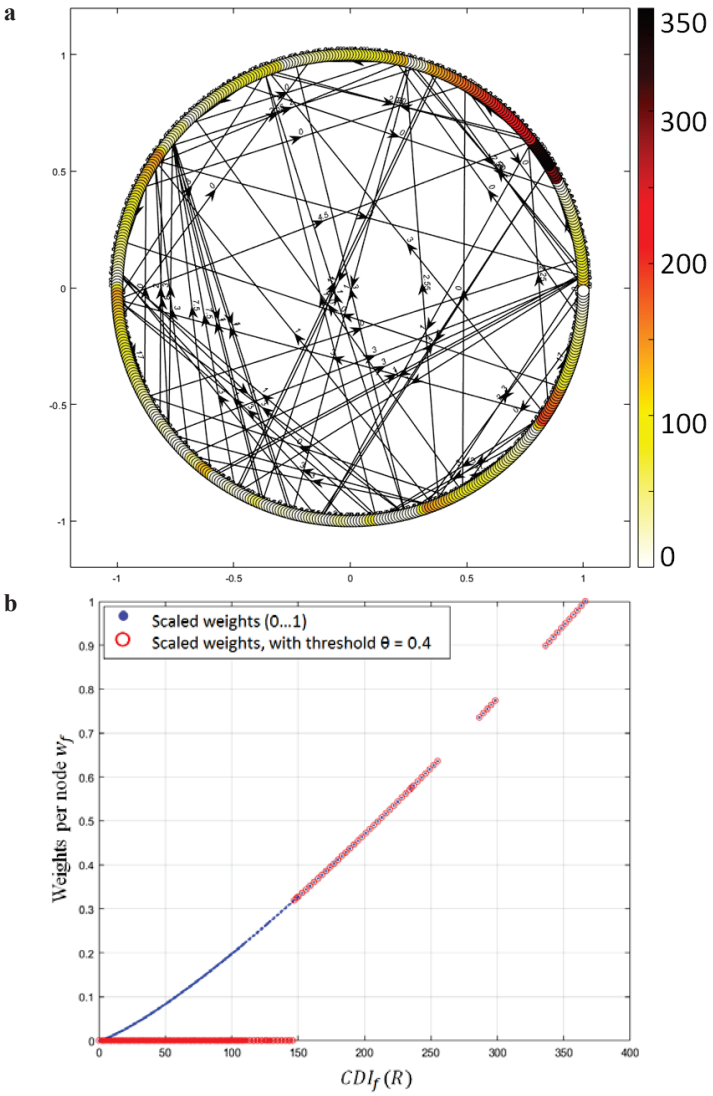


Figure 4: a) The values of the $CDI_f(R)$ for all event nodes (1 to 500) of the timetable

after step 1 indicated by a colour code ranging from 0 (low impact) to 350 (high impact) and the interdependencies between the event nodes. b) shows the weights (calculated according to equation 10), normalized to values in the range of 0 to 1 for all event nodes.

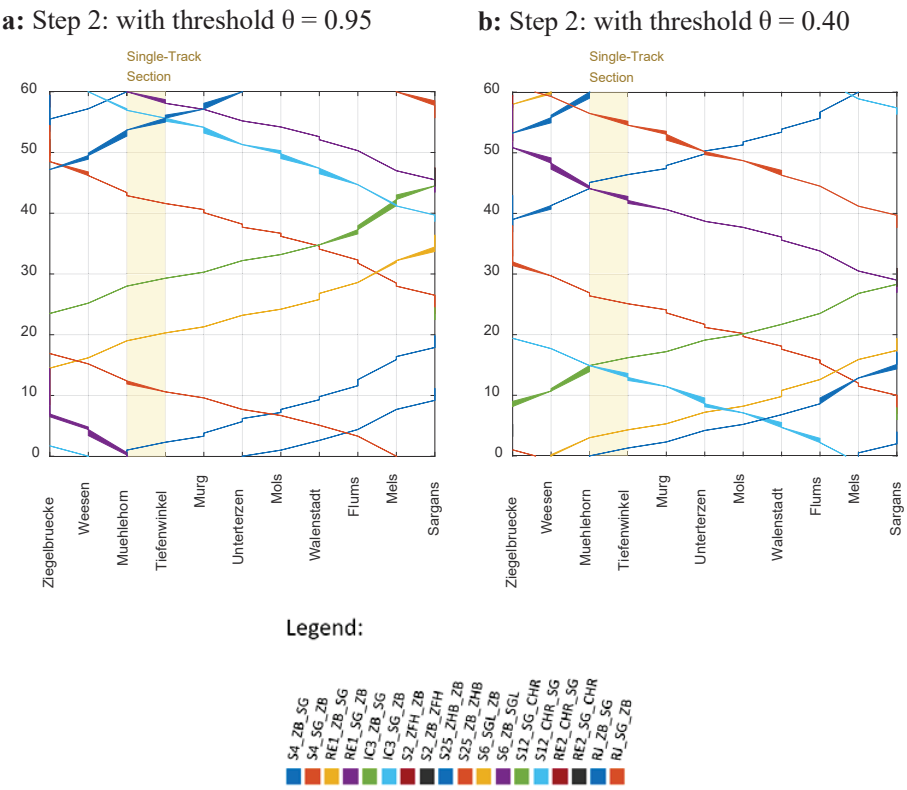


Figure 5: Time-distance diagram for the second iteration of the timetable calculation π. a) with a threshold θ = 0.95 and a resulting low number of weights w_f selected. b) with a threshold θ = 0.40 and a resulting rather high number of weights w_f selected. The line colours are the same as in Figure 3).

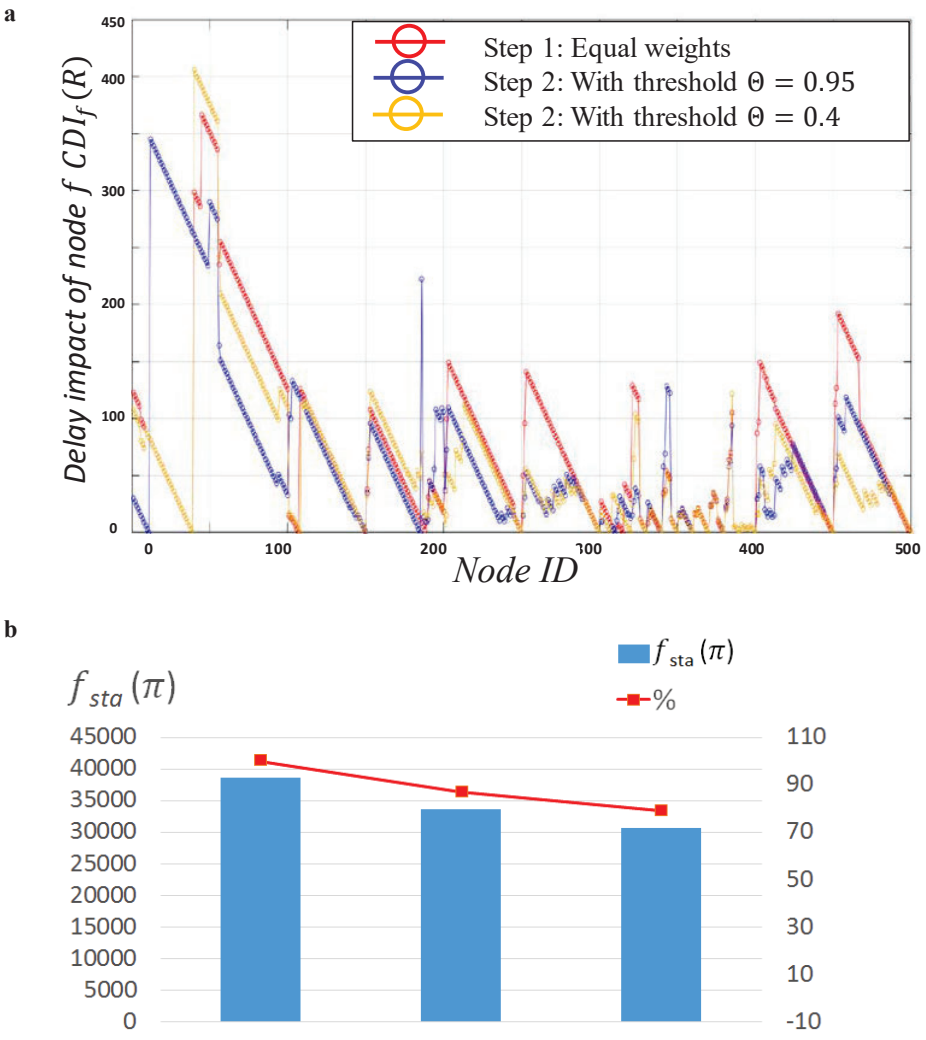


Figure 6: a) Distribution of the cumulative delay impact values $CDI_f(R)$ across nodes f for timetable after step 1 with equal weights applied (red curve), and for timetable after step 2 with few selected weights (blue curve, see text for selection criteria) and after step 2 with all weights applied (orange curve). b) Improvement of step 2 for $\theta = 0.95$ (middle bar) and for $\theta = 0.4$ (right bar) relative to step 1 (left bar).

The $f_{sta}(\pi)$-value in Figure 6a and 6b indicated as 'equal weights' was calculated with weights equal to 1 for all events in step 1.

4 Conclusions

The aim of this research was to introduce an extension of FPESP with track selection and a heuristic improvement of solutions based on max-plus algebraic stability analysis. The track choice extension of the FPESP approach accounts for a mesoscopic infrastructure level of detail which is an additional requirement for generating operational timetable scenarios. Temporary changes of infrastructure properties like the number or the maximum allowed speed of tracks and switches reduce the available capacity for track assignments to train runs. For this reason, we introduce an extension of the FPESP model that we call 'TCFPESP' model. The TCFPESP model allows to make a target-oriented adjustment of event flexibility by applying weights to the TCFPESP objective function. We obtain those weights from the calculation of the cumulative delay impacts for all timetable events and use them in an iterative manner for improving timetable stability.

However the SI-based timetable calculation only results in feasible solutions of the TCFPESP model described in section 2.2 and a stability improvement by applying the iteration scheme presented in section 2.4 if the temporary restrictions of infrastructure properties are not too severe. If this is the case, the SI has to be relaxed (especially the functional requirements part of it). This requires eventually adaptations to the underlying line concept. This is a different use case than the one, that we described here. In Wüst et al. (2018a) we show, that using the SI for specifying the functional and non-functional input for maintenance timetabling, we can generate different timetables for different maintenance scenarios without having to change the functional part of the SI. This has the advantage, that communicating only earliest departure and latest arrival times, the commercial timetable can remain unchanged for the whole planning period. The complete application concept has been described in Wüst et al. (2018b).

With our results we demonstrate the operational benefit that can be obtained by utilizing the max-plus stability analysis for TCFPESP based re-planning. To make the effect of the iterative stability improvement more clear and to illustrate the interoperability between the TCFPESP and max-plus framework we do not apply event time constraints at the boundary nodes. This makes that the resulting timetables of each iteration differ significantly from each other as can be seen from Figure 5 a and b. It is however rather easy to add additional event time constraints in TCFPESP to force the solution of a successive stability improvement step be close to the previous one. As those additional constraints limit the range for the reduction of the $CDI_f(R)$, we did without them in the presented case study.

We show results for a few example scenarios which demonstrate that we can reduce the overall delay impact of timetable events by a significant amount (a reduction of more than 20% in the second iteration compared to the first iteration). We consider these preliminary results as promising for making target-oriented improvements of timetable stability, especially in cases where variability of process times is high and cannot be reduced by operational measures. Timetable events that have a strong influence on many other timetable events should be planned with more flexibility than those with low cumulated impact. On one side the use case that we selected is based on operational requirements and the mesoscopic data level for the scenario is characteristic for maintenance timetable planning. On the other side, we wanted to point out the strong impact of the stability analysis-based re-planning iterations. This is the reason, why we did not use time-dependencies to fix pass times at the boundary nodes of the corridor what would be more typical for the use case of maintenance timetable planning. In future

studies we would like to closer investigate the potential for stability improvement also under conditions when existing timetables must be altered. In this context, we also want to further investigate the presented observations with the help of simulations on microscopic level. Our aim is to develop more specific application rules for the presented framework.

Acknowledgements

This research has been funded by the SBB-Research Fund and was part of the SmartRail 4.0 Development Initiative in order to introduce new railway technologies and improve the efficiency of a competitive Swiss Public Transport. We are very grateful also to IBM Research, who contributed with financial support to M. Laumanns and J. Szabo to develop the concept for the TCPESP-model. The technical implementation of the TCPESP-model and the $CDI_f(R)$ as well as the execution of this case study was done by IDP in collaboration with the SBB project team of SR4.0-TMS-PAS. Here we want to thank especially Thomas Wieland and Thomas Künzi for many inspiring discussions.

References

Bešinović, N., Goverde, R.M.P., Quaglietta, E., Roberti, R., 2016. *An integrated micro-macro approach to robust railway timetabling.* Transportation Research Part B: Methodological, vol. 87, pp. 14-32.

Caimi, G.,2009. *Algorithmic decision support for train scheduling in a large and highly utilised railway network*. Diss. ETH Zürich Nr. 18581

Caimi, G., Fuchsberger, M., Laumanns, M., Schüpbach, K., 2011. "Periodic railway timetabling with event flexibility" *Networks*, vol. 57, issue 1, pp. 3-18

de Fabris, S., Longo, G., Medeossi, G., Pesenti, R., 2014. "Automatic generation of railway timetables based on a mesoscopic infrastructure model" *Journal of Rail Transport Planning & Management*, vol. 4, pp. 2–13

Friedrich, M., Hartl, M., Schiewe, A. and Schöbel, A., 2017. *Integrating passengers' assignment in cost-optimal line planning.* Technical Report 2017-5, Preprint-Reihe, INAM, Georg-August Universität Göttingen

GAMS, 2018. General Algebraic Modeling System GAMS, GAMS Software GmbH, Frechen, Germany, https://www.gams.com/. last accessed: 19 September 2018

Goverde, R.M.P., 2005. *Punctuality of Railway Operations and Timetable Stability Analysis.* TRAIL Thesis Series no. T2005/10. Delft University of Technology, Delft, The Netherlands

Goverde, R.M.P., 2007. "Railway timetable stability analysis using max-plus system theory" *Transportation Research Part B*, vol. 41, no. 2, pp. 179-201

Hansen, I.A., Pachl., J. (eds.), 2008. *Railway Timetable & Traffic. Analysis – Modelling – Simulation.* Eurail Press, Hamburg, Germany

Herrigel, S., 2015. *Algorithmic decision support for the construction of periodic railway timetables*, Diss. ETH Zürich Nr. 22548, Department Bau, Umwelt und Geomatik, ETH Zürich, Switzerland. https://doi.org/10.3929/ethz-a-010412035

Lamorgese, L., Mannino, C., Piacentini, M., 2016. "Optimal Train Dispatching by Benders'-Like Reformulation" *Transportation Science*, vol 50, issue 3, pp. 910-925. DOI: https://doi.org/10.1287/trsc.2015.0605

Liebchen, C., Möhring, R.H., 2007. "The modeling power of the periodic event scheduling problem: Railway timetables – and beyond" In: Geraets, F., Kroon, L., Schöbel, A., Wagner, D., Zaroliagis, C. (Eds.), *Algorithmic Methods for Railway Optimization,* Lecture Notes in Computer Science, vol. 4359, pp. 3–40, Springer, Berlin/Heidelberg, Germany

RailNetEurope, 2017. Glossary of Terms Related to Network Statements [online]. RailNetEurope.http://www.rne.eu/rneinhalt/uploads/RNE_NetworkStatementGlossary__V8_2016_web.pdf [Accessed 22 Jan. 2019].

Serafini, P., Ukovich, W., 1989. "A mathematical model for periodic scheduling problems", *SIAM Journal on Discrete Mathematics*, vol. 2, pp. 550-581. http://dx.doi.org/10.1137/0402049

SMA, 2018. *Viriato - software for railways.* http://www.sma-partner.ch, Zurich. last accessed 2018/07/01

Van Aken, S., Bešinović, N., Goverde, R.M.P., 2017a. "Designing alternative railway timetables under infrastructure maintenance possessions" *Transportation Research Part B* vol. 98, pp. 224–238

Van Aken, S., Bešinović, N., Goverde, R.M.P., 2017b. "Solving large-scale train timetable adjustment problems under infrastructure maintenance possessions" *Journal of Rail Transport Planning & Management* vol. 7, pp. 141-156

Wüst, R.M., Laube, F., Roos, S., Caimi, G., 2008. *Sustainable Global Service Intention as objective for Controlling Railway Network Operations in Real Time,* In: Proceedings of the WCRR 2008. Seoul

Wüst, R.M., Bütikofer, S., Ess, S., Gomez, C, Steiner A., Laumanns, M., Szabo, J., 2018a. *Periodic timetabling with 'Track Choice'-PESP based on given line concepts and mesoscopic infrastructure* In: Operations Research Proceedings 2018, In press, Springer, Berlin/Heidelberg, Germany

Wüst, R.M., Bütikofer, S., Ess, S., Gomez, C, Steiner A., 2018b. *Development of a prototype for the automated generation of timetable scenarios specified by the transport service intention.* Research Report of SBB Research Fund St.Gallen http://www.hsg.ch

Study on Station Buffer Time Allocation According to Delay Expectation

Xiong Yang[a,b], Yafei Hou[a,b], Li Li[a,b,1], Chao Wen [a,b]

[a] National Engineering Laboratory of Integrated Transportation Big Data Application Technology, Chengdu Sichuan 610031

[b] National United Engineering Laboratory of Integrated and Intelligent Transportation, Southwest Jiaotong University, Chengdu Sichuan 610031

[1] E-mail: speciallili@swjtu.edu.cn, Phone: +86 18581887274

Abstract

Trains are inevitably subject to interference from the external environment and internal systems during operation, leading to delays and conflicts. In this regard, there are usually buffer times allocated at (in) the station (section) in the train timetable, to recover delays. Most of the existing methods that deal with the buffer time allocation mainly consider the length of the section and the traffic density. These methods usually fail to consider the impact of the actual delay of trains, and the buffer time allocation (BTA) is unreasonable. The integration of the actual delay effects into the BTA needs to be resolved. Based on this, in this work, a delay time distribution model was established, and the models were compared according to the standard error of each parameter in the model. Subsequently, based on the delay distribution, a BTA model with weighted average delay expectation time as the objective function was constructed in which the weight coefficients were determined based on the delay strength, and the model was solved by a mathematical analysis method. Different allocation models were designed for different ranges of the total buffer time values. Finally, taking the Dutch railway network trunk section Maarssen–Utrecht Centraal (Mas–Ut) as an example, the results show that the buffer time after redistribution of the BTA model reduces the expected delay time in the segment by 5.25% compared with the original buffer time of the station, indicating that the BTA is reasonable.

Keywords

Buffer time, Delay distributions model, Delay strength, Mathematical analysis

1 Introduction

Trains are inevitably subject to interference from the external environment or internal

systems during operation. When the disturbance intensity is high, the train is delayed. The buffer time set in the train timetable is usually used to eliminate or reduce delays. To make the timetable have enough strain capacity and ensure the punctuality of the train, when the train is in disorder, it can restore normal operation order as soon as possible and make the timetable more flexible. It is often necessary to reserve a certain "buffer" time between the train running lines, which is called the buffer time between the train running lines. The buffer time set in the train timetable is usually used to eliminate or reduce delays.

Zhang et al. (1997) collected a large number of data about the average delay time and buffer time of trains for statistical analysis, and they obtained the change law of the average delay time of trains with the buffer time shown in Figure 1. On the one hand, I in Figure 1 is the train-tracking interval, while the minimum train-tracking interval is $I_{min} = 5$ min, and the buffer time of each train is $I - I_{min}$; On the other hand, the horizontal axis shows the redundant parking time of the train station. The figure shows that the average delay time of trains with various train interval ($I = 6$, 7, and 10 min) and different stopping buffer times tends to decrease with the increase of buffer time. When the station stop buffer time is 6 min, the average train delay time is 10 min when the tracking interval buffer time is 5 min and 20 min when the tracking interval buffer time is 1 min. Buffer time plays an active role in alleviating the fluctuation of the train interval running time and train delays caused by various random factors during train operation. The setting of the buffer time is conducive to improving the stability of the train timetable and enhancing the anti-interference ability of the train timetable.

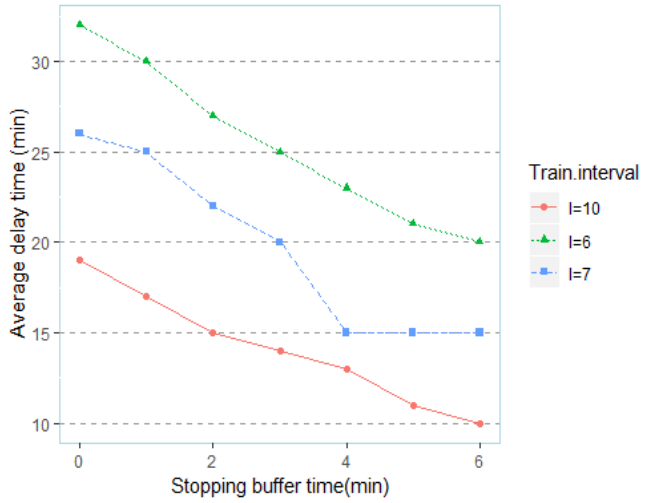


Figure 1: Variation of average train delay time with buffer time

The buffer time has incomplete accumulation, which means that the buffer time is limited to the use of a given station and section. It is shown that the buffer time is used only when the train is disturbed, deviates from the operation plan, and needs to be adjusted. When the train operation adjustment is performed in the current section and station, the buffer time of the previous section and station cannot be stored in the current section and station and has no effect on the train operation adjustment. Similarly, the buffer time that is not fully utilized in the current section and station cannot be accumulated in the station and section ahead of the train operation. Therefore, the excessive buffer time in the train timetable affects the capacity of the section and reduces the efficiency of the transportation organization.

According to the above analysis, to make full use of the buffer time and not waste the capacity, the buffer time allocation (BTA) should consider the actual demand of train delay recovery. In this study, based on the operational performance data, a model of delay time distribution was constructed. Based on this model and considering the impact of actual delay, a BTA model was established with the objective function of minimizing weighted average delay expectation time. In the process of solving the model, the corresponding allocation models were solved by using a mathematical analytic algorithm, aiming at different value ranges of the total buffer time. Finally, the model is validated by a case study. The results show that the established BTA model can reduce the delay expectation time by 5.25%.

The remaining sections of this work are arranged as follows. Section 2 gives an analysis of the current research on buffer time. In Section 3, the relationship between buffer time utilization and delay recovery is discussed, and the rules for buffer time are summarized. In Section 4, a BTA model is established with the objective function of minimizing the weighted average expected delay time, based on the established delayed distribution model and combined with the buffer time, and the model is tested by an example. The conclusions and direction of future research are described in Section 5.

2 Literature Review

Delays seriously affect the order of railway operation. To eliminate or reduce delays, many experts and scholars have done corresponding research. Buffer time is considered the main resource of delay recovery and is closely related to delay recovery. Abril et al. (2008) took Spanish railway infrastructure as an example to analyze the main indicators affecting railway capacity. The results show that railway capacity varies with train speed, train stopping point, distance between railway signals, and robustness of the train timetable. The concept of elasticity was proposed to measure the ability of a railway system to absorb

interference and recover interference (Adjetey-Bahun et al., 2016). The train timetable must be designed with appropriate travel time and be able to withstand delays, disturbances, and changes in operating conditions to achieve a higher level of service during operation (Goverde et al., 2013). Yuan et al. (2007) proposed a new stochastic model for train delay propagation analysis at stations. The model was validated by the example of the Hague Holland Spoor in the Netherlands. The study found that when the planned buffer time between trains at level crossings decreases, the average knock-on delay of all trains increases exponentially. It was pointed out that buffer time in a train timetable has a significant effect on solving and reducing train interference, and the allocation scheme of buffer time affects the possibility of interference (Yuan et al., 2008).

Statistical methods and computational theory have become the main research methods in studying the effect of buffer time on delay recovery. Liebchen et al. (2009) introduced restorable robustness into the study of delay recovery and optimized recovery plans and strategies under resource constraints. In this study, it was assumed that the uncertainty of the time required for train operation and stopping can be obtained from historical data. The proposed method was applied to the Palermo Central Station, and the results show that delay propagation can be largely reduced. Khadilkar et al. (2016) proposed a data-based stochastic model to evaluate the robustness of train timetables that considers delayed recovery. Buffer time and station running time are often used to absorb delay, and the efficiency of delay recovery can be estimated statistically based on empirical data. The average recovery rate obtained from the arrival and departure records of more than 38,000 trains in the Indian Railway Network was 0.13 min/km. However, the number of data in this study was too small — only 15 days of empirical data were available, and it was difficult for the fixed average recovery rate to reflect the real recovery capacity of different sections and stations.

The BTA has become a research hotspot in recent years. In terms of BTA, relevant literature has been studied and some conclusions have been drawn. The buffer time allocated for a single train is generally considered proportional to the section distance of the train, and the average weighted distance was proposed as the basis for BTA (Vromans, 2005; Fischetti et al., 2009). According to the guide “UIC CODE 451-1 OR” (2000), the BTA needs to be calculated according to the train running distance or the average travel time, and the [min/km] or [%] is used to determine the BTA at (in) stations (sections). However, this kind of statistical method does not allocate buffer time by trains, stations, and sections in accordance with specific conditions. Kroon et al. (2008) distributed the buffer time by establishing a stochastic optimization model to increase the robustness of the train timetable. The model was tested and verified with the Dutch train passenger train timetable. Vansteenwegen et al. (2007) calculated the ideal BTA in sections by using negative

exponential distributions. They constructed the delay loss equation on this basis and optimized the timetable by using a linear programming method. Carey et al. (2007) applied probability theory to determine the reasonable buffer time under the condition of train operation performance, but they did not consider the impacts of delay. Krasemann et al. (2012) used the depth-first greedy algorithm to assist with train operation adjustment planning. The buffer time is a tool to eliminate random interference of train operation, but there is no in-depth study of the BTA. Carey (2007) and Dewilde (2013) have made a series of studies on how to allocate buffer time in the process of compiling train timetables and achieved certain results.

Because of the difficulty in acquiring and processing operational performance data, the above literature seldom addressed the BTA based on operational performance data. In recent years, more and more researchers have used machine-learning methods to study the BTA. Huang et al. (2018) established a data-driven BTA model based on the Wuhan–Guangzhou high-speed railway. Based on the utilization of buffer time, the model redistributes buffer time, which provides a new research method for BTA. Wen et al. (2016) proposed a data-driven method based on a multiple linear regression model and stochastic forest model to solve the problem of delay recovery of high-speed rail trains after initial delays. In addition, under the same explanatory variables and datasets, the stochastic forest regression proposed is superior to the over-limit learning machine and stochastic gradient descent methods (Bottou, 2010; Huang et al., 2004). Therefore, on the premise of data availability, it has become an inevitable trend to discover rules from data and construct models to study the BTA.

However, the existing literature on BTA mainly considers the length of the interval and the driving density, and rarely considers the influence of the actual delay strength. It is especially important to integrate the delay effects into the BTA, and the delay distributions can effectively evaluate the delay effects, which can be used as an entry point for the BTA. Therefore, it is of great significance to study the BTA based on the delay distributions.

3 Relationship between Buffer Time Utilization and Delay Recovery

The BTA needs to consider various factors comprehensively to achieve the scientific and rational selection of buffer time. The International Railway Union standardized the selection of train operation buffer time. In terms of operating mileage, it is 1.5 min for every 100 km of single-engine passenger trains. For multimachine traction, it is compensated for 1 min per 100 km. In terms of travel time, the buffer time needs to be based on the running speed of the train, which ranges from 3% to 7% of the total travel time.

Generally, delay recovery mainly depends on buffer time, which can be used to restore

the train to the planned train timetable as soon as possible. As shown in Figure 2, t_j^i represents the minimum stop operation time of train i at station s_j ; $t_{j,j+1}^i$ represents the minimum running time of train i between station s_j and station s_{j+1} ; b_j^i and $b_{j,j+1}^i$ represent

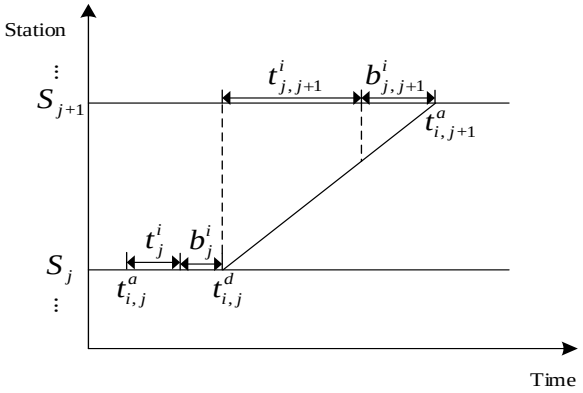


Figure 2: Schematic diagram of buffer time utilization in stations (sections)

the buffer time of the station and interval, respectively; $t_{i,j}^a$ represents the actual arrival time of train i at station s_j ; and $t_{i,j}^d$ represents the actual departure time of train i at station s_j .Then, there is

$$b_j^i = t_{i,j}^d - t_{i,j}^a - t_j^i. \tag{1}$$

$$b_{j,j+1}^i = t_{i,j+1}^a - t_{i,j}^d - t_{j,j+1}^i. \tag{2}$$

Based on buffer time, train delay recovery can be described as follows.

(1) If the delay time of the train at station s_j is $d_j^i \le b_j^i$, it indicates that the delay time can be absorbed by the buffer time of s_j , thus achieving the effect of delay recovery.

(2) If $b_j^i < d_j^i \le b_j^i + b_{j,j+1}^i$, it shows that the delay cannot be absorbed completely by the station buffer time, but the part that is not absorbed completely can be absorbed by the interval buffer time, so as not to affect the station arrival time, thus achieving the delay recovery effect.

(3) If $d_j^i > b_j^i + b_{j,j+1}^i$, it indicates that the delay cannot be absorbed by the station buffer time and interval buffer time, and the delay is propagated at station s_{j+1} .

The analysis of three cases of train delay recovery clearly shows that the buffer time has the effect of delay recovery, but the effect is closely related to the length of the specific delay time. Considering the buffer time separately from the delay situation either wastes the buffer time or makes the delay recovery effect not obvious. Therefore, in this work, the BTA was studied by comprehensively considering the actual impact of delay. First, the buffer

time was combined into the delay time distribution model, and the expected delay time was calculated based on the distribution model. Then, taking the delay strength as the weight coefficient of the expected delay time of each station, a BTA model with the minimum expected total delay time as the objective function was established. Finally, the buffer time after reallocation could be obtained by solving the model.

4 BTA Model

4.1 Model establishment

Delays in the section will show up at the station. For example, when the train runs in the section, it is delayed 2 min. If the buffer time in the section is not considered, the delay will be expressed as the train arrival delay at the station, and the arrival delay time is also 2 min. Therefore, the delay of the section can be analyzed by the station, and the buffer time of the sections can be summarized as the station buffer time — that is, the running time of the train in all the sections is assumed to be the minimum running time, and the train is assumed to be on time at the originating station.

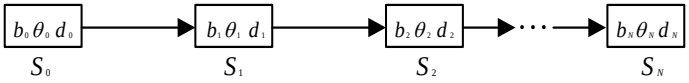


Figure 3: Schematic diagram of train operation

Figure 3 is a schematic diagram of a train operating at N stations, using the collection T record of the station where the train arrives, the delay time distributions at the station, and the buffer time at the station, which is $T_i = \{S_i, b_i, \theta_i, d_i \mid i = 0,1,2,\cdots N\}$. Here, S_i indicates the i-th station, b_i shows the buffer time assigned to station i, and θ_i shows the delay distributions in the i-th station, and d_i is the delay time at the i-th station ($d_i \geq 0$).

In the BTA model, delays in the interval are generalized to delays at the station, thus simplifying the BTA model. The problem of BTA at (in) the station (section) is transformed into a whole for research and analysis, which has no effect on the BTA result.

Assuming that the total amount of buffer time is constant, there are:

$$\sum_{i=1}^{N} b_i = b. \quad and \quad b_i \geq 0. \tag{3}$$

where b represents the total buffer time.

Making $f_i(\theta)$ indicate the delay distributions density function of station i, then the probability that train departures from S_0 are on time to S_1 with a delay time d_1 that less than

or equal to x is:

$$P\{d_1 \le x\} = \int_0^{x+b_1} f_1(\theta)d\theta. \tag{4}$$

The delay mathematics expectation at S_0 is:

$$E(d_1) = \int_0^{+\infty} \theta f_1(\theta + b_1)d\theta. \tag{5}$$

When the train is running, if the delay time d_{i-1} is generated at S_{i-1}, and $d_{i-1} > b_{i-1}$, then the delay will spread to S_i. Therefore, the probability that the delay time of the train at S_i is less than or equal to x is:

$$P\{d_i \le x\} = \int_0^{x+b_i} \int_0^{b_{i-1}+x+b_i-\theta} f_i(\theta) f_{i-1}(\eta)d\theta d\eta. \tag{6}$$

The delay mathematics expectation at S_i is:

$$E(d_i) = \int_0^{+\infty} \int_0^{+\infty} \theta f_i(\theta + b_i)\eta f_{i-1}(\eta + b_{i-1})d\eta d\theta. \tag{7}$$

Therefore, the average delay expectation of the train during the entire operation can be calculated as:

$$E(d) = \frac{1}{N}\sum_{i=1}^{N} E(d_i). \tag{8}$$

Because the delay strength can be used to evaluate the frequency and severity of the delay, different weights are given to the delay mathematical expectation of each station according to the delay strength. Then, Eq. (8) is amended to the Eq. (9):

$$\begin{cases} E(\bar{d}) = \sum_{i=1}^{N} w_i E(d_i) \\ \sum_{i=1}^{N} w_i = 1 \\ w_i > 0 \end{cases}. \tag{9}$$

In Eq. (9), w_i is the expected weight coefficient of the delay at S_i ,which is determined based on the delay strength. Therefore, if $E(\bar{d})$ in Eq. (9) is minimized, the BTA function can be obtained as follows:

$$\min E(\bar{d}). \tag{10}$$

In summary, Eq. (10) is a BTA function, and Eq. (3) to Eq. (9) are constraints.

4.2 Delay distribution model

In the construction of the BTA model, the key is to solve the problem of the delay time distribution. This part focuses on the construction of the delay time distribution model. The research idea is to select the common data distribution model to fit the delay time based on the delay time data and take the standard error of each parameter in the distribution model as the model comparison criterion, to select the optimal delay time distribution model.

Based on the train operation performance data in Maarssen–Utrecht Centraal (Mas–Ut) of the Dutch railway network trunk section, the BTA under the condition of continuity was studied. This section contains three stations: Maarssen (Mas), Utrecht Zuilen (Utzl), and Utrecht Centraal (Ut). The time span of operational performance data in the segment was three months, and the data volume was 122,480, of which there were 27,728 delay records. After the screening and noise reduction of the delay data, the delay time distribution model at the station was established based on this. The lognormal distribution, exponential distribution, and Weber distribution models were selected to study the delay distributions. Based on the station delay data, the above models were used to fit the station delay data.

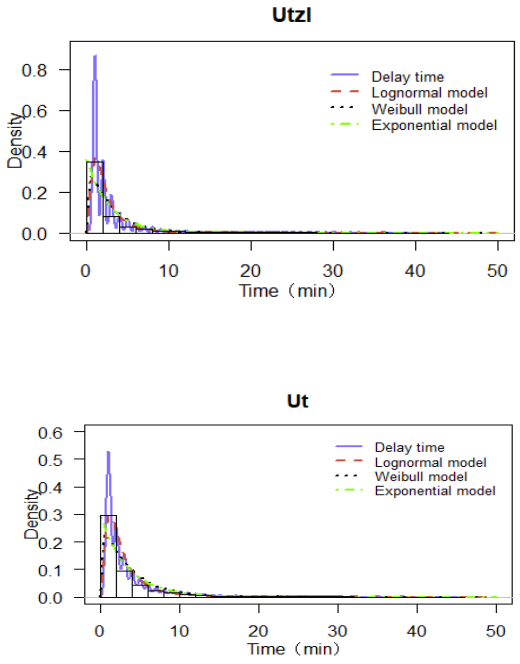


Figure 4: Fitting diagram of station delay distributions

Figure 4 is a schematic diagram of the lognormal distribution, Weber distribution, and exponential distribution used to fit the probability density of the station delay time. To compare the above models, the optimal delay distribution model was determined by comparing the standard error of parameters in each model as the criterion (Maas, 2004). The standard error of each model parameter was calculated, and the results are shown in Table 1.

According to the results in Table 1, compared with other models, the standard error of the model parameters of the exponential distribution model is the smallest, so the exponential distribution model was selected as the station delay distribution model.

Table 1: Standard error of model parameters

Station	Distribution Model	Parameters	Standard error
Utzl	Exponential distribution	rate	0.0038
	Lognormal distribution	meanlog	0.0081
		sdlog	0.0058
	Weber distribution	shape	0.0076
		scale	0.0301
Ut	Exponential distribution	rate	0.0025
	Lognormal distribution	meanlog	0.0071
		sdlog	0.0050
	Weber distribution	shape	0.0063
		scale	0.0295

After determining the station delay distribution model, the maximum-likelihood algorithm was used to solve the parameters of the exponential distribution model, and the station delay distribution model was obtained, as shown below.

$$f_1(\theta)=\begin{cases}\lambda_1 e^{-\lambda\theta}, \theta\geq 0\\ 0 \quad\quad , \theta<0\end{cases}=\begin{cases}0.293e^{-0.293\theta}, \theta\geq 0\\ 0 \quad\quad\quad\quad , \theta<0\end{cases}. \tag{11}$$

$$f_2(\theta)=\begin{cases}\lambda_2 e^{-\lambda_2\theta}, \theta\geq 0\\ 0 \quad\quad , \theta<0\end{cases}=\begin{cases}0.316e^{-0.316\theta}, \theta\geq 0\\ 0 \quad\quad\quad\quad , \theta<0\end{cases}. \tag{12}$$

where $f_1(\theta)$ and $f_2(\theta)$, respectively, represent the delay distribution density function of Utzl and Ut. $\lambda_1=0.293$ and $\lambda_2=0.316$ are, respectively, parameters of the delay distribution density function.

4.3 Delay expectation time model based on buffer time optimization

A delay expectation model considering buffer time optimization was built based on the delay time distribution model. The redistributed buffer time can be obtained by solving the model. For the convenience of the following statement, stations Utzl and Ut are replaced with S_1 and S_2, respectively. The buffer time allocated by S_1 and S_2 is represented by b_1 and b_2, respectively. From Eq. (3), there are $b = b_1 + b_2$ and $b_1 \geq 0$, $b_2 \geq 0$.

The probability that the delay time d_1 of train at S_1 is less than or equal to x is:

$$P\{d_1 \leq x\} = \int_0^{x+b_1} \lambda_1 e^{-\lambda_1 \theta} d\theta = 1 - e^{-\lambda_1 (b_1 + x)}. \tag{13}$$

Then, after increasing the buffer time b_1, the delay probability density function of S_1 is:

$$g_1(x) = \frac{dP\{d_1 \leq x\}}{dx} = \lambda_1 e^{-\lambda_1 (b_1 + x)}. \tag{14}$$

According to Eq. (14), the expected delay time of the train at S_1 is:

$$\begin{aligned} E(d_1) &= \int_0^{+\infty} x g_1(x) dx = \int_0^{+\infty} x \lambda_1 e^{-\lambda_1 (b_1 + x)} dx \\ &= \frac{1}{\lambda_1} e^{-\lambda_1 b_1}. \end{aligned} \tag{15}$$

For S_2, it is necessary to consider the delay time generated on S_1. Figure 2 shows that delays generated on S_1 can be absorbed through the buffer time of S_1 and S_2, while delays on S_2 can only be absorbed through the buffer time by S_2. Therefore, the probability of the train at S_2 with a delay time $d_2 \leq x$ is:

$$\begin{aligned} P\{d_2 \leq x\} &= \int_0^{x+b_2} \int_0^{b_1 + x + b_2 - \theta} f_2(\theta) f_1(\eta) d\eta d\theta \\ &= \int_0^{x+b-b_1} \int_0^{x+b-\theta} \lambda_2 e^{-\lambda_2 \theta} \lambda_1 e^{-\lambda_1 \eta} d\eta d\theta \\ &= 1 - e^{-\lambda_2 (b - b_1 + x)} - \frac{\lambda_2}{\lambda_1 - \lambda_2} e^{-\lambda_1 (b + x)} [e^{(\lambda_1 - \lambda_2)(b - b_1 + x)} - 1]. \end{aligned} \tag{16}$$

The delay probability density function of S_2 is:

$$\begin{aligned} g_2(x) &= \frac{dP\{d_2 \leq x\}}{dx} \\ &= \lambda_2 e^{-\lambda_2 (b - b_1)} e^{-\lambda_2 x} + \frac{{\lambda_2}^2}{\lambda_1 - \lambda_2} e^{-\lambda_1 b_1 - \lambda_2 (b - b_1)} e^{-\lambda_2 x} - \frac{\lambda_1 \lambda_2}{\lambda_1 - \lambda_2} e^{-\lambda_1 (b + x)}. \end{aligned} \tag{17}$$

According to Eq. (17), the expected delay time of the train at S_2 can be obtained as follows:

$$\begin{aligned} E(d_2) &= \int_0^{+\infty} x g_2(x) dx \\ &= \frac{1}{\lambda_2} e^{-\lambda_2 b} e^{\lambda_2 b_1} + \frac{1}{\lambda_1 - \lambda_2} e^{-\lambda_2 b} e^{(\lambda_1 - \lambda_2) b_1} - \frac{\lambda_2 e^{-\lambda_1 b}}{\lambda_1 (\lambda_1 - \lambda_2)}. \end{aligned} \tag{18}$$

After $E(d_1)$ and $E(d_2)$ are obtained, the expected delay time weighting coefficients of station S_1 and S_2 are determined according to the delay strength of station. The calculation formula for the delay strength is shown in Eq. (19).

$$q = \frac{m * k}{c * l * z}. \tag{19}$$

In Eq. (19), q is the delay strength, which is an indicator of the influence of the delayed train number; m indicates that delays affect the number of trains; c represents the traffic volume; l represents the length of the sections; z represents the effective working day; and k is a constant, and its role is to convert the value of q to (0, 1). Based on the delay strength and combined with Eq. (20) , the weight of delay expectation of S_1 and S_2 are determined to be w_1 and w_2.

$$w_i = \frac{q_i}{\sum_{i=1}^{n} q_i}. \tag{20}$$

To sum up, the expected weighted delay time of trains in this trunk line section is:

$$E(\bar{d}) = w_1 E(d_1) + w_2 E(d_2). \tag{21}$$

Substitute Eqs. (15) and (18) into Eq. (21) to obtain:

$$E(\bar{d}) = w_1 \frac{1}{\lambda_1} e^{-\lambda_1 b_1} + w_2 [\frac{1}{\lambda_2} e^{-\lambda_2 b} e^{\lambda_2 b_1} + \frac{1}{\lambda_1 - \lambda_2} e^{-\lambda_2 b} e^{(\lambda_1 - \lambda_2) b_1} - \frac{\lambda_2 e^{-\lambda_1 b}}{\lambda_1 (\lambda_1 - \lambda_2)}]. \tag{22}$$

Then, one can solve Eq. (22) to obtain the minimum value of b_1^*, that is, the optimal buffer time at S_1, and the optimal buffer time on S_2 is $b - b_1^*$. Take the derivative of b_1 in Eq. (22) and set the derivative result equal to 0, which is:

$$w_2 (e^{\lambda_1 b_1} - 1) = w_1 e^{\lambda_2 (b - b_1)}. \tag{23}$$

Eq. (23) shows that the function on the left side of the equation increases as b_1 increases, and the function on the right side of the equation decreases as b_1 increases. Then, in $0 \le b_1 \le b$, there is an optimal solution, that is, Eq. (23) is solvable, but it is not easy to solve Eq. (23) directly, and it can be solved by the approximate estimation method.

(1) When $0 \le b < 1$ is equal to $0 \le b_1 < 1$, the Taylor formula is used to expand and simplify the exponential function to obtain:

$$\lambda_1 \lambda_2 w_2 b_1^2 + \lambda_1 w_2 b_1 - w_1 e^{\lambda_2 b} = 0. \tag{24}$$

By solving Eq. (24), one can obtain:

$$\begin{cases} b_1^* = \dfrac{-\lambda_1 w_2 + \sqrt{(\lambda_1 w_2)^2 + 4\lambda_1 \lambda_2 w_2 w_1 e^{\lambda_2 b}}}{2\lambda_1 \lambda_2 w_2} \\ b_2^* = b - b_1^* \end{cases}. \tag{25}$$

(2) When $b \geq 1$, the approximate estimation of Eq. (23) is:

$$\begin{cases} w_1 = e^{\lambda_1 b_1} - 1 \\ e^{\lambda_2 (b - b_1)} = w_2 \end{cases}. \tag{26}$$

By solving Eq. (26), one can obtain:

$$\begin{cases} b_1^* = \dfrac{\lambda_2 \ln w_1 + \lambda_1 (b\lambda_2 - \ln w_2)}{2\lambda_1 \lambda_2} \\ b_2^* = b - b_1^* \end{cases}. \tag{27}$$

4.4 Case study

The BTA was studied by taking the main line section Mas–Ut as an example. The buffer times allocated by the stations S_1 and S_2 were 3 and 2 min, respectively, and the buffer time allocated in the section was 5 min. That is, $b_1 = 3$, $b_2 = 2$, and $b = 5$, which can be used to calculate $E(\overline{d})$=2.170.

The expected weight coefficients of delay of S_1 and S_2 were determined to be 0.58 and 0.42, respectively, through Eq. (19), that is, w_1=0.58 and w_2=0.42. With the established BTA model, $b_1^* = 2.943$ and $b_2^* = 2.057$ can be obtained; then, $E(d_1) = 1.441$ and $E(d_2) = 2.905$ can be calculated, and, finally, $E(\overline{d^*}) = 2.056$.

Figure 5 shows that the delay expectation $E(\overline{d^*})$ after the BTA model is 0.114 min lower than the delay expectation $E(\overline{d})$ without the model — it was 5.25% lower. Therefore, the BTA model is effective. What is more, the buffer time focuses on the allocation of S_2. This measure can effectively reduce the expected delay time in the segment, provide a relevant basis for scheduling decisions, and help improve the efficiency of the work organization at (in) the stations (sections).

In conclusion, the BTA model established can consider the actual impact of delays. It provides a relevant research idea for the research of buffer time allocation based on operational performance data. Although the model only analyzes the BTA of several stations, the application of multiple stations remains to be studied. However, the results of the case study show that the model is reasonable and can be used to allocate buffer time between main stations in the trunk section.

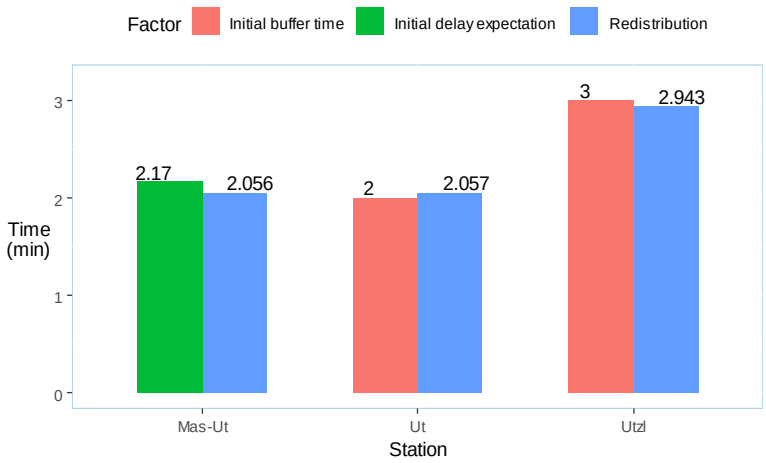


Figure 5: Comparison chart after BTA model

5 Conclusions

According to the delay distributions, a BTA model was established with the expected delay time as the objective function, realizing the redistribution of station buffer time, and the BTA model was verified by the Mas–Ut trunk section of the Dutch railway network. The results indicate the following.

(1) For the case where the total buffer time of the trunk section was different, the formula for assigning the station buffer time is given in Eqs. (24) and (26). The BTA formula shows that the delay distributions and delay strength have a certain influence on the BTA.

(2) The BTA model based on the delay distributions has a good effect on the redistribution of buffer time. By redistributing the buffer time of the stations Utzl and Ut, compared with the buffer time allocated before the station, after the BTA model, the total delay expectation time of the trunk segment decreased by 5.25%.

In conclusion, the dispatcher can adjust the work organization of the station according to the buffer time after BTA, to reduce the occurrence of station delays and improve the work efficiency of the station. Planned future work is the study of the BTA of the operation route and local network based on the BTA model of the trunk section. It is expected that the redistribution of buffer time can effectively reduce the delay of the operation route and local network and improve the delay recovery ability of the operation route and local road network.

Acknowledgments

This work was supported by the National Nature Science Foundation of China [grant number 71871188] and the Science & Technology Department of Sichuan Province [grant number 2018JY0567]. We are grateful for the contributions made by our project partners.

References

Zhang, X. C, Hu, A. Z., 1997. Computer simulation study on the distribution model of scattering reserve capacity on operation line, *Journal of Beijing Jiaotong University*, 603-608.

Abril, M., Barber, F., Ingolotti, L., Salido, M. A., Tormos, P., Lova, A., 2008. An assessment of railway capacity. *Transportation Research Part E: Logistics and Transportation Review*, 44(5), 774-806.

Adjetey-Bahun, K., Birregah, B., Châtelet, E., Planchet, J. L., 2016. A model to quantify the resilience of mass railway transportation systems, *Reliability Engineering & System Safety*, 153, 1-14.

Goverde, R.M.P., Hansen, I. A., 2013. Performance indicators for railway timetables, In *Intelligent Rail Transportation (ICIRT), 2013 IEEE International Conference on* (pp. 301-306). IEEE.

Yuan, J., Hansen, I.A., 2007. Optimizing capacity utilization of stations by estimating knock-on train delays, *Transportation Research Part B: Methodological*, 41(2), 202-217.

Yuan, J., Hansen, I.A., 2008. "Closed form expressions of optimal buffer times between scheduled trains at railway bottlenecks," In *Intelligent Transportation Systems, 2008. ITSC 2008. 11th International IEEE Conference on* (pp. 675-680). IEEE.

Liebchen, C., Lübbecke, M., Möhring, R., Stiller, S., 2009. The concept of recoverable robustness, linear programming recovery, and railway applications, In *Robust and online large-scale optimization* (pp. 1-27). Springer, Berlin, Heidelberg.

Khadilkar, H., 2016. Data-enabled stochastic modeling for evaluating schedule robustness of railway networks, *Transportation Science*, 51(4), 1161-1176.

Vromans, M., 2005. *Reliability of railway systems* (No. 62).

Fischetti, M., Salvagnin, D., Zanette, A., 2009. Fast approaches to improve the robustness of a railway timetable, *Transportation Science*, 43(3), 321-335.

UIC, U., 2000. Code 451-1 Timetable Recovery margins to guarantee timekeeping-Recovery margins.

Kroon, L., Maróti, G., Helmrich, M.R., Vromans, M., Dekker, R., 2008. Stochastic improvement of cyclic railway timetables, *Transportation Research Part B:*

Methodological, 42(6), 553-570.

Vansteenwegen, P., Van Oudheusden, D., 2006. Developing railway timetables which guarantee a better service, *European Journal of Operational Research*, 173(1), 337-350.

Carey, M., Crawford, I., 2007. Scheduling trains on a network of busy complex stations, *Transportation Research Part B: Methodological*, 41(2), 159-178.

Krasemann, J.T., 2012. Design of an effective algorithm for fast response to the re-scheduling of railway traffic during disturbances, *Transportation Research Part C: Emerging Technologies*, 20(1), 62-78.

Carey, M., Carville, S., 2000. Testing schedule performance and reliability for train stations, *Journal of the Operational Research Society*, 51(6), 666-682.

Dewilde, T., Sels, P., Cattrysse, D., Vansteenwegen, P., 2013. Robust railway station planning: An interaction between routing, timetabling and platforming, *Journal of Rail Transport Planning and Management*, 3(3), 68-77.

Huang, P., Wen, C., Peng, Q., Lessan, J., Fu, L., Jiang, C., 2018. A data-driven time supplements allocation model for train operations on high-speed railways, *International Journal of Rail Transportation*, 1-18.

Wen, C., Lessan, J., Fu, L., Huang, P., Jiang, C., 2017. "Data-driven models for predicting delay recovery in high-speed rail," In *Transportation Information and Safety (ICTIS), 2017 4th International Conference on* (pp. 144-151). IEEE.

Bottou, L., 2010. Large-scale machine learning with stochastic gradient descent, In *Proceedings of COMPSTAT'2010* (pp. 177-186). Physica-Verlag HD.

Huang, G.B., Zhu, Q.Y., Siew, C. K., 2004. Extreme learning machine: a new learning scheme of feedforward neural networks, In *Neural Networks, 2004. Proceedings. 2004 IEEE International Joint Conference on* (Vol. 2, pp. 985-990). IEEE.

Maas, C.J., Hox, J.J., 2004. The influence of violations of assumptions on multilevel parameter estimates and their standard errors, *Computational statistics & data analysis*, 46(3), 427-440.

Assessment of energy and emissions saving solutions in urban rail-based transport systems

Mohammad Hassan Davoudi Zavareh[a], Stefano Ricci[a]
[a] Sapienza Università di Roma - DICEA
Via Eudossiana, 18 – 00184 Roma, Italy
[1] E-mail: stefano.ricci@uniroma1.it, Mh.davoudi93@gmail.com
Phone: +39 06 44585144

Abstract

Global warming and climate change are indisputable theories. Since the Industrial Revolution, the mean temperature of the planet has increased by 1°C. Now, temperatures are approaching a higher stage of +1.5°C and the attention is on both CO_2 emissions and energy consumption. Transportation is a major component of the environmental impact, accounting for approximately 30% of air pollution and energy consumption. Due to the rapid urbanization in the EU, with an estimated 74.3% of the population living in cities, forecasted to rise to 80% by 2050, urban mobility is dramatically increasing its relevance. Therefore, a reduction in energy consumption and pollutant emissions is a crucial factor to consider in developing urban transportation and particularly rail-based systems, able to provide energy saving transport services by improving urban environment. Several methods and techniques are under development to improve the energy performance of Light Rail Transport (LRT), which spread from different typologies of power supply to improving energy efficiency. The purpose of this paper is to start from the last developments and innovative energy sources for LRT systems. The focus is on two parts: a) trams running on Hydrogen in parallel with on board batteries with energy saving control techniques, b) potential renewable energy sources to meet power demand. The comparison is with traditional power sources and equipment (e.g. Catenary-based). The methods, based on selected indicators, are under development and test by calculations and simulations with reference to the case study of the new tramlines in the city of Brescia (Italy).

Keywords
Railways, Tram, Urban Transit, Renewable Energy, Fuel cell, Hydrogen

Introduction

Energy demand globally increased by 2.1% in 2017, according to IEA, more than twice the average growth rate over the previous five years, which was 0.9%. According to Global Energy and CO2 Status Report (OCED International Energy Agency, 2018) energy-related CO2 emissions grew by 1.4% in 2017, which was a record through the history, after three years of emissions remaining flat worldwide.
On the other hand, renewable energies had the highest growth rate of any energy source in 2017, meeting a quarter of global energy demand growth last year. China and the United States together accounted for half of the increase in renewables-based electricity generation, followed by the European Union (8%), Japan and India (with 6% of growth each).
The EU set an ambitious target of 40% greenhouse emission reduction by 2030, and 80% by 2050 (EC, 2016a). Based on the Paris agreement, adopted on 12 December 2015, at COP21 and signed by 195 states in 2016, the EU is promoting the following target (Council of the EU, 2016): *Holding the increase in the global average temperature to well below*

2°C above pre-industrial levels and limiting the temperature increase to 1.5°C. To foster low carbon transition, a framework strategy for a resilient Energy Union links the transport and energy systems (EC, 2016a, 2016b). Its key features are:

- Reduction of the dependency on particular fuels, energy suppliers and routes;
- Full integration of the internal energy market and more efficient energy consumption;
- Decarbonisation of the economy.

Mobility within cities and between suburban areas and towns is significantly important, since transport represents more than 30% of the final energy consumption in Europe (EC, 2016b) and the majority of EU population are urbanised. According to UITP (2016), in 2014 urban rail accounted for 44.3% of all local public transport journeys in Europe (13.6% suburban rail, 16.2% metro, 14.5% tram/LRT).
As the European economy and transport demand are continuing to grow, the mentioned aims are only achievable if attentions of policy makers, local and companies' authorities pulled them in. In this framework, the article promotes:

- Firstly, a comparative assessment of the traditional power supply (catenary based) in LRT systems with modern renewable energy sources;
- Secondly, new methods for a better climate adoptable, enhanced passenger comfort and finally improve urban environment by removing catenary-based infrastructure.

The innovative concept includes the possibility to either transfer energy supply to street ground surface with such systems like third rail electrification and magnetic fields or moving the energy source on board to practically remove the catenary infrastructure.
The most common on-board source are nowadays batteries, but they are expensive, heavy, they required the extensive use of rare earth metals and the production of lithium-ion batteries itself is an energy-intensive process; Furthermore, charging them take a long time. Another way to carry clean energy source is using the most abundant element of universe, Hydrogen, that already used in automobile industry but rarely in railway, especially urban. Hydrogen has specific energy up to 40,000 Wh/kg, comparing to only 278 Wh/kg for batteries, which is 236 times more and makes using fuel cell vehicles a feasible choice.
Toyota Mirai (2014), Fuel cell-powered 113 kW with a total range of almost 500 km with only 5 kg of Hydrogen, was one of the first mass production Fuel Cell Vehicle (FCV) with a combination of power train, runs on both Hydrogen and Battery with variable energy consumption. The development of FCVs followed by other automotive companies (Honda, Hyundai) and Locomotive companies.
In September 2018, Alstom commercialised the world's first hydrogen powered train, the Coradia iLint that entered passenger service in Lower Saxony, Germany. The two pre-series trains, homologated by the German Federal Railway Association in July, are now running over the cities of Cuxhaven, Bremerhaven, Bremervörde and Buxtehude. The train is able to operate over a daily range of 1000 km. Alstom and the local transport authority of Lower Saxony (LNVG) signed a contract for the delivery of 14 hydrogen fuel cell trains by 2021.
Moreover, the most relevant FCV in mass urban transit ran in October 2017 in China: CRRC Tangshan Railway Company unveiled a prototype low-floor LRV powered by Canadian

supplier Ballard Power Systems' hydrogen fuel cell technology, FCveloCity, which trialled on the new 14 km light rail line in Tangshan, China.
Ballard's fuel cell technology work in combination with batteries and super-capacitors over a range of 40 km, top speed of 70 km/h and capacity of 336 passengers to offer entirely catenary-free operation on the line. The LRVs have a range of 40 km on a single 12 kg hydrogen fill-up, which takes 15 minutes to complete. The four-station line includes a 100 kg capacity hydrogen refilling station.
Finally, for better understanding differences, potentials and drawbacks of traditional catenary based system and fuel cells, the comparison, calculation and simulations are under development for the new tramline project in Brescia (Italy) basing on infrastructural and operational data concerning both line and vehicles.
The new tramway network includes three main sections (Fig.1) of double track lines: T1, T2 and T3 for a total extension of about 23 km (46 km of single equivalent track), 65% shared with current urban road including approximately 41 signalized intersections. The project cost would be 450,000,000€, forecasted to be operated by 2026 with a total number of 14040 pax/h in rush hours with estimated yearly demanded power of 9,000,000 kW.

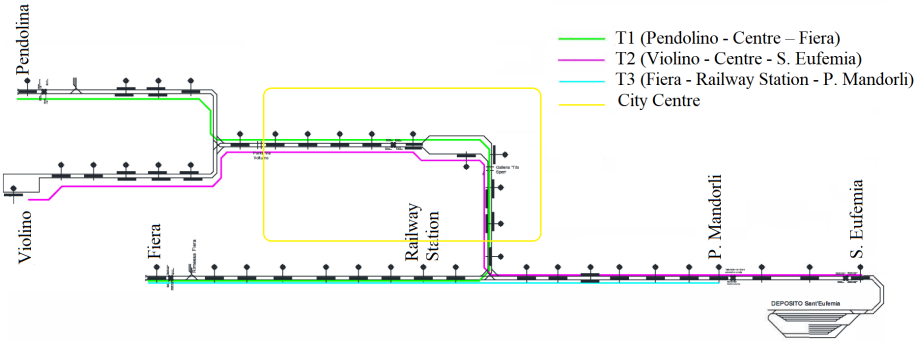


Figure 1: Track layout

The paper will forecast possible environmentally friendly clean local power sources to cover at least 20% of demanded power. Moreover, it promotes low-cost technology to make renewable hydrogen using sunlight and any source of water (Hyper Solar Inc.) directly at or near the depot area, to make a self-sustained renewable zero carbon Hydrogen powered urban transit combined with intelligent energy management (Fig.2).

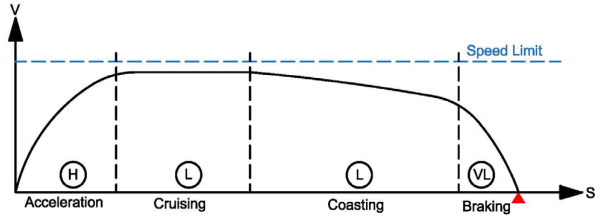


Figure 2: Energy-saving driving strategies
(Energy demand level H: High, L: Low, VL: Very Low)

Comparisons are following by discussing the most important issues concerning using

Hydrogen, such as safety, infrastructure and cost. Simulations are carried out by OPEUS simulator which is part of the Shift2Rail and stands for "Modelling and strategies for the assessment and OPtimisation of Energy Usage aspects of rail innovation", and is aiming to develop a simulation methodology and accompanying modelling tool to evaluate, improve and optimise the energy consumption of rail systems with a particular focus on in-vehicle innovation. The OPEUS concept is based on the need to understand and measure the energy being used by each of the relevant components of the rail system and in particular the vehicle. This includes the energy losses in the traction chain, the use of technologies to reduce these and to optimise energy consumption (e.g. ESSs). Specifically, the OPEUS approach has three components at its core: The energy simulation model, the energy use requirements (e.g. duty cycles) and the energy usage outlook and optimisation strategies recommendation.

Why Hydrogen?

Electricity production is carbon intensive, releases massive heat and noise at local power plants, gives rise to negative impacts on the environment and human health throughout all stages of its lifecycle, from resource extraction to electricity use. Impacts on climate change, air and water quality, direct and indirect impacts on land resources, etc. Impacts stemming from electricity production depend on the (fossil) fuel employed, how it was extracted and processed, the actual technology (and its efficiency) used to produce electricity, as well as the use of abatement technologies. An almost full decarbonisation of the electricity sector will be needed in order to meet the EU's objective of reducing greenhouse gas emissions by 90% by 2050. Increasing electricity generation and use throughout Europe without reforming the current energy system will lead to higher overall health and environmental impacts. Nevertheless, an increase in electricity consumption in the transport sector might signal a positive modal shift towards rail transport or a higher penetration of electric vehicles. The carbon intensity of total electricity generation in EU in 2016 was 296 g CO2/kWh. 1207 million tonnes of CO2 emitted from electricity generation in EU in 2013, leading by Germany 332 Mt, UK 163 Mt and Italy with 111Mt. An increase in electricity consumption in transport sector (mainly railways) arose in countries such as France and Italy.

Instead, Hydrogen and Fuel cell technology can contribute significantly towards reducing emissions and facilitating the necessary green energy transition in EU regions and cities. Fuel-cell technology usage can improve air quality and create positive health impacts for local population, hence enhancing life quality. Regarding recent study, about 90 cities in EU planned to invest about 1.8 billion euros in coming five years to deploy different H_2 transport modes and electrolysers for H_2 production and power generation. This conversion is expected to have not only environmental but also local economic effects.
For example, according to Unione Petrolifera, in 2017 Italy imported 15.9 million tonnes of refined petroleum products. However, the petroleum industry employs relatively few people; historical data shows that 1 million euro of value added in the petroleum sector in Italy created only 3.5 jobs in 2017, while hydrogen sectors are almost 5 times more labour intensive. Overall, the transition towards low and zero carbon economy has a net positive impact on employment (19,225 additional jobs in 2030), and will create opportunities for the adaptation and transformation of workers. As another example, German state of Baden-Württemberg *FCV and H_2 for green energy in EU cities and regions* estimate a value added of around 680 million euros by the year 2030. The Hydrogen council's vision is to create

around 30 million additional jobs globally as well as sales of approximately 2000 billion $ by 2050 should Hydrogen become a global energy carrier. Which could then serve up to 18% of global energy demand.

Why Fuel cell Tram?

Fuel cells have achieved enough maturity to support railway sector as they are already on tracks in Germany and is being considered for operation in France and UK as inter-city trains. A fuel cell approach in urban area has the following characteristics and benefits, compared to overhead caternary systems:

- Less expensive due to much less infrastructure;
- Less impacting on urban area surroundings and operations;
- Visually more attractive, specially in old city centres and tourist attractions;
- Extendable to additional locations without additional infrastructure;
- Operable in power outage, independently powered;
- Enabling other independently powered alternatives in future;
- Operating with zero emissions and less pollution at electricity production plants.

This approach comparing to traditional Overhead Caternary System (OCS) eliminates the need for the overhead catenary wires, support poles, notching the existing tunnels and electrical substations. Costs related to removing or trimming trees and relocating existing electrical and telecommunication infrastructure, etc. would be reduced or eliminated. In the case study, the elimination of 13 electrical substations, with pitch of about 2 km each, provide power in two electrical zones through 750 Volt power supply, one dedicated transformer in each substation for auxiliaries. Estimated saving is 30 million euros (cost of OCS infrastructure). Additionally, caternary-free approach could make use of abandoned existing tram tracks with reduced or no additional infrastructure upgrades or costs.
A Fuel-cell tram could continue to operate as long as stored hydrogen fuel is available during a power outage. On site hydrogen production during periods of low electricity demand or such new methods like solar hydrogen generation with pumped in waste water and other equipment powered by solar panels this system could be *off-grid*. In contract, the heaviest electrical usage in an OCS system is during peak daytime hours. Considering geographical coordinates of Brescia, there is solar power production potential of 50.17 kwh/m² a year. With today's solar technology in market and available area of almost 30,000m² in depot, approximate 1.5 GWh production of electricity is feasible, which could run on site green hydrogen production with CO_2 emissions of zero kg for each kg of H_2.

Powertrain Technology

By early 2018, the Alstom Coradia iLint fuel cell entered service on a 100 km route in Germany. This train has a maximum velocity of 140 km/h, range of between 800 and 1000 km, a capacity of 150 sitting or 300 standing passengers.
In Brescia situation, an urban light fuel-cell tram would meet these requirements. On line T1 (Pendolina- Tangenziale Fiera) with length of 11.4 km, 23 stops, 143 trips per day on both directions in the most critical scenario, 5 units are running with expected range of 350 km a day each and max speed of 50 km/h.
The consumed energy for a single run is 71 kWh (Fig.3) considering total trips and number of operating units, each tram should be capable to store at least 2350 kWh per day.

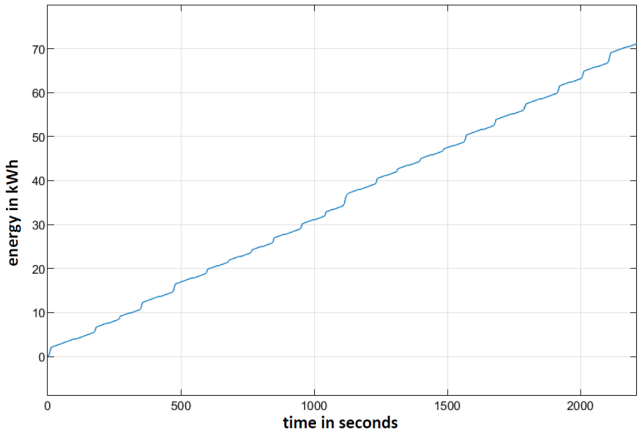


Figure 3: Energy consumption of a single run on studied line

The energy consumption on rail transport mainly depends on the rolling stock features, stop spacing and track profile. The local trains are heavier but they stop less frequently than the trams. The higher stop density includes more accelerations and braking than longer stop spacing (Fig.4). HyPM HD30 (33 kW) with efficiency map and technical data (Table 1) is the considered fuel cell component on the catenary-free tram. Battery and super-capacitors together are providing enough energy and power for traction and auxiliaries. On the other side, fuel cells and recovered braking energy (about 6.8 kWh) (Fig.5) are charging the battery with no sudden variation in output or a steady trend. By providing a balanced state of charge on our single branch battery, we need 11 fuel cells onboard. Hydrogen consumption of simulated tram is 0.3 kg/km and approximately 3.3 kg for one run.

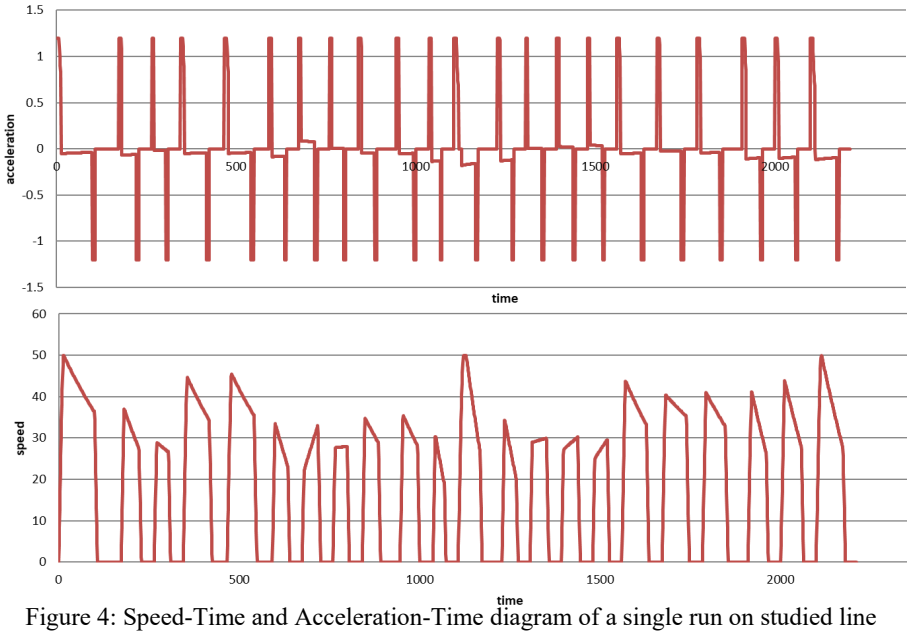


Figure 4: Speed-Time and Acceleration-Time diagram of a single run on studied line

Table 1: Parameters of HyPM HD30 (33 kW) Fuel Cell

Performances parameters	**Unit**	**Value**
Rated (Max continuous) Power	kW	30 (33)
Dimensions (L x W x H)	mm	950 x 1630 x 265
Mass	kg	≤ 70
Gravimetric Power Density	kW/kg	0.5
Operating Current	A_{dc}	0 to 500
Operating Voltage	V_{dc}	60 to 120
Peak Efficiency	$\%_{LHV}$	55
Stack Operating Pressure	kPa	< 120

Carrying Hydrogen enough for a day would require massive hydrogen tanks, which would make the tram too heavy and will affect the autonomy, hence a refuel approach of less than 15 minutes at terminus after each cycle is a promising solution. Therefore, 10 kg of hydrogen compressed to 350 bar can be stored preferably in dual tanks for longer charging life cycle and reliability with total weight of 80 kg and a volume of 120 l each. Tanks are on the top and providing mechanical safety valves, which let the tram to release H_2 into atmosphere in case of high temperatures, furthermore they are able to indicate any leaking in the system.

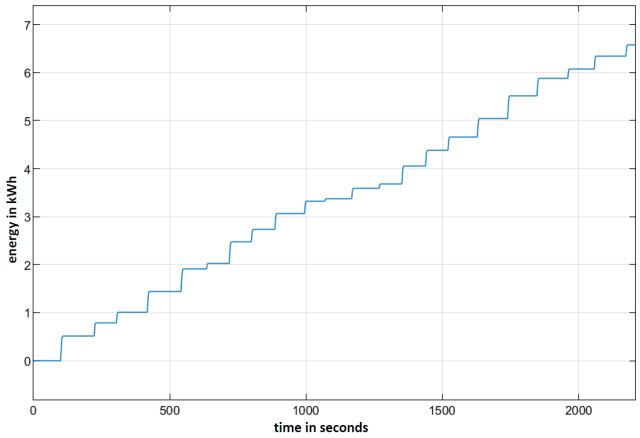


Figure 5: Energy recuperation in braking of a single run on studied line

In figure 6, you can see the suggested hybrid propulsion system, with a combination of super-capacitors, batteries (B) and fuel cell (Fc). The energy management system in hybrid powertrain enables the amount of needed power from each energy source to achieve high efficiencies, high performance and low consumption and take advantage of the components features. Batteries have high specific energy, and super capacitors (Sc) high specific power. Moreover, Sc provides energy for more charge/discharge cycles. In high demand of energy (Fig.2) Sc and B provide enough power and energy to supply traction motors. In low and very low energy demand phases recovered energy and surplus provided by the Fc with steady trend and no sudden output variation, charge B and Sc in cycles.

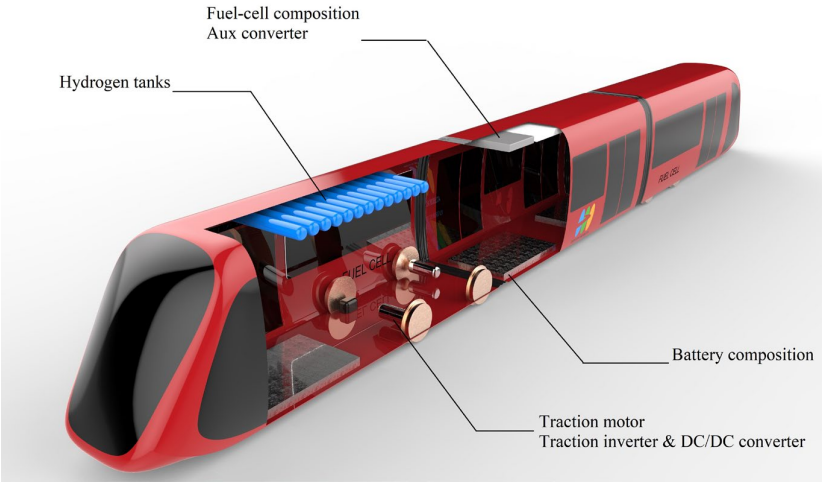


Figure 6: Schematic demonstration of conceptual simulated hybrid fuel-cell tram

Conclusions

The paper discusses new idea in a real-life hydrogen approach by taking into account Brescia tram project as a joint model for comparison, in light rail transport sector by highlighting benefits of using Fuel-cell tram instead of traditional catenary-based system. Generally, transition from pollutant electricity generation towards Hydrogen and Fuel-cell technologies would have direct and indirect benefits. Direct such as, zero local pollutant, reduced noise level, zero CO_2 emissions and increased used of renewables. Followed by indirect benefits, boosting research and innovation, attracting new businesses, creating new jobs, attracting skilled workforce, boosting local tourism, improving image as *green city* and increasing the quality of life. Hydrogen approach is perfectly in line with EU energy road map for 2050, reduction of emissions by at least 80% from their 1990s level, improving the security of energy supply and the flexibility of energy systems and infrastructures.

References

Kalenoja H. - *Energy consumption and environmental effects of passenger transport modes- A life cycle study on passenger transport modes* - Tampere University of Technology

Fragiacomo P., Piraino F. (2017) - *Energy performance of a Fuel Cell hybrid system for rail vehicle propulsion*

European Climate Foundation (2017) - *Fueling Italy's future*

Unione Petrolifera (2017) *annual report*, Italy

European Environment Agency - *Overview of electricity production and use in Europe*

Council of the EU (2016) - Council Decision (EU) 2016/1841 of 5 October 2016 on the conclusion of the Paris Agreement adopted under the United Nations Framework Convention on Climate Change - http://data.europa.eu/eli/dec/2016/1841/oj

EC. (2016a) - *The implementation of the 2011 White Paper on Transport "Roadmap to a Single European Transport Area – towards a competitive and resource-efficient transport system" five years after its publication: achievements and challenges* (Working document). SWD(2016) 226 final - https://ec.europa.eu/transport/sites/transport/files/themes/strategies/doc/2011_white_paper/swd%282016%29226.pdf

EC (2016b) - *Clean Energy for All Europeans (Communication).COM (2016) 860 final.* - http://eurlex.europa.eu/resource.html?uri=cellar:fa6ea15b-b7b0-11e6-9e3c-01aa75ed71a1.0001.02/DOC_1&format=PDF

EC (2016c) - *A European Strategy for Low-Emission Mobility (Communication). COM (2016) 501 final* - http://eurlex.europa.eu/resource.html?uri=cellar:e44d3c21-531e-11e6-89bd-01aa75ed71a1.0002.02/DOC_1&format=PDF

European Union's Horizon 2020 research and innovation programme, agreement No 730827 (2019) *OPEUS project,* Available at: *http://opeus-project.eu/*

UITP (2016) – *Statistics Brief: Local Public Transport in the European Union OPEUS_WP01 _D01.1 _ urban rail systems energy requirements in Europe _Final Page 40 / 40 Transport (UITP) -* http://www.mobiuitp.org/PermaLinkRecord.htm?archive=164530298271&file=UITP_Statistics_PT_in_EU_DEF_print+.pdf

OCED International Energy Agency (2018). *Global Energy & CO_2 Status Report*

Yoshida T, Kojima K. (2015). *Toyota MIRAI Fuel Cell Vehicle and Progress toward a Future Hydrogen Society -* https://www.electrochem.org/dl/interface/sum/sum15/sum15_p45_49.pdf

Yang X., Li X., Ning B., Tang T. (2015) - *A Survey on Energy-Efficient Train Operation for Urban Rail Transit -* https://www.researchgate.net/publication/282552115_A_Survey_on_Energy-Efficient_Train_Operation_for_Urban_Rail_Transit

Alstom (2018) - *Coradia iLint – the world's 1st hydrogen powered train -* https://www.alstom.com/coradia-ilint-worlds-1st-hydrogen-powered-train

Roland Berger (2018) *Fuel Cells and Hydrogen for Green Energy in European Cities and Regions*, Sederanger 1 80538 Munich Germany: Roland Berger GmbH.

International Railway Journal (2018) - *CRRC Tangshan trials new hydrogen-fuelled tram* - https://www.railjournal.com/passenger/light-rail/crrc-tangshan-trials-new-hydrogen-fuelled-tram/

Hypersolar.com (2018) - *HyperSolar solutions* - http://www.hypersolar.com/solutions.php

A Hybrid Forewarning Algorithm for Train Operation under Adverse Weather Conditions

Jun Zhang [a], Yuling Ye [a,1], Yunfei Zhou [a]
[a] College of Transportation Engineering, Tongji University,
The Key Laboratory of Road and Traffic Engineering, Ministry of Education
P.O. Box 201804, 4800 Cao'an Highway, Shanghai, China
[1] E-mail: yyling71@163.com, Phone: +86 131 6256 6602

Abstract
This paper presents a combined method of fuzzy theory and rough sets theory for the early warning of high-speed railway (HSR) under adverse weather conditions. Based on the monitoring data of meteorological indicators along the railway, a fuzzy c-means (FCM) clustering is first applied in order to figure out the fuzzy distribution of sample data and to fit the corresponding membership function of every indicator. According to the clustering results, every original sample is transformed into its cluster level as string data for the subsequent application of rough sets theory. Then a series of effective rough rules between conditional indicators and the decision indicator is extracted after attribute reduction by the Rosetta toolkit, where the decision indicator is represented by the train deceleration rate. Since the value of an indicator may correspond to several fuzzy levels, the multiple combinations of different conditional indicators will activate multiple rough rules. In order to forecast a clear value of the decision indicator, a centroid-based Max-Min compound arithmetic is applied to clarify relevant rules and determine the warning level. Using the designed algorithm, a case analysis of an HSR line section is conducted to verify the feasibility of the combined method, all meteorological data and operation records are provided by the Shanghai Railway Bureau in China. The results prove that the hybrid algorithm can be applied in the real-time forewarning of HSR train operation, with a global accuracy over 86%.

Keywords
High-speed railway, Forewarning algorithm, Adverse weather, Fuzzy theory, Rough sets

1 Introduction

High-speed railway (HSR) has recently become an important share of the transport market in China, with advantages of comfort, convenience and punctuality. Currently, in view of the high service frequency and high management demand, developing the forewarning system has become an essential way to proactively secure the train operation and guarantee the transport efficiency. The early warning of HSR operation has been extensively explored in the literature. Risk factors of HSR accidents usually come from railway infrastructure, train equipment, operation management and external environmental conditions (Goverde and Meng, 2011; Li et al., 2018). The infrastructure failure, train equipment malfunction and operation error are usually unexpected and are uncertain with emergency responses (Fan et al., 2015; Ouyang et al., 2010), while the escalation of environmental conditions can be predictable under the real-time monitoring. Meanwhile, adverse weather conditions such as wind gust and heavy rainfall have strong effects on

high-speed train operation according to the aerodynamic analysis (Baker, 2010; Shao et al., 2011; Du and Ni, 2016), and relevant possibility of derailment is higher. Xia et al. (2013) also found that the train arrival punctuality and cancellation rate become worse under bad weather conditions. Therefore, it is necessary to establish an effective forewarning method to guide the HSR operation under bad weather.

Forewarning methods such as the decision tree algorithm, Bayesian training network and support vector machine (SVM) algorithm are frequently used in training datasets and predicting the impacts of occurring event, based on the data of high nonlinearity and dynamicity (Castillo et al., 2016; Jiang et al., 2017; Annelies et al., 2018; Yan et al., 2018). In addition, An et al. (2016) applied fuzzy analytical hierarchy process (AHP) approaches in the railway risk decision making process. Hu et al. (2018) constructed a rough measurement model to describe the safety of high-speed train operation. However, when faced with the forewarning of train operation under adverse weather conditions, some algorithms will have limitations. To our best knowledge, the decision tree algorithm is unable to distinguish the noisy datasets from valid datasets (Oates and Jensen, 1997). The SVM is a learning method for small sample data (Yang et al., 2018), and it is hard to deal with complex multi-dimensional data of weather conditions. Meanwhile, the Bayesian network model requires that the data obey a Gaussian distribution (Xie et al., 2017), which doesn't meet the abrupt changes of meteorological indicators such as rain intensity and wind speed, meanwhile prediction failure occurs when a real-time data is outside of the original training set.

On the basis of above mentioned limitations, this study presents a hybrid algorithm of fuzzy theory and rough sets theory, composed of fuzzy c-means (FCM) clustering, fuzzy distribution fitting, attribute reduction, rough rules extraction and Max-Min compound arithmetic. This combined method was designed with advantages of mitigating the influence of noisy data for efficient forecasting, due to the correlation between indicators. This algorithm has been applied to an HSR section (shown in Figure 1) of Shanghai Railway Bureau in China, based on the historical monitoring data of meteorological conditions and operation records.

Figure 1: The railroad section of Beijing-Shanghai HSR

The remainder of this paper is structured as follows. Section 2 first outlines the algorithm framework; Section 3 describes the details of models in fuzzy theory and rough sets theory. Following this, Section 4 performs a case analysis using real monitoring data under adverse weather, where the results are fully discussed. Eventually, Section 5 reaches some conclusions and makes suggestions for future work and research aspects.

2 Algorithm design

In China, current forewarning system for train operation under adverse weather is designed according to the *Regulations on Railway Technical Management* and the *Detailed Rules on Organization of Train Operation*, where speed limits have been regulated for train operation under windy weather and rainy weather seperately, as shown in Table 1. Since the speed limits for wind speed are inconsistent with the hourly rainfall, we found difficulties in train dispatching when faced with complex weather conditions, especially the wind-driven rain.

Table 1: Speed-limit standards under windy weather and rainy weather

Wind Speed (m/s)	Top Speed (km/h)	Rainfall (mm/h)	Top Speed (km/h)
[0, 20]	300	[0, 30]	300
(20, 25]	200	(30, 45]	250
(25, 30]	120	(45, 60]	120
> 30	Stop	> 60	45

Therefore, it is essential to propose a forewarning algorithm to support the decision making of train operation under complex weather conditions. For this purpose, a hybrid algorithm of fuzzy theory and rough sets is then developed, shown in Figure 2.

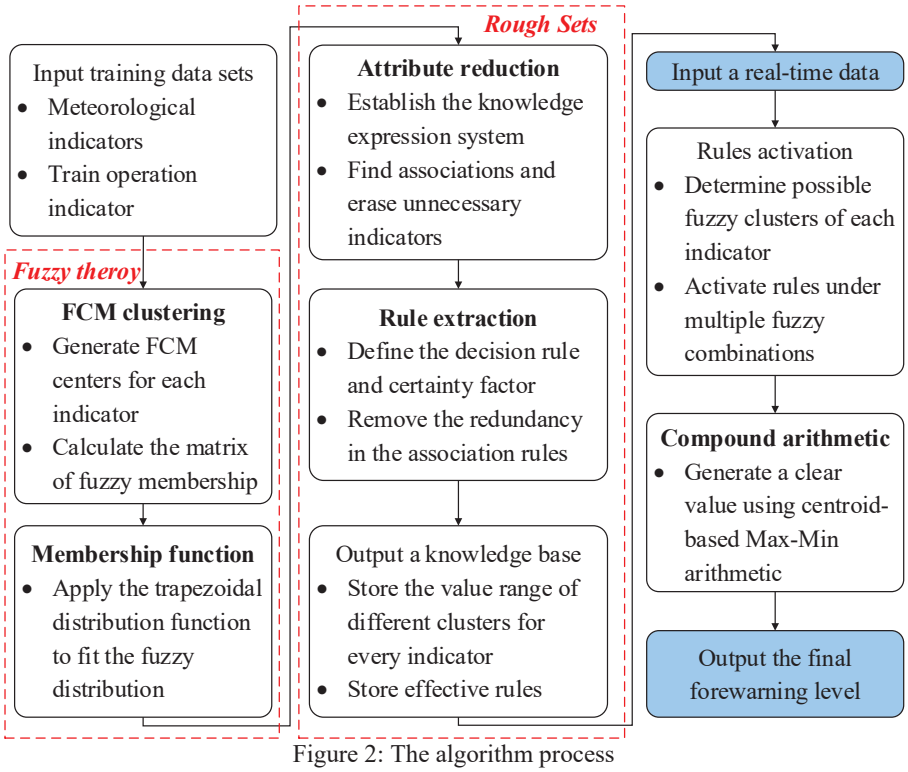


Figure 2: The algorithm process

Five major steps in the algorithm procedure are listed below.

Setp1. FCM clustering. For each indicator (including conditional indicators and decision indicator), several cluster centers and fuzzy membership grades are generated from corresponding original data.

Step2. Membership function fitting. The membership function of each conditional indicator is fitted to the fuzzy membership grades. To simplify the calculation, the trapezoidal distribution function is applied to each cluster level of indicators.

Step3. Attribute reduction. The FCM is an independent fuzzy classification of each indicator, while the attribute reduction can efficiently find out associations and erase unnecessary indicators according to the rough sets theory.

Step4. Rule extraction. This step is a further analysis to remove the redundancy in the association rules, and to reduce the impact of noisy data. Upon introduced certainty factor, effective rules between conditional indicators and the decision indicator are figured out.

Step5. Compound arithmetic. Given a real-time monitoring data of conditional indicators, an intersection set is naturally output from the multiple combinations of fuzzy levels. The centroid-based Max-min compound arithmetic is applied in order to defuzzy the calculation and to get a clear value for judging the forewarning level.

3 The combination of fuzzy theory and rough sets

3.1 Fuzzy theory

Fuzzy c-means clustering

As compared to traditional clustering models like k-means method and density-based methods, the FCM can better accommodate the rough sets theory in discretizing the original datasets and in performing a comprehensive arithmetic based on the value of fuzzy membership. For each indicator, the FCM clustering is performed based on the original data set, which is a column vector. Assuming that n is the number of samples in the original data set, and $P = \{p_1, p_2, \cdots, p_n\}$ is the values set, the problem of FCM clustering can be formulated as:

$$T = \min\{\sum_{j=1}^{K}\sum_{i=1}^{n} v_{ij}^{\alpha} \left\| p_i - m_j \right\|^2\}, \quad \alpha \geq 1 \tag{1}$$

$$s.t. \quad \begin{cases} \sum_{j=1}^{K} v_{ij} = 1 \\ 0 \leq v_{ij} \leq 1 \end{cases}, \quad i = 1, 2, \cdots, n \tag{2}$$

where v_{ij} denotes the probability when the i^{th} sample belongs to the j^{th} clustering center, namely the fuzzy membership, m_j represents the value of the j^{th} clustering center, K is the number of clustering centers, and α is the fuzzy parameter with a positive relation to the fuzziness (Gong et al., 2005). It is important to note that initial centers are random selected from P, and the value of K should consider practical significance.

Fuzzy membership function

According to the coverage of clustering centers, the distribution patterns of membership function include left type, right type and center type, where the triangular function and trapezoidal function are included in the center type distribution, shown in Figure 3. Since the left type and right type distribution are two special cases of the trapezoidal distribution (Botzheim et al., 2001), the center type is selected to fit the fuzzy distribution.

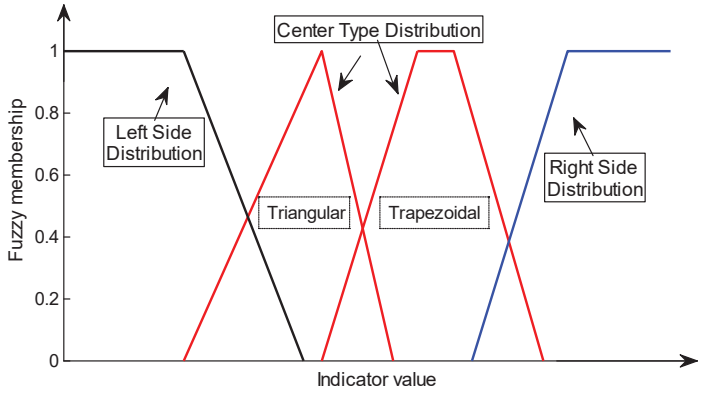


Figure 3: Three typical distribution patterns of membership function

3.2 Rough sets theory

Attribute reduction

The rough sets theory was first proposed by Pawlak (1982), which can be applied in fields of machine learning, knowledge acquisition, decision analysis and process control (Pawlak, 2002). Before attribute reduction, a knowledge expression system (KES) of the rough sets should be established, which is defined as:

$$S = (U, A, V, f) \tag{3}$$

In this equation, U is the set of samples defined as $U = \{x_1, x_2, \cdots, x_n\}$, where x_i is a row vector representing the i^{th} individual sample. A is the set of attributes including conditional indicators (denoted by C) and the decision indicator (denoted by D). V is the set of value ranges of all attribute indicators. f represents the information function. It is noted that every indicator's value of x_i is uniquely determined in V.

Based on the discernible matrix from original decision table, attributes should get reducted to erase the linearity between conditional indicators as much as possible. The decision table is defined by $T = (U, A, C, D)$, and the corresponding discernible matrix is denoted as an $n \times n$ matrix $M(T)$. Any element in $M(T)$ is determined by:

$$c_{ij} = \begin{cases} \left\{ a \middle| a \in C \ \& f(a, x_i) \neq f(a, x_j) \right\}, & (x_i, x_j) \notin ind(D) \\ \phi, & (x_i, x_j) \in ind(D) \end{cases} \tag{4}$$

where c_{ij} represents the set of attributes which can distinguish sample x_i from sample x_j, and ind (D) is the indiscernible set of samples with the same attributes values of D. Obviously, c_{ij} is an empty set when samples x_i and x_j belong to the same indiscernible set.

Rough rules extraction

Rough rules extraction is in a critical position between the attribute reduction and compound arithmetic, aiming to output decision rules from conditional indicators to the decision indicator (Maji and Garai, 2013). For example, if a decision table contains 2 conditional indicators and 1 decision indicator, assuming every indicator has 3 clustering levels, then there are 27 decision rules in an exhaustive way, while the number of rules will get significantly reduced by the rules extraction considering the certainty of each rule.

During rules extraction, the decision rule is defined as:

$$r_{ij}: des(C_i) \rightarrow des(D_j), C_i \bigcap D_j \neq \phi \tag{5}$$

$$des(C_i) = \left\{(a, V_a) \middle| f(x,a) = v_a, \forall a \in C\right\} \tag{6}$$

$$des(D_j) = \left\{(a, V_a) \middle| f(x,a) = v_a, \forall a \in D\right\} \tag{7}$$

Also, the corresponding certainty factor of rule r_{ij} is therefore determined by:

$$\mu_{ij} = \frac{card\left\{D_j \bigcap C_i\right\}}{card\left\{C_i\right\}}, \quad C_i \bigcap D_j \neq \phi \tag{8}$$

where μ_{ij} denotes the certainty factor, ranging from 0 to 1. Rule r_{ij} is a certain decision rule when μ_{ij} is 1, otherwise it is uncertain. Decision rules with high certainty are output into a knowledge base to improve the calculation efficiency of subsequent work.

3.3 Compound arithmetic

Given a real time monitoring data, values of conditional indicators correspond to different levels and will activate different rough rules in the knowledge base. The compound arithmetic is a centroid-based Max-Min arithmetic (Wang, 2009) used to forecast a clear value of decision indicator under different rough rules activated by the same sample data of conditional indicators. The basic function centroid-based Max-Min arithmetic is:

$$x^* = \frac{\sum_{i=1}^{P} \int_{U_x^i} x \cdot v_D(x) \mathrm{d}x}{\sum_{i=1}^{P} \int_{U_x^i} v_D(x) \mathrm{d}x} \tag{9}$$

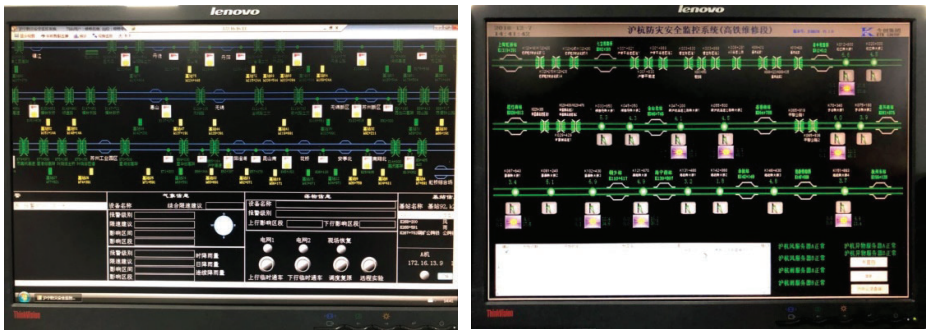
$$\left\{U_x^i\right\} = \max\{\min[v_{C1}(x), v_{C2}(x), \cdots, v_{Ck}(x), v_D(x)]\} \tag{10}$$

where x^* denotes the clear value of the decision indicator, $v_D(x)$ is the fuzzy distribution function of the decision indicator, $v_{Ci}(x)$ is the fuzzy distribution function of the i^{th} conditional indicator, U_x^i represents the domain set activated by the i^{th} rough rule, and P is the number of activated rough rules under current sample data.

4 The Case Analysis

4.1 Data collection

The original monitoring data of weather conditions and train operation under adverse weather conditions of an HSR section (see Figure 1) are provided by the Shanghai Railway Bureau in China. As shown in Figure 4, rainfall indicators and wind indicators are two key targets in current safety monitoring system of HSR. With the help of this system, continuous monitoring data of meteorological indicators are easily associated with train operation records under adverse weather conditions. The date of collected data is 10th June, 2017, a day during stormy weather.



(a) Monitoring rainfall indicators (b) Monitoring wind indicators

Figure 4: The disaster prevention and safety monitoring system of HSR

The meteorological indicators function as the conditional indicators, including wind speed (WS), wind direction (WD), rainfall intensity (RI), hourly rainfall (HR), daily rainfall (DR) and continuous rainfall (CR). The actual deceleration rate (AD) functions as the decision indicator to determine the level of early warning. The training data of 297 valid samples under bad weather have been studied, as shown in Table 2.

Table 2: Original sample data of HSR operation under adverse weather conditions

A	Conditional indicators						Decision indicator
	C_1	C_2	C_3	C_4	C_5	C_6	D_1
n	WS (m/s)	WD [0,180°]	RI (mm)	HR (mm/h)	DR (mm/day)	CR (mm)	AD (m/s^2)
1	12.1	46.4	0.0	0.0	0.3	0.3	0.11
2	12.8	39.4	0.0	0.0	0.3	0.3	0.14
3	13.6	40.7	0.0	0.0	0.3	0.3	0.17
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
175	15.3	88.1	25.5	53.0	44.1	150.9	0.97
176	15.6	88.0	25.4	52.4	44.3	151.2	1.17
177	15.8	87.0	22.5	51.9	44.5	151.4	0.94
178	16.0	93.5	14.5	51.4	44.8	151.6	1.03
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
295	9.7	60.6	0.0	1.0	27.8	155.6	0.08
296	9.6	65.4	0.0	1.0	27.8	155.6	0.00
297	9.4	60.4	0.0	0.9	27.8	155.6	0.03

4.2 Algorithm application

Based on the sample data, FCM is first performed to obtain the fuzzy membership distribution of each indicator, shown in Figure 5. Using the indicator of wind speed as an example, data of wind speed have been classified into level I, II, III and IV, and the corresponding function curves are plotted by different colors. Similarly, indicators of wind direction, rainfall intensity, hourly rainfall, daily rainfall and continuous rainfall are classified into 3, 4, 5, 3 and 3 levels respectively, where the number of levels are carefully determined to satisfy relevant HSR technical regulations.

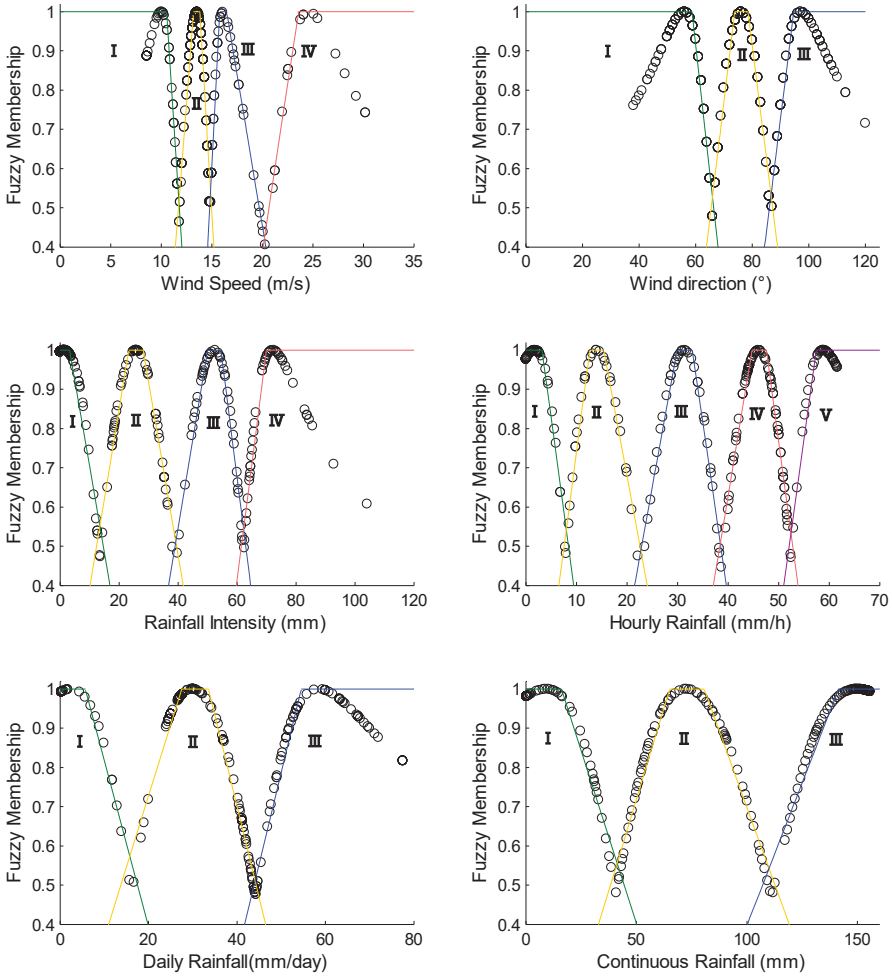


Figure 5: The fuzzy membership distribution of 6 conditional indicators

Based on the results of fuzzy clustering of each indicator, original numeric data of attributes are converted into string type data for the analysis in Rosetta toolkit. Through attribute reduction, the indicator C_6 (CR) is removed from the original data set, and 66 rough rules have been generated. Given a sample data set {21.2, 67.9, 47.5, 52.3, 86.7}, six rough rules are activated, shown in Table 3.

Table 3: Activated rough rules

Activated rules (Inputs → Output)	Fuzzy Membership Grade					Min value
	WS	WD	RI	HR	DR	
III, I, II, IV, III → V	0.271	0.401	0.148	0.548	0.718	0.148
III, II, II, IV, III → I	0.271	0.645	0.148	0.548	0.718	0.148
III, II, III, V, III → V	0.271	0.645	0.883	0.519	0.718	0.271
IV, I, II, V, III → IV	0.591	0.401	0.148	0.519	0.718	0.148
IV, II, II, IV, III → V	0.591	0.645	0.148	0.548	0.718	0.148
IV, II, III, V, III → V	0.591	0.645	0.883	0.519	0.718	0.519

Then the max-min compound arithmetic is applied based on the 6 activated rules. In combination with the fuzzy membership function of deceleration rate, the max-min area is designated by the shaded area, as shown in Figure 6. The clear value of decision indicator DR is 0.994 m/s^2 according to the centroid arithmetic in Equations (9) and (10); the corresponding forewarning level is IV.

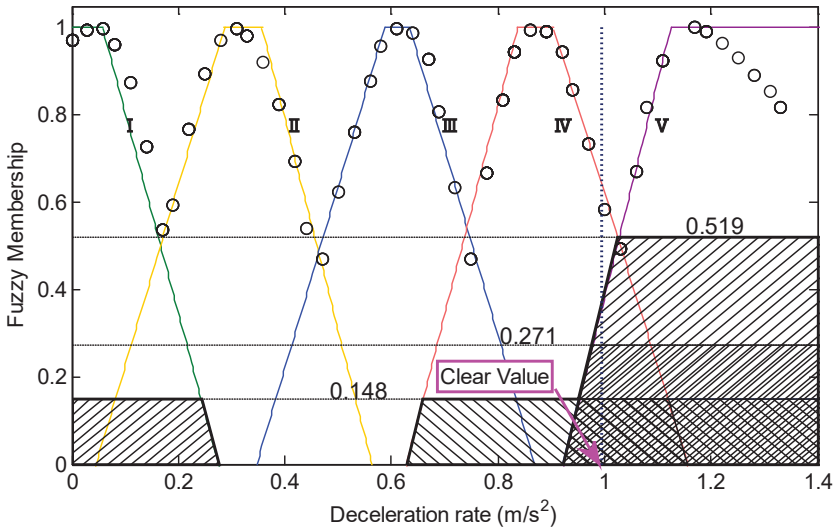


Figure 6: The max-min area of decision indicator DR under 6 activated rough rules

4.3 Discussion

Before evaluating the accuracy of this hybrid algorithm, samples are divided into 2 groups (Group 1 and Group 2) according to the actual deceleration rate. Under current early warning system, the service braking curve is frequently applied to HSR train operation. Based on the braking curves of CRH2 series train (Shangguan et al., 2011) at an initial velocity 300 km/h (see Figure 7), the average deceleration rate under service braking is approximately 0.83 m/s^2. Based on this, Group 1 contains sample data with actual deceleration rates below this average value, and Group 2 contains sample data with deceleration rates above the average value.

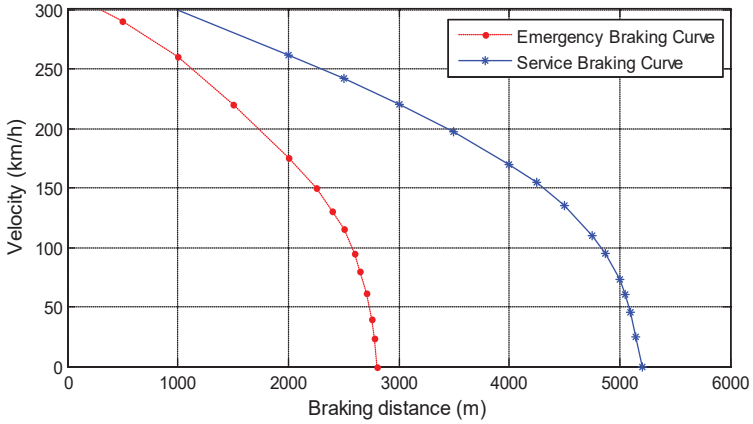


Figure 7: Braking curves of CRH2 series train

The accuracy is defined as the proportion of samples whose forecasting levels are consistent with practical levels. To acquire the accuracy rate, the hybrid forewarning algorithm is applied to all 297 samples, and the results are shown in Table 4.

Table 4: Forewarning accuracy of samples in Group 1 and Group 2

Group	Number of samples	Forewarning accuracy	Global accuracy
Group 1 (AD<0.83)	174	83.33%	86.53%
Group 2 (AD≥0.83)	123	91.06%	

As indicated in Table 4, we have 174 samples in Group 1 with a forewarning accuracy of 83.33%, and 123 samples in Group 2 with a forewarning accuracy 91.06%. Meanwhile, it is obvious that Group 2 has a higher forewarning accuracy as compared to Group1. As we know, the actual deceleration rate is correlated with weather conditions, and the deceleration rate increases with weather conditions getting worse. Since the actual deceleration rates of Group 2 are bigger than Group 1, the weather conditions of samples in Group 2 are worse than Group 1.

Based on the above analysis, the hybrid model seems better suited to data under extremely adverse weather conditions. The phenomenon may be explained the indicator level is sensitive under extremely adverse weather, which is easy to identify by the hybrid algorithm. In general, the global accuracy of all 297 samples is approximately 86.53%.

Meanwhile, the suggested top speed can be obtained by combining the threshold intervals of forewarning levels with the characteristics of the service braking curve, shown in Table 5. In the case study, the final forewarning level is IV, meaning that the corresponding top speed for train operation is suggested to be 200 km/h.

Table 5: Suggested top speed under each forewarning level of DR

Level	I	II	III	IV	V
Threshold Interval (m/s^2)	[0,0.17)	[0.17,0.46)	[0.46,0.74)	[0.74,1.03)	[1.03, 1.4]
Top Speed (km/h)	300	300	250	200	100

The algorithm is logically divided into parts of offline computation and real-time computation. Basic decision rules get extracted or updated by training and analyzing historical datasets in the offline computation, while the compound arithmetic is operated efficiently in the real-time computation with a little computational load.

5 Conclusions

In this paper, authors contribute to the forewarning method for train operation under adverse weather conditions. It is a combined algorithm of fuzzy clustering and rough sets, where monitoring data of meteorological indicators like wind speed and rain intensity are used for the data training and analysis. Main novelties introduced by this paper are the adoption of combining the fuzzy theory with rough sets theory, characterized by: (a) a fuzzy distribution of original conditional indicators and the decision indicator; (b) a set of reducted indicators after attributes reduction; (c) effective rough rules represented by the level of conditional indicators and the decision indicator; (d) a clear value output by the compound arithmetic under activated rough rules.

The application of this early warning method has indicated the feasibility of decision rules. The global forewarning accuracy is approximately 86.53%, where the accuracy is higher for 123 samples under extremely adverse weather conditions. Nonetheless, due to the difficulty in data collection considering some confidentiality and privacy, the number of valid samples is below our expectation. Given more sample data, the conditional attributes will get fully reducted, and the rough rules will describe the relationship between conditional attributes and decision attributes more precisely, thus the algorithm can guarantee the accuracy of the forewarning level.

Further developments will be focused on the expansion of conditional indicators such as atmospheric pressure and ambient temperature, and additional efforts has to be spent in the modification of reduction algorithm to guarantee the nonlinearity among conditional indicators. Nevertheless, the authors believe that the proposed method can be applied in the revision of corresponding rules and regulations, and the hybrid algorithm can provide basic support for HSR train operation under complex adverse weather conditions.

Acknowledgements

This research work was partially supported by the National Key R&D Program of China under Grant No. 2018YFB1201403 and partially by the CRRC Corporation Limited under Grant No.2016YFB1200400. We deeply thank Yongjian Zhang at Shanghai Railway Bureau for his assistance in data collection.

References

An, M., Qin. Y., Jia, M. L., Chen Y., 2016. “Aggregation of group fuzzy risk information in the railway risk decision making process”. *Safety Science*, vol. 82, pp. 18-28. https://doi.org/10.1016/j.ssci.2015.08.011

Annelies, B., Luciana, D. N., Steven S., Justine M., Saul B., 2018. “An impact-oriented Early Warning and Bayesian-based Decision Support System for flood risks in Zeebrugge harbor”, *Coastal Engineering*, vol. 134, pp. 191-202. https://doi.org/10.1016/j.coastaleng.2017.10.006

Baker, C. J., 2010. “The simulation of unsteady aerodynamic cross wind forces on trains”, Journal of Wind Engineering & Industrial Aerodynamics, vol. 98, pp. 88-99. https://doi.org/10.1016/j.jweia.2009.09.006

Botzheim, J., Hámori, B., Kóczy L. T., 2001. “Extracting Trapezoidal Membership Functions of a Fuzzy Rule System by Bacterial Algorithm”, In: *International Conference on Computational Intelligence*. Dortmund, Germany. https://doi.org/10.1007/3-540-45493-4_25

Castillo, E., Andrade Z. G., Calviño, A., 2016. “Bayesian networks-based probabilistic safety analysis for railway lines”, *Computer-Aided Civil and Infrastructure Engineering*. vol. 31, pp. 681-700. https://doi.org/10.1111/mice.12195

Du, M. L., Ni S. L., 2016. “Influence of Aerodynamic Characteristics of a High-Speed Train under Rain and Lower Atmospheric Boundary Layer Crosswind Conditions”, *Journal of Dalian Jiaotong University*, vol. 37, pp. 56-61.

Fan, Y., Li, Z., Pei, J., Li, H., Sun, J., 2015. “Applying systems thinking approach to accident analysis in china: case study of ‘7.23’ yong-tai-wen high-speed train accident”, *Safety Science*, vol. 76, pp. 190-201. https://doi.org/10.1016/j.ssci.2015.02.017

Gong, G. Y., Gao, X. B., Wu, Z. D., 2005. “An optimal choice method of parameter m in FCM clustering algorithm”, *Fuzzy Systems and Mathematics*, vol. 19, pp. 143-148.

Goverde, R. M. P., Meng, L., 2011. “Advanced monitoring and management information of railway operations”, *Journal of Rail Transport Planning & Management*, vol. 1, pp. 69-79. https://doi.org/10.1016/j.jrtpm.2012.05.001

Hu, Q. Z., Tan, M. J., Lu, H. P., Zhu, Y., 2018. “A rough set-based measurement model study on high-speed railway safety operation”, *PLOS ONE*, vol. 13, pp. 1-14. https://doi.org/10.1371/journal.pone.0197918

Jiang, D. Z., Gong, J., Garg, A., 2017. “Design of early warning model based on time series data for production safety”, *Measurement*, vol. 101, pp. 62-71. https://doi.org/10.1016/j.measurement.2017.01.033

Li, S. H., Cai, B. G., Liu, J., Wang, J., 2018. “Collision risk analysis based train collision early warning strategy”, *Accid Anal Prev*, vol. 112, pp. 94-104. https://doi.org/10.1016/j.aap.2017.11.039

Maji, P., Garai, P., 2013. “On fuzzy-rough attribute selection: criteria of max-dependency, max-relevance, min-redundancy, and max-significance”, *Applied Soft Computing*, vol. 13, pp. 3968-3980. https://doi.org/10.1016/j.asoc.2012.09.006

Oates, T., Jensen, D., 1997. “The effects of training set on decision tree”, In: *Procedings of the 14th International Conference on Machine Learning*, Morgan, Kaufman.

Ouyang, M., Hong, L., Yu, M. H., Fei, Q., 2010. “Stamp-based analysis on the railway accident and accident spreading: taking the China-Jiaoji railway accident for example”, *Safety Science, 48*(5), vol. 48, pp. 544-555. https://doi.org/10.1016/j.ssci.2010.01.002

Pawlak, Z., 1982. “Rough sets”, *International Journal of Computer & Information Sciences,* vol. 11, pp. 341-356.

Pawlak, Z., 2002. “Rough sets and intelligent data analysis”, *Information Science*, vol. 147, pp. 1-12. https://doi.org/10.1016/S0020-0255(02)00197-4

Shangguan, W., Cai, B. G., Wang, J. J., Wang, J., Wang, L., 2011. “Braking mode cure arithmetic of high-speed train above 250 $km \cdot h^{-1}$”, *Journal of Traffic and Transportation Engineering*, vol. 11, pp. 41-54.

Shao, X. M., Wan, J., Chen, D. W., Xiong, H. B., 2011. “Aerodynamic modeling and stability analysis of a high-speed train under strong rain and crosswind conditions”, *Journal of Zhejiang University-Science A(Applied Physics & Engineering)*, vol. 12, pp. 964-970. https://doi.org/10.1631/jzus.A11GT001

Wang, Y. M., 2009. “Centroid defuzzification and the maximizing set and minimizing set ranking based on alpha level sets”, *Computers & Industrial Engineering*, vol. 57, pp. 228-236. https://doi.org/10.1016/j.cie.2008.11.014

Xia, Y., Van Ommeren, J. N., Rietveld, P., Verhagen, W., 2013. “Railway infrastructure disturbances and train operator performance: the role of weather”, *Transportation Research Part D: Transport and Environment,* vol. 18, pp. 97-102. https://doi.org/10.1016/j.trd.2012.09.008

Xie, J. M., Choi, Y. K., 2017. “Hybrid traffic prediction scheme for intelligent transportation systems based on historical and real-time data”, *International Journal of Distributed Sensor Networks*, vol. 13, pp. 1-12. https://doi.org/10.1177/1550147717745009

Yan, R. H., Wu, C., Wang, Y., 2018. “Exploration and evaluation of individual difference to driving fatigue for high-speed railway: a parametric SVM model based on multidimensional visual cue”, *Iet Intelligent Transport Systems*, vol. 12, pp. 504-512. http://doi.org/10.1049/iet-its.2017.0289

Yang, J., Liu, Z., Jiang, G., Zhu, L., 2018. “Two-phase model of multistep forecasting of traffic state reliability”, *Discrete Dynamics in Nature & Society*, pp. 1-12. https://doi.org/10.1155/2018/7650928

A Heuristic Algorithm for Re-Optimization of Train Platforming in Case of Train Delays

Yongxiang Zhang [a], Qingwei Zhong [a], Chao Wen [a,b,1], Wenxin Li [a], Qiyuan Peng [a]
[a] School of Transportation and Logistics, Southwest Jiaotong University
Chengdu, 610031, China
[b] High-speed Railway Research Center, University of Waterloo
Waterloo, N2L3G1, Canada
[1] E-mail: c9wen@uwaterloo.ca, Phone: 1-2269788096

Abstract
Train platforming is critical for ensuring safety and efficiency of train operations within the stations, especially when train delays occur. This paper studies the problem of re-optimization of train platforming, where the train station is modeled using discretization of the platform track time-space resources. To solve the re-optimization problem, we propose a binary integer programming model which minimizes the weighted sum of total train delays as well as platform track utilization costs, subject to constraints defined by operational requirements. Moreover, we design an efficient heuristic algorithm to solve the model with a good precision. A real-world case is taken as an example to show the effectiveness of the proposed model and algorithm. The results show that the model established in this paper can describe re-optimization of train platforming accurately and can be solved quickly by the proposed heuristic algorithm. In addition, the model and algorithm developed in this paper can provide an effective computer-aided decision-making tool for the train dispatchers in case of train delays.

Keywords
Train platforming; Train delay; Re-optimization; Discretization; Heuristic algorithm

1 Introduction

Train operations of the trains at stations, including arrival, dwell, and departure or passing through, are usually optimized by solving the train platforming problem (Lusby et al., 2011). In general, due to the hierarchal planning process of the railway, train timetable is specified first, and then train platforming problem is optimized with given train arrival and departure times (Lusby et al., 2011). Train platforming is a classic NP-hard combinational optimization problem (Kroon et al., 1997), and a lot of work has been done to generate high-quality train operation plan within stations. Zwaneveld et al. (1996) defined the train route as a collection of station equipment traveled by a train from inbound to the outbound of the station, and they built a mixed integer linear programming (MILP) model based on node packing problem to maximize the number of trains routed through the station. Chakroborty and Vikram (2008) developed a MILP model for optimally allocating trains to the platform tracks, where the accurate train arrival times can only be available shortly before the train

arrives at the station such that trains could be reassigned to different platforms. Besides, the headway between two trains was also considered while delaying the train arrival and departure times. Caprara et al. (2011) assumed that the arrival and departure times of trains could be slightly flexible, and they presented a quadratic binary integer programming model to solve the train platforming problem. Later, Lusby et al. (2013) built MILP models based on the set packing model to maximize total revenue and minimize the total costs of all trains. Sels et al. (2014) developed a MILP model to solve the train platforming problem from strategic and tactical levels.

Trains may suffer from all kinds of disturbances and disruptions, such as bad weather, equipment failure, management factors, etc. When train delays occur, the scheduled train timetable within stations needs to be re-optimized in real time. However, very few researchers have focused on the problem of re-optimization of train platforming in case of train delays. In this study, we aim to re-optimize the train platforming in case of train delays and generate a new train operation plan within the station in real time. Our solution is to develop a Mixed-Integer Linear Programming (MILP) model, where the train station is modeled using the discretized platform track time-space resources, and to propose an efficient heuristic algorithm. The goal of the proposed model and algorithm is to simultaneously minimize the deviation from the train timetable and the total train operating costs, realizing the coherence between train operations and the station management.

The contributions of this study include the following three aspects. First, the train arrival and departure times and the train platform assignment are optimized simultaneously in order that the negative influence of train delays can be minimized. Second, the novel modeling method based on the discretized platform track time-space resources can describe the train conflicts accurately, where the complex binary train sequencing variables in the big-M modeling framework can be avoided (Chakroborty and Vikram, 2008). Third, an efficient heuristic algorithm is designed to quickly obtain the near-optimal solutions for the real-time re-optimization of train platforming.

2 Analysis of platform track time-space resources

In the planned horizon $\mathscr{T}$, we handle the time resources as small time units $\Delta\tau$. The number of time units is equal to $|T|=[\mathscr{T}/\Delta\tau]$ in the entire planned horizon. In addition, the number of platform tracks in a station is denoted by $|I|$, i.e., the maximum spatial capacity. Hence, the platform track time-space resources of a station can be represented by a two-dimensional matrix X,

$$X=\left[x_{i,t}\right]=\begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,|T|} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,|T|} \\ \cdots & \cdots & \cdots & \cdots \\ x_{|I|,1} & x_{|I|,2} & \cdots & x_{|I|,|T|} \end{bmatrix},$$

where i and t are the indexes of the platform track and the time unit, respectively, and a binary variable $x_{i,t}$ in matrix X denotes the occupation and vacancy state of the platform track time-space resource $(i,\ t)$, where

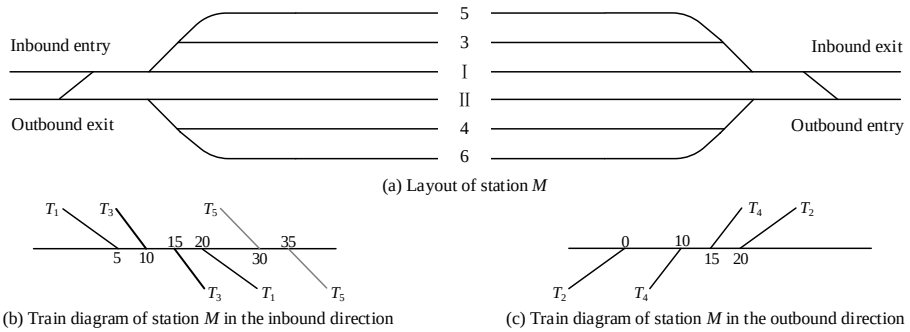
$$x_{i,t}=\begin{cases} 1, & \text{platform track time-space resource } (i,\ t) \text{ is used} \\ 0, & \text{othewise} \end{cases}$$

Fig. 1 provides an illustrative example of the modeling method of platform track time-space resources. Suppose that 5 trains successively arrive at or depart from station *M* which has 6 platform tracks within the planned horizon of 60 min. The detailed schedules of train operations in both directions are given in Fig. 1 (*b*) and (*c*), and the time unit $\Delta\tau$ is set to 5 min. Platform track time-space resources and the corresponding matrix X of a feasible usage plan are described in Fig. 2 (*d*) and (*e*), respectively. The application requirements of the time-space resources modeling method for the re-optimization of train platforming problem can be formulated as follows:

(1) Inseparability. A train must occupy only one platform track and cannot occupy more than one platform track simultaneously.

(2) Exclusivity. One platform track can only store one train at any time unit.

(3) Continuity. A train operation on one platform track with the duration equal to Δt time units cannot be interrupted. If one train starts to use track i at time *t*, it continues to occupy the platform track i until $t+\Delta t$, i.e., $x_{i,t} = x_{i,t+1} = x_{i,t+2} = ... = x_{i,t+\Delta t} = 1$.

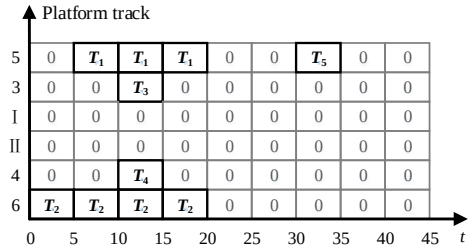


(a) Layout of station *M*

(b) Train diagram of station *M* in the inbound direction

(c) Train diagram of station *M* in the outbound direction

Figure 1: Layout and train schedules of station *M*



(a) Usage of platform track time-space resources

Platform track

5	0	1	1	1	0	0	1	0	0
3	0	0	1	0	0	0	0	0	0
I	0	0	0	0	0	0	0	0	0
II	0	0	0	0	0	0	0	0	0
4	0	0	1	0	0	0	0	0	0
6	1	1	1	1	0	0	0	0	0
t	5	10	15	20	25	30	35	40	45

(b) Platform track time-space resource matrix

Figure 2: Platform track time-space resource usage plan and time-space resource matrix

3 Modeling of re-optimization of train platforming

3.1 Parameters description

Parameters of this study are defined in Table 1. We assume that all parameters and values related to time are multiplies of the time unit $\Delta\tau$.

Table 1: Illustration of Parameters

Symbol	Definition
L	Set of trains, indexed by l
L_1	Set of delayed trains
I	Set of platform tracks, indexed by i
$c_{l,i}$	Cost of train l assigned to platform track i
π_l	The 0-1 parameter equals to 1 if train l is running in the inbound direction; 0, otherwise
$q_{l,i}$	The 0-1 parameter equals to 1 if train l was initially assigned to platform track i before a delay occurs; 0, otherwise
T	Length of the planning horizon
S	The time when train delays' information is updated
$t_{l,a}$	The scheduled arrival time of train l
$t_{l,d}$	The scheduled departure time of train l
$t_{l,a}^1$	The estimated arrival time of train l when a delay occurs
$t_{l,d}^1$	The estimated departure time of train l when a delay occurs
Δ_l	Dwell time of train l
P_l	Priority of train l
D	Safety time interval for platform track operation
MT	Sum of the length of the planning horizon T and the safety time interval for platform track operation D
$\Delta_{\max}$	Maximum dwell time among trains
h_a	Headway between two arrival trains running in the same direction
h_d	Headway between two departure trains running in the same direction
α	Objective function weighting factor
M	A sufficiently large number

3.2 Variable definitions

For each train $l, k \in L$, each platform track i ($i \in I$), and each moment t ($1 \le t \le MT$), the following variables are defined in the model.

(1) Platform track choice variable $w_{l,i}$ and $z_{l,k}$

$$w_{l,i} = \begin{cases} 1, & \text{train } l \text{ chooses platform track } i \\ 0, & \text{othewise} \end{cases}$$

$$z_{l,k} = \begin{cases} 1, & \text{train } l \text{ and train } k \text{ chooses the same platform track} \\ 0, & \text{othewise} \end{cases}$$

(2) Platform track occupancy state variable $x_{l,i,t}$

$$x_{l,i,t}=\begin{cases}1, & \text{train } l \text{ occupies platform track } i \text{ at moment } t\\ 0, & \text{othewise}\end{cases}$$

(3) Platform track occupancy state variable $u_{l,i,t}$ and clearance state variable $v_{l,i,t}$

In order to describe arrival and departure process of train l, platform track occupancy state variable $u_{l,i,t}$ as well as platform track clearance state variable $v_{l,i,t}$ are defined to denote the state of platform track i when train l arrives at and leaves from platform track i.

$$u_{l,i,t}=\begin{cases}1, & \text{train } l \text{ has not yet arrived at platform track } i \text{ at moment } t\\ 0, & \text{othewise}\end{cases}$$

$$v_{l,i,t}=\begin{cases}1, & \text{train } l \text{ has left platform track } i \text{ at moment } t\\ 0, & \text{othewise}\end{cases}$$

(4) Train sequence variables $\lambda_{l,k}$ and $\mu_{l,k}$

In order to describe the sequences of trains arriving at and departing from stations, the train sequence variables $\lambda_{l,k}$ and $\mu_{l,k}$ are defined as follows.

$$\lambda_{l,k}=\begin{cases}1, & \text{train } l \text{ arrives at the station before train } k\\ 0, & \text{othewise}\end{cases}$$

$$\mu_{l,k}=\begin{cases}1, & \text{train } l \text{ departs from the station before train } k\\ 0, & \text{othewise}\end{cases}$$

3.3 Objective function

The objective function in equation (1) contains the weighted sum of two parts. The former part is the sum of train arrival and departure delays, considering the train priority P_l and the weighting factor α, and the latter part is the total platform track occupancy costs.

$$\min z=\alpha\sum_{l\in L}P_l\left[(y_{l,a}-t_{l,a})+(y_{l,d}-t_{l,d}-D)\right]+\sum_{l\in L}\sum_{i\in I_l}w_{l,i}c_{l,i} \tag{1}$$

3.4 Constraints

According to definitions of $x_{l,i,t}$, $u_{l,i,t}$, and $v_{l,i,t}$, the relationship among those variables can be expressed in constraint (1). Constraints (2) and (3) show that values of actual arrival time $y_{l,a}$ and actual departure time $y_{l,d}$ of train l can be inferred from $u_{l,i,t}$ and $v_{l,i,t}$. Constraint (5) requires that each train l can only be assigned to one platform track. Constraint (6) ensures that any platform track i can only be occupied by at most one train at any time t. Constraints (7)–(9) guarantee that train operations on the platform tracks should be consecutive, by enforcing the condition that the values of variables $u_{l,i,t}$ and $v_{l,i,t}$ are continuous. Constraints (10)–(15) impose the required safety headway between two arriving or departing trains running in the same direction, and the train sequence variables $\lambda_{l,k}$, $\mu_{l,k}$ as well as the platform track choice variable $w_{l,i}$ and

$z_{l,k}$ are also embedded in those constraints. Constraint (16) enforces the minimum dwell time for each train l. Note that the safety time interval for platform track operation D is included in the right side of the constraints so that the required safety time interval for trains assigned to the same platform track is imposed. In addition, the minimum dwell time Δ_l of a train *l* is a deterministic value. Constraints (17)–(19) specify that the actual arrival and departure times of the trains should be no less than the corresponding planned arrival and departure times, respectively. Constraints (20)–(26) assign initial values to the variables $u_{l,i,t}$, $v_{l,i,t}$, $w_{l,i}$, $y_{l,a}$ and $y_{l,a}$, so that all trains adhere to their original plan before the train delay occurs. Finally, constraints (27)–(29) define the domain of variables. Note that $x_{l,i,t}$, $y_{l,a}$ and $y_{l,d}$ are intermediate variables to facilitate model definition, and their values can be inferred from $u_{l,i,t}$ and $v_{l,i,t}$.

$$x_{l,i,t} = 1-(u_{l,i,t}+v_{l,i,t}) \tag{2}$$

$$y_{l,a} = MT - \sum_{i\in I}\sum_{t=1}^{MT}(1-u_{l,i,t}) \tag{3}$$

$$y_{l,d} = MT - \sum_{i\in I}\sum_{t=1}^{MT} v_{l,i,t} \tag{4}$$

$$\sum_{i\in I} w_{l,i} = 1, \quad \forall l \in L \tag{5}$$

$$\sum_{l\in L} x_{l,i,t} \le 1, \quad \forall i \in I, \ \forall 1 \le t \le MT \tag{6}$$

$$u_{l,i,t} \ge u_{l,i,t+1} + w_{l,i} - 1, \quad \forall l \in L, \forall i \in I, \forall 1 \le t < MT \tag{7}$$

$$v_{l,i,t} \le v_{l,i,t+1} - w_{l,i} + 1, \quad \forall l \in L, \forall i \in I, \forall 1 \le t < MT \tag{8}$$

$$u_{l,i,t} \le u_{l,i,t+1} + w_{l,i}, \quad \forall l \in L, \forall i \in I, \forall 1 \le t < MT \tag{9}$$

$$y_{l,a} - y_{k,a} \ge (1-z_{l,k})h_a + z_{l,k}D - \lambda_{l,k}M, \quad \forall l,k \in L : l \ne k, \pi_l = \pi_k \tag{10}$$

$$y_{l,d} - y_{k,d} \ge (1-z_{l,k})h_d + z_{l,k}D - \mu_{l,k}M, \quad \forall l,k \in L : l \ne k, \pi_l = \pi_k \tag{11}$$

$$z_{l,k} \ge w_{l,i} + w_{k,i} - 1, \quad \forall l,k \in L, \forall i \in I_l \cap I_k : k > l, \pi_l = \pi_k \tag{12}$$

$$z_{l,k} = z_{k,l}, \quad \forall l,k \in L : k > l, \pi_l = \pi_k \tag{13}$$

$$\lambda_{l,k} + \lambda_{k,l} = 1, \quad \forall l,k \in L : k > l, \pi_l = \pi_k \tag{14}$$

$$\mu_{l,k} + \mu_{k,l} = 1, \quad \forall l,k \in L : k > l, \pi_l = \pi_k \tag{15}$$

$$\sum_{t=1}^{MT} x_{l,i,t} \ge w_{l,i} \times (\Delta_l + D), \quad \forall l \in L, \forall i \in I \tag{16}$$

$$y_{l,a} \ge t_{l,a}, \quad \forall l \in L \tag{17}$$

$$y_{l,d} \ge t_{l,d} + D, \quad \forall l \in L \tag{18}$$

$$y_{l,d} \ge y_{l,a} + \Delta_l + D, \quad \forall l \in L \tag{19}$$

$$u_{l,i,1} = 1, \quad \forall l \in L, \forall i \in I \tag{20}$$

$$v_{l,i,1} = 0, \quad \forall l \in L, \forall i \in I \tag{21}$$

$$w_{l,i} = q_{l,i}, \quad \forall l \in L, \forall i \in I : t_{l,a} < S \tag{22}$$

$$y_{l,a} = t_{l,a}, \quad \forall l \in L : t_{l,a} < S \tag{23}$$

$$y_{l,d} = t_{l,d}, \quad \forall l \in L : t_{l,a} < S \tag{24}$$

$$t_{l,a} = t^1_{l,a}, \quad \forall l \in L_1 \tag{25}$$

$$t_{l,d} = t^1_{l,d}, \quad \forall l \in L_1 \tag{26}$$

$$w_{l,i} = \{0,1\}, \quad \forall l \in L, \forall i \in I \tag{27}$$

$$u_{l,i,t}, v_{l,i,t} = \{0,1\}, \quad \forall l \in L, \forall i \in I, \forall 1 \le t \le MT \tag{28}$$

$$z_{l,k},\ \lambda_{l,k},\ \mu_{l,k} = \{0,1\}, \quad \forall l,k \in L : l \ne k, \pi_l = \pi_k \tag{29}$$

3.5 Valid equalities

Valid equalities are constraints that can strengthen the model formulation, as shown in constraints (30)–(33).

$$u_{l,i,t} \ge 1 - w_{l,i}, \quad \forall l \in L, \forall i \in I, \forall 1 \le t \le MT \tag{30}$$

$$v_{l,i,t} \le w_{l,i}, \quad \forall l \in L, \forall i \in I, \forall 1 \le t \le MT \tag{31}$$

$$x_{l,i,t} \le w_{l,i}, \quad \forall l \in L, \forall i \in I, \forall 1 \le t \le MT \tag{32}$$

$$x_{l,i,t} = 0, \quad \forall l \in L, \forall i \in I, \forall t < t_{l,a} \text{ or } t > t_{l,d} + \Delta_{\max} + D \tag{33}$$

Principle of valid inequalities (30), (31) and (32) are similar. For example, in valid inequality (30), if train l occupies platform track i, then valid inequality (30) is equivalent to $u_{l,i,t} \ge 0$ which turns out to be ineffective. However, if train l does not occupy the platform track i, then valid inequality (30) is equivalent to $u_{l,i,t} \ge 1$ which implies $u_{l,i,t}$=1. Valid inequality (33) considers when the station capacity is not sufficient and two conflicting trains need to be assigned to the same platform track, then one of the two trains with lower priority can be delayed at most by $\Delta_{\max}$, which means $x_{l,i,t}$ can be constrained to 0 when $t < t_{l,a}$ or $t > t_{l,d} + \Delta_{\max} + D$.

4 Genetic and simulated annealing hybrid algorithm

In order to recover the train operations as soon as possible in case of train delays, a genetic and simulated annealing hybrid algorithm (GSAHA) is designed to solve the optimization model efficiently and effectively (Xing et al., 2006). The GSAHA algorithm combines the advantages of genetic algorithm (GA) and simulated annealing algorithm (SA). Moreover, GSAHA is robust on the convergence performance while avoiding being trapped into the local optimal solutions. The implementation details for the components of GSAHA are illustrated as follows.

4.1 Chromosome representation

Fig. 3 shows the one-dimensional real-value encoding method that is used to represent chromosomes. Each chromosome denotes a platform track assignment plan, i.e., if the value

of the l^{th} gene is equal to i, then the l^{th} train is assigned to platform track i with its scheduled arrival and departure time. The length of each chromosome is equal to the number of trains $|L|$, and the genes in a chromosome are numbered in decreasing order according to the scheduled arrival time of trains, where the value range of each gene is located within the range $[1, |I|]$, and there could be $|I|^{|L|}$ chromosomes in total.

Platform track	5	2	3	6	...	4	7	1
Train	1	2	3	4	...	$\|L\|-2$	$\|L\|-1$	$\|L\|$

Figure 3: Illustration of chromosome representation

4.2 Generate initial population

Considering diversity and rationality of individuals in the initial population, the following strategies are proposed to generate the initial population.

Step 1. Denote platform tracks whose number is smaller than the number of platform tracks $|I|$ as the set I_1.

Step 2. Select $\lfloor |L| / (|I|-1) \rfloor$ trains that have not been selected yet and assign those trains to one of the unassigned platform tracks in set I_1.

Step 3. Repeat Step 1 until all platform tracks in set I_1 are assigned, and assign the rest $|L| - \lfloor |L| / (|I|-1) \rfloor \cdot (|I|-1)$ trains to the last platform track numbered as $|I|$.

Step 4. Repeat Step 2 and Step 3 until all individuals in the initial population are generated.

4.3 Obtain a feasible solution

The chromosome designed in Section 4.1 only assigns trains to platform tracks, i.e., to determine the platform track spatial resources that each train occupies. However, it is still possible that two trains may conflict with each other on the occupation of platform track temporal resources due to the violation of safety headway requirements, namely, the headway between two trains assigned to the same platform track D, headway between two arrival trains running in the same direction h_a, and headway between two departure trains running in the same direction h_d. Hence, a heuristic rule is designed to resolve the temporal conflicts according to the constraints in Section 3.4:

Step 1. Sort all trains in decreasing order by their scheduled or estimated arrival time and number them from 1 to $|L|$.

Step 2. Use **Algorithm 1** to resolve the temporal conflicts between any two trains in the order given in Step 1. Note that **Algorithm 1** will not lead to a deadlock between trains where trains can always be delayed to resolve the temporal conflicts.

Algorithm 1: heuristic method to resolve the temporal conflicts with given train order

For each train i $(1 \le i \le |L|)$

  For each train j $(1 \le j < i)$

    If train i conflicts with train j

Fix the arrival and departure times of train i and record the weighted-sum delay amount ϑ_i after resolving the conflicts of trains number before train i;
Fix the arrival and departure times of train j and record the weighted-sum delay amount ϑ_j after resolving the conflicts of trains number before train i;
If $\vartheta_j \le \vartheta_i$
Adopt the adjust method by fixing the arrival and departure times of train i;
Else
Adopt the adjust method by fixing the arrival and departure times of train j;
End If $\vartheta_j \le \vartheta_i$
End If train i conflicts with train j
End For each train j $(1 \le j < i)$
End For each train i $(1 \le i \le |L|)$

Step 3. Calculate the weighted sum of arrival and departure delays compared to the scheduled or estimated arrival and departure times. This operation considers all trains in set L and the platform track occupancy costs. The value calculated during this step serves as the objective value of the corresponding chromosome.

4.4 Fitness function

The fitness function in equation (34) is designed to evaluate each individual such that the algorithm can achieve a better convergence performance:

$$f_i(t_k) = \exp\left\{-\frac{f(i) - f_{\min}}{t_k}\right\}, \tag{34}$$

where t_k represents the temperature at the k^{th} generation, $f(i)$ represents the objective value of the i^{th} chromosome, $f_{\min}$ represents the minimal objective value at the k^{th} generation, and $f_i(t_k)$ represents fitness value of the i^{th} chromosome when the temperature is t_k. Fitness function in equation (34) is an important feature of the simulated annealing (SA) algorithm, and it has a good capacity to accelerate the convergence of the algorithm.

4.5 Temperature decline function

After determining the initial temperature T, the temperature decline function in equation (35) is used to lower the temperature at each iteration:

$$t_k = T \cdot \alpha^k, \tag{35}$$

where t_k represents the temperature at the k^{th} generation, and the constant α

represents the temperature decline rate in the SA algorithm.

4.6 Genetic operators

Neighborhood Search

Neighborhood search operator is applied to every chromosome. For instance, neighborhood search operator modifies the value of one gene in chromosome i randomly to generate a new chromosome j, and the objective value $f(j)$ of chromosome j is recalculated. Chromosome j is accepted or rejected to replace chromosome i according to the probability $P_{ij}(t_k)$ in equation (36).

$$P_{ij}(t_k) = \min\left\{1, \exp\left(-\frac{f(j) - f(i)}{t_k}\right)\right\} \tag{36}$$

If $P_{ij}(t_k)$ is greater than the random number r_1 generated within the range [0, 1), then chromosome i is replaced by chromosome j. Neighborhood search operator is another important feature of the SA algorithm and it can enlarge the search space with the probability of resulting in better solutions. Moreover, neighborhood search operator is one of the main operators that can increase population diversity when the algorithm is trapped into local optimal solutions.

Selection

Roulette method is adopted to select parents according to the cumulative probability, as shown in equation (37):

$$C_i = \sum_{k=1}^{i} f_k \Big/ \sum_{k=1}^{N} f_k \quad , \tag{37}$$

where N represents the number of individuals in the population. A random number r_2 is generated within [0, 1), if $r_2 \in (C_i, C_j)$, then chromosome j is chosen as a parent. The elitism strategy is used to reduce randomness of the algorithm. Additionally, individuals are restricted to be consecutively chosen as parents to avoid the situation when the algorithm is trapped into a local optimal solution too early.

Crossover

Two individuals are chosen as parents each time and a random number r_3 is generated within the range [0, 1). If r_3 is greater than or equal to the given crossover rate, then the crossover operator is not used and the two parents are reserved as children directly; otherwise, 2-point crossover operator is performed.

Mutation

For each gene of a chromosome, a random number r_4 is generated within the range [0, 1). If r_4 is smaller than the given mutation rate, then the mutation operator is applied, i.e., a different platform track is randomly assigned to the gene.

5 Numerical experiments

The proposed model is applied to a railway passenger station as shown in Fig. 4, with five

platform tracks (I, 3, 5, 7, 9, 11) in the inbound direction, and four platform tracks (II, 4, 6, 8, 10) in the outbound direction. The time unit $\Delta\tau$ is set as 1 min. There are 70 inbound and outbound trains which need to conduct the necessary operations from 16:00 to 22:00. Trains have assigned priorities from 1 to 3, and the initially scheduled train operation plan within the station is illustrated as shown in Table 2 and Fig. 5. Additionally, the platform track occupancy costs for the inbound and outbound trains are given in Tables 3 and 4. There is a penalty of 10,000 for trains which use the platform tracks in the opposite direction. Moreover, it is known that 6 inbound trains and 4 outbound trains are delayed at 18:38, and the estimated arrival and departure times of those delayed trains are given in Table 5. The maximum dwell time $\Delta_{\max}$ is 43 min, and the length of the scheduled horizon T is 360 min. The safety interval time on the platform track D is 6 min, and the headway between two arriving or departing trains in the same direction are set as 5 min. The weighting factor α is set as 200. Please note that the value of α can be flexible adjusted by the train dispatchers. In addition, we believe that keeping trains on time with guaranteed train service quality is more important than assigning the trains to their preferred platform tracks, and thus the penalty parameters on train delays are relatively larger than the platform track occupancy costs.

First, we use the commercial solver CPLEX 12.7.0 to solve the model in section 2. The test computer is an Intel(R) Core(TM) i7-5600U 2.6GHZ CPU with 12G RAM. CPLEX can obtain optimal solutions after *608 seconds* with the *objective value of 17,059*. Table 7 shows that arrival and departure times of 11 trains are delayed after the re-optimization, and 14 trains are assigned different platform tracks. The new train operation plan within the station is shown in Table 6 and Fig. 6, where all safety headway requirements are satisfied. Please note that the trains in Table 6 with bold fonts represent that those trains have been reassigned to different platform tracks.

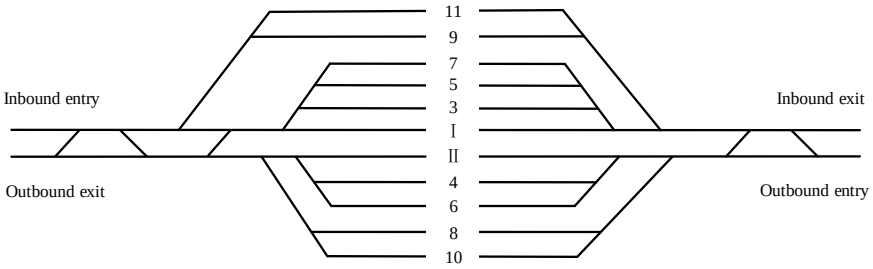


Figure 4: Layout of the railway passenger station

Table 2: Initial platform tack assignment plan between 16:00 and 22:00

Platform track number	Occupation trains
11	T_9, T_{29}
9	T_5, T_{19}, T_{31}, T_{41}, T_{49}
7	T_{11}, T_{21}, T_{27}, T_{33}, T_{43}, T_{47}, T_{55}, T_{63}, T_{69}
5	T_1, T_7, T_{15}, T_{25}, T_{35}, T_{39}, T_{53}, T_{61}, T_{67}, T_{73}
3	T_3, T_{13}, T_{17}, T_{23}, T_{37}, T_{45}, T_{51}, T_{57}, T_{59}, T_{65}, T_{71}, T_{75}
I	
II	
4	T_2, T_8, T_{18}, T_{22}, T_{30}, T_{34}, T_{40}, T_{42}, T_{48}, T_{54}, T_{60}, T_{64}
6	T_4, T_{12}, T_{16}, T_{24}, T_{32}, T_{38}, T_{44}, T_{50}, T_{56}, T_{62}

8	T_6, T_{14}, T_{20}, T_{28}, T_{36}, T_{46}, T_{52}, T_{58}
10	T_{10}, T_{26}

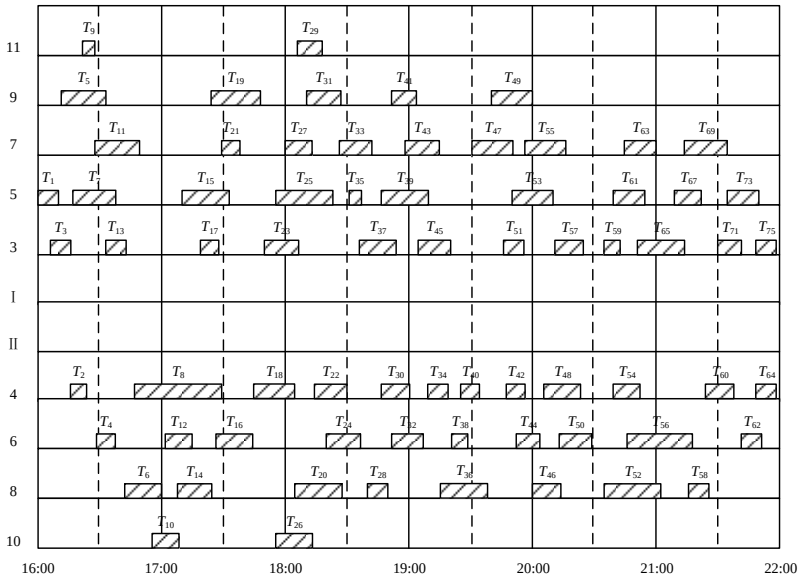


Figure 5: Arrival and departure track utilization scheme between 16:00 and 22:00

Table 3: Platform track occupancy costs for inbound trains with different priorities

Train direction	Train priority	Platform track number					
		I	3	5	7	9	11
Inbound	1	600	6	12	24	48	96
	2	300	3	6	12	24	48
	3	200	2	4	8	16	32

Table 4: Platform track occupancy costs for outbound trains with different priorities

Train direction	Train priority	Platform track number				
		II	4	6	8	10
Outbound	1	600	6	12	24	48
	2	300	3	6	12	24
	3	200	2	4	8	16

Table 5: Estimated arrival and departure times for delayed trains

Train	Arrival delay	Expected arrival time	Expected departure time	Dwell time	Priority
T_{39}	20	187	210	23	1
T_{41}	25	197	209	12	3
T_{43}	25	203	220	17	3
T_{45}	27	211	227	16	1
T_{47}	30	240	260	20	3
T_{55}	30	266	286	20	3

T_{32}	30	202	217	15	3
T_{36}	32	227	250	23	3
T_{38}	35	235	243	8	3
T_{44}	40	272	283	11	1

Table 6: Platform tack assignment plan after re-optimization with CPLEX

Platform track number	Occupation trains
11	T_9, T_{29}
9	T_5, T_{19}, T_{31}, T_{41}, T_{49}, $\boldsymbol{T_{55}}$
7	T_{11}, T_{21}, T_{27}, T_{33}, T_{43}, $\boldsymbol{T_{53}}$, T_{63}, T_{69}
5	T_1, T_7, T_{15}, T_{25}, T_{35}, $\boldsymbol{T_{45}}$, $\boldsymbol{T_{47}}$, T_{61}, T_{67}, T_{73}
3	T_3, T_{13}, T_{17}, T_{23}, T_{37}, $\boldsymbol{T_{39}}$, T_{51}, T_{57}, T_{59}, T_{65}, T_{71}, T_{75}
I	
II	
4	T_2, T_8, T_{18}, T_{22}, $\boldsymbol{T_{28}}$, T_{34}, T_{40}, T_{42}, T_{48}, $\boldsymbol{T_{44}}$, T_{60}, T_{64}
6	T_4, T_{12}, T_{16}, T_{24}, $\boldsymbol{T_{30}}$, T_{32}, T_{38}, T_{50}, $\boldsymbol{T_{54}}$, $\boldsymbol{T_{58}}$, T_{62}
8	T_6, T_{14}, T_{20}, T_{46}, $\boldsymbol{T_{56}}$
10	T_{10}, T_{26}, $\boldsymbol{T_{36}}$, $\boldsymbol{T_{52}}$

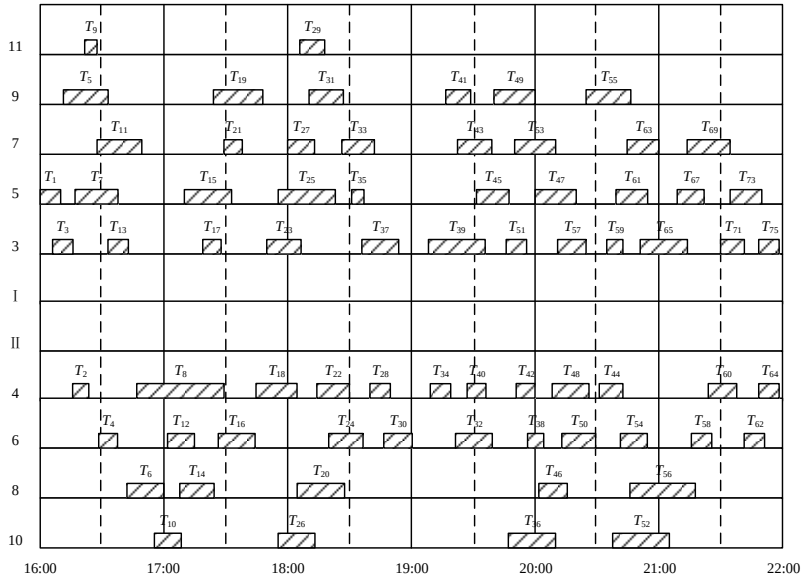


Figure 6: Platform track use scheme after re-optimization with CPLEX

Table 7: Amount of secondary delay for the trains obtained by CPLEX

Train	Priority	Secondary arrival delay(min)	Secondary departure delay (min)
T_{39}	1	0	4
T_{55}	3	0	2
T_{32}	3	0	3

T_{38}	3	2	2
T_{40}	3	2	2
T_{42}	1	5	5
T_{46}	2	2	2
T_{48}	1	2	2
T_{50}	1	0	1
T_{52}	3	2	2
T_{54}	1	2	2

Parameters for GSAHA are set as follows. The number of individuals in the population is 50, the maximum number of generations is 300, the crossover rate is 0.98, the mutation rate is 0.1, the initial temperature *T* is 8000 ℃, the temperature decline rate α is 0.9, and the temperature is increased to 4000 ℃ if the objective value of the best individual in the current generation remains unchanged for 3 iterations. The GSAHA is implemented in C++, and the average objective value of GSAHA for total 20 runs is *17,612*, which is only 3.24% higher than the optimal solution of CPLEX. In addition, the average running time of the GSAHA is only *27 seconds*. The convergence process of the simulated annealing hybrid algorithm for a specific run is shown in Fig. 7, where the algorithm can reach the near-optimal solution only after 70 iterations.

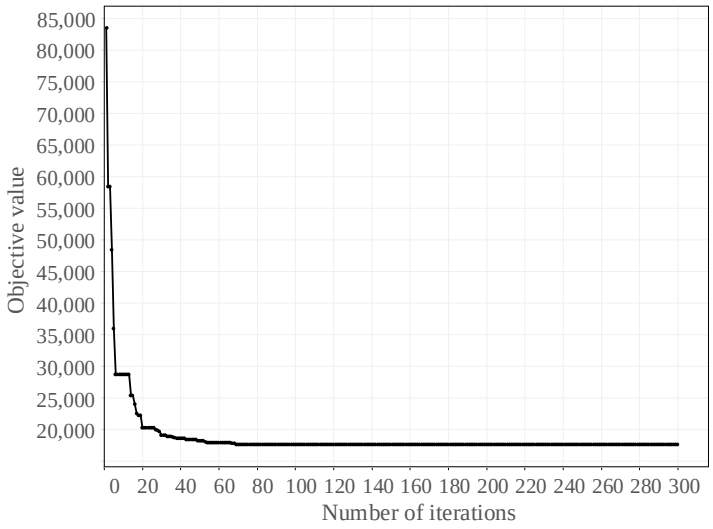


Figure 7: Convergence process of GSAHA

Meanwhile, the sensitivity analysis of different values of weighting factor α is performed by increasing the value of α from 40 to 440 with the step size equal to 40. The optimization results of CPLEX and GSAHA are listed in Table 8, and the parameter settings for the GSAHA remain unchanged, and the objective value of GSAHA takes the average results of 20 times. It can be shown that the objective values of CPLEX and GSAHA increase as the value of α increases, and the solution times of CPLEX range from *329 to 764 seconds* while the solution times of GSAHA only range from *27 to 29 seconds*. In addition, the objective values of GSAHA are *2.80%–5.10%* larger than that of CPLEX.

Hence, the stable performance of GSAHA regarding the solution quality and solution times show that our proposed GSAHA is suitable to serve as an effective computer-aided decision-making tool for the train dispatchers in case of train delays.

Table 8: Optimization results of CPLEX and the GSAHA with different weighting factors

Weighting factor α	CPLEX		GSAHA		
	Objective value	CPU time (sec)	Objective value	CPU time (sec)	GAP with CPLEX (%)
40	3939	740	4140	28	5.10
80	7219	589	7507	28	3.99
120	10499	764	10914	28	3.95
160	13783	447	14233	28	3.26
200	17059	679	17612	27	3.24
240	20339	388	20951	27	3.01
280	23619	360	24342	27	3.06
320	26899	596	27681	28	2.91
360	30179	329	31069	28	2.95
400	33459	340	34402	29	2.82
440	36739	412	37808	28	2.91

6 Conclusions

The problem of re-optimization of the train platforming is essential in recovering the train operations within the station and minimizing the negative influences of train delays. This paper proposes a MILP re-optimization model, where the train station is represented using discretized platform track time-space resources. The resulting model is solved by CPLEX and the designed heuristic algorithm GSAHA. The effectiveness of the proposed MILP model is verified by using the CPLEX solver, and the proposed heuristic algorithm further speeds up the solving process with near-optimal solutions. In addition, the performance of GSAHA is stable when the values of weighting factor α vary from 40 to 440.

The work in this paper can be extended in several interesting directions. First, instead of ensuring the arrival and departure safety headway between two different trains (Chakroborty and Vikram, 2008), the explicit consideration of train entrance and exit route conflicts can increase the station throughput capacity and reduce the train delays (Zwaneveld et al., 1996). Second, the MILP model and the heuristic algorithm GSAHA proposed in this paper can be further developed to consider different station types, such as the terminal station where trains need to perform the turn-around movement which makes the train platforming problem more complicated. Third, the effectiveness of the heuristic algorithm GSAHA can be tested and improved for bigger railway stations with more complex station layout structure.

Acknowledgements

This work is supported by the National Key R&D Program [grant no. 2017YFB1200700] and National Nature Science Foundation of China [grant no. 71871188, grant no.

U1834209]. The first and second authors gratefully acknowledges the financial support from the China Scholarship Council [201707000080, 201707000041].

References

Caprara, A., L. Galli, P. Toth., 2011. “Solution of the train platforming problem”. *Transportation Science*, Vol. 45, No. 2, pp. 246-257.

Chakroborty, P., Vikram, D., 2008. “Optimum assignment of trains to platforms under partial schedule compliance”. *Transportation Research Part B: Methodological*, 42(2), 169-184.

Kroon, L. G., H. Edwin Romeijn, P. J. Zwaneveld., 1997. “Routing trains through railway stations: complexity issues”. *European Journal of Operational Research*, Vol. 98, No. 3, pp. 485-498.

Lusby, R. M., J. Larsen, M.E. David Ryan., 2011. “Railway track allocation: models and methods”. *OR spectrum*, Vol. 33, No. 4, pp. 843-883.

Lusby, R., J. Larsen, M. Ehrgott., 2013. “A set packing inspired method for real-time junction train routing”. *Computers & Operations Research*, Vol. 40, No. 3, pp. 713-724.

Sels, P., P. Vansteenwegen, and T. Dewilde. The train platforming problem: the infrastructure management company perspective. *Transportation Research Part B: Methodological*, Vol. 61, 2014, pp. 55-72.

Xing, W., Xie, J., 2006. “Modern optimization methods (2nd Edition)”. Beijing: Tsinghua University Press, 113-147.

Zwaneveld, P. J., L. G. Kroon, H. Edwin Romeijn, 1996. “Routing trains through railway stations: model formulation and algorithms”. *Transportation Science*, Vol. 30, No. 3, pp. 181-194.

The Comparison of Three Strategies in Capacity-oriented Cyclic Timetabling for High-speed Railway

Xin Zhang[a], Lei Nie[a,1], Yu Ke[a]
[a] School of Traffic and Transportation, Beijing Jiaotong University, Beijing, China
[1] E-mail: lnie@bjtu.edu.cn, lnie8509@yahoo.com, Phone: + 86 010 51688173

Abstract
The expansion of the scale of high-speed railway networks and the growth of passenger demand imply a high frequency of high-speed trains in China, i.e. higher railway capacity utilization. Based on given infrastructures and train line plans, there are some timetabling strategies which affect the capacity utilization, e.g. changing train departure sequence at origin stations, overtakings between trains, and adding new train stop at stations. Nowadays, managers of high-speed railway in China are eager to find out that what kind of impact these strategies have on the capacity utilization. In this study, new variables of train stops and constraints of overtakings are proposed with an extended cyclic timetabling model based on the periodic event scheduling problem (PESP). Minimum cycle time, train travel time and the total number of train stops are calculated as objectives to measure the differences between the strategies. The effectiveness of the three timetabling strategies are compared and presented by a series of experiments based on one real-world rail line in China. According to our results, with flexible train departure sequence at the origin stations and train overtakings, the possibility of acquiring good capacity utilization can be higher, but too many overtakings will have negative effect on the quality of timetable. The effectiveness of adding new stops on the capacity utilization depends on the ways of adding stops, i.e. which train is allowed to be added new stops and which stations can be selected to stop at.

Keywords
Cyclic timetabling, Capacity utilization, Train sequence, Train stop, Overtaking

1 Introduction

With the expansion of the scale of high-speed railway networks in China, the exchange among the different regional areas causes the passenger flow volumes to expand, which implies more high-speed trains, i.e. better capacity utilization. In general, there are two kinds of ways to improve the capacity utilization, i.e. upgrade railway infrastructure and equipment, and increase the efficiency of transportation. Compared with the former, improving train operating plans for the efficiency of transportation, e.g. improving line plans and timetables, can be low-cost, since upgrading infrastructure/equipment always needs more time and money. Therefore, capacity-oriented timetabling is necessary for improving railway capacity utilization and transport management.

Railway timetabling and railway capacity analysis has been deeply studied in recent years. Based on given infrastructures and train line plans, there are some timetabling strategies which affect the capacity utilization, e.g. changing the train departure sequence in the origin stations, overtakings between trains and adding new train stops at intermediate stations. Nowadays, the railway company of China is eager for higher

capacity utilization, i.e. operating more trains in limited time period. However, what kind of strategies is better or easier to improve the capacity utilization with acceptable cost is not studied deeply. Sparing and Goverde (2013) discussed the capacity utilization as follows:

"*The relationship between the nominal and the minimum cycle time describes the capacity utilization of the timetable (Hansen and Pachl (2008)): the timetable is stable exactly if the minimum cycle time T is less than the nominal cycle time T_0, i.e. $T<T_0$, and the larger T_0 -T is, the more time reserve there is available.*"

In this paper, the minimum cycle time of operating a series of trains are considered as one index of the capacity utilization, i.e. the smaller the minimum cycle time is, the more opportunities that we have to design smaller time period to operate the trains. In other words, based on given operating time period, if the minimum cycle time of a series of trains is small, more other trains can be operated in the remaining time of the given operating time period, i.e. it is possible to operate more trains in the given operating time period and the capacity utilization can be increased. In this study, with given train line plan, a cyclic timetabling model based on the periodic event scheduling problem (PESP) is built. New variables and constraints to modified train stop plans and describe train overtakings at stations are introduced. On the one hand, the three timetabling strategies, i.e. train departure sequence at the origin stations, overtakings at stations and new train stops, are described in different constraints, and each strategy can be considered by using the corresponding constraint in the model. On the other hand, the minimum cycle time, train travel time and the number of new train stops are used as objectives to measure the differences between the strategies.

This study is a further study of our previous paper, i.e. Zhang and Nie (2016). Literature review is presented in Section 2. The cyclic timetabling problem and the model are displayed in Section 3 and 4, respectively. New variables and the constraints of the three timetabling strategies are included. Experiments based on one real-world case of high-speed rail corridor in China and detailed conclusions are presented in Section 5. Finally, conclusions are included in Section 6.

2 Literature Review

In recent years, many remarkable studies have been devoted to train timetabling (e.g., Caprara et al. (2006); Zhou and Zhong (2007); Salido and Barber (2009); Goverde (2010); Cacchiani and Toth (2012); Harrod (2012); Schmidt and Schöbel (2015)). Among the research performed on cyclic train timetabling, models based on the periodic event scheduling problem (PESP), which was introduced by Serafini and Ukovich in 1989 (Peeters (2003)), have demonstrated great power in periodic railway timetabling. A PESP-based model for the cyclic railway timetabling problem (CRTP) was first considered in 1993, and a stronger model, the cycle periodicity formulation (CPF) was introduced. The PESP and the CPF are based on the construction of an auxiliary graph, whose nodes correspond to events (train departures and arrivals) and whose arcs model the constraints acting on the time separations between those events (Cordone and Redaelli (2011)). This auxiliary graph, known as the event-activity network (EAN), which is also used in this paper, has been widely applied in the literatures on train timetabling (e.g., Kroon and Peeters (2003); Schöbel (2007); Liebchen et al. (2010); Schachtebeck and Schöbel(2010)).

Many extended models and effective algorithms based on the PESP have been studied in depth in recent years (e.g., Kroon and Peeters (2003); Liebchen (2004); Mathias (2008); Xie and Nie (2009); Caimi et al. (2011); Cordone and Redaelli (2011); Kroon et al.

(2013)). With regard to operating rule constraints, Peeters (2003) and Caimi et al. (2011) discussed a non-collision constraint with variable trip time to prevent overtaking between successive stations. With regard to the objective function, an objective for the PESP based on the minimum cycle time T (i.e., the minimum period length of one regular timetable) was presented by Sparing and Goverde (2013, 2017), where the stability of the timetable is considered. Regarding applicable algorithms, Siebert and Goerigk (2013) studied a series of experimental comparisons of various extended PESP models (the Origin Destination aware PESP (ODPESP) and the Extended PESP (EPESP)) and three different methods based on the modulo simplex algorithm proposed by Nachtigall and Opitz (2008), which is a powerful heuristic for solving the PESP (Goerigk and Schöbel (2013)). For an in-depth overview of the PESP, the CRTP, and the CPF, we refer to Peeters (2003) as well as Liebchen (2006), Liebchen and Möhring (2007) and Liebchen et al. (2010). In particular, based on Heydar et al. (2013), Petering et al. (2015) presented an innovative mixed-integer linear programming model, which falls outside the framework of the PESP, of a cyclic train timetabling and platforming problem. The new model and their pre-processing techniques have great potential to analyse the railway capacity utilization based on various factors and the computation time is reasonable.

In many capacity analysis studies of cyclic timetables which have the same setting as ours, influencing factors such as train speed, line plan specifications (train stop plans), overtaking and train heterogeneity have been discussed (e.g., Burdett and Kozan (2006); Abril et al. (2008); Landex et al (2008); Zhu et al. (2009); Dicembre and Ricci (2011); Lindfeldt (2011); Lai and Wang (2012); Petering et al. (2015)). However, to our knowledge, this paper is the first study to build one cyclic timetabling model based on the PESP which includes new variables of adding train stops. Based on the model, it is possible to modify train stops while cyclic timetabling.

3 The Cyclic Timetabling Problem Defined

We now formally introduce the problem. Stations are presented by nodes in our cyclic timetabling problem. There is only one rail line for one direction and no sidings in block sections, so it is impossible for trains to overtake each other between two successive stations. In order to define the cyclic train timetabling problem, the event-activity network is presented first.

3.1 Event-Activity Network and sets

In cyclic timetabling based on the PESP, mathematical formulations are typically constructed in terms of events and activities. Before introducing these formulations, we assume that a public transportation network (PTN) and a *line* have been determined a priori.

Notation 1. A public transport network PTN=($\boldsymbol{S}$, $\boldsymbol{T}$) (where $\boldsymbol{S}$ is the set of nodes and $\boldsymbol{T}$ is the set of edges) is a simple, undirected graph in which the nodes represent stations and the edges represent connections between them. A line $\boldsymbol{l}$ is a path in the PTN, and $\boldsymbol{f}$ is the corresponding frequency of the line (Siebert and Goerigk, 2013). For cyclic timetables, the time horizon on which trains are scheduled, such as one hour or two hours, is usually considered to be the ***cycle time***.

The goal of our model is to determine the departure times and arrival times such that the cycle time, the number of new stops or total train travel time can be minimized.

Assume that a line plan is known, i.e., the stop plans of the lines (sequences of stations at which trains stop) and their corresponding frequencies are given. Then, a given line $\boldsymbol{l_*}$ can be transformed into its individual trains according to its frequency $\boldsymbol{f}$ (i.e. $\boldsymbol{l_{*,1}}$, $\boldsymbol{l_{*,2}}$,..., $\boldsymbol{l_{*,f}}$), and the PTN is thus transformed into EAN $= (\varepsilon, A)$, where ε is the set of events and A is the set of activities. Events can be arrivals at or departs from stations (define ε_{dep} as the set of departure event and ε_{arr} as the set of arrival event), i.e., $\varepsilon = (\varepsilon_{dep}, \varepsilon_{arr})$, and activities are the transitions between pairs of events. To distinguish different types of train operating behavior, the corresponding activity sets can be described as follows (see Table 1). Moreover, Figure 1 presents an example of the EAN.

Table 1: Sets in the EAN

Symbol	Definition
ε	Set of events (nodes)
ε_{dep}	Set of departure events, $\varepsilon_{dep} \subset \varepsilon$
ε_{arr}	Set of arrival events, $\varepsilon_{arr} \subset \varepsilon$
A	Set of activities (arcs)
A_{run}	Set of running activities, $A_{run} \subset A$
A_{run-o}	Set of running activities of trains which depart from their origin stations to intermediate stations (e.g. running activities from station A to B in Fig. 1), $A_{run-o} \subset A_{run}$
A_{run-d}	Set of running activities of trains which depart from intermediate stations to their destination stations (e.g. running activities from station C to D in Fig. 1), $A_{run-d} \subset A_{run}$
A_{run-n}	Set of running activities between intermediate stations of trains (e.g. running activities from station B to C in Fig. 1), $A_{run-n} \subset A_{run}$
A_{dwell}	Set of all dwelling activities at stations (i.e. one dwelling activity is from one arrival event to one departure event, and the train may stop at the stations), $A_{dwell} \subset A$
$A_{dwell-a}$	Set of alternative dwelling activities at stations (i.e. trains may stop at the stations or not), $A_{dwell-a} \subset A_{dwell}$
$A_{dwell-c}$	Set of common dwelling activities at stations (i.e. trains have to stop at the stations), $A_{dwell-c} \subset A_{dwell}$, $A_{dwell} = A_{dwell-a} \cup A_{dwell-c}$
A_{dwell}^{r}	Set of all dwelling activities of the same train r, $A_{dwell}^{r} \in SA_{dwell}$, $r = 1,2, \ldots$
A_{pass}	Set of passing activities at stations (i.e. one passing activity is from one arrival event to one departure event, but the times of the arrival and the departure events must be the same since the train passes the station), $A_{pass} \subset A$
A_{safe}	Set of safety activities between trains (i.e. connections between any two arrival events or departure events that interact with each other because they occupy the same physical infrastructure at minimum headway times), $A_{safe} \subset A$
$A_{regular}$	Set of regularity activities between two trains at their origin stations (i.e. connecting two departure events between successive trains of the same line), $A_{regular} \subset A$

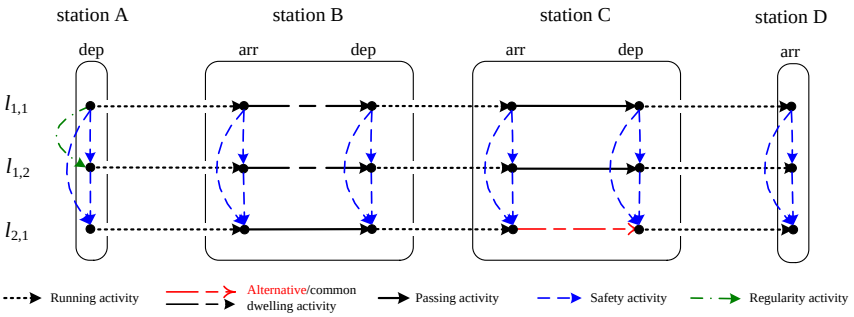


Figure 1: an example of the EAN

3.2 Parameters and Variables

Based on the assumptions and the EAN above, our problem is the train cyclic timetabling problem with stop planning (CTP-SP) of one rail corridor. Based on the structure of the PESP model, new variables of train stops are introduced and the CTP-SP model can be considered to be an extended model of the traditional PESP model. However, train stop plans are not allowed to be "regenerate" when timetabling, but modified according to the line plan, i.e. we are only adding limited number of train stops in this study. Meanwhile, the number of new train stops will be restricted in the model. It is also assumed that all trains will depart from the same station (i.e. the first station of the corridor according to one operation direction), such that the strategy of "train departure sequence" can be analysed. Table 2 and Table 3 present the subscripts, parameters and decision variables in the CTP-SP model, respectively. Mathematical formulations are presented in Section 4.

Table 2: Subscripts and parameters in the CTP-SP model

Symbol	Definition
i, j, i', j'	Indexes for the events
a	Activity, $a \in A$, $a_{ij} = (i, j)$, $i, j \in \varepsilon$
l_a	Lower duration bound of activity a, $l_a \in \mathbb{N}$, $0 \leq l_a \leq T - 1$
u_a	Upper duration bound of activity a, $u_a \in \mathbb{N}$, $0 \leq u_a \leq T - 1$ (different from the traditional PESP model, since our model can be linear only in this way)
$h_a(h'_a)$	Minimum headway time of activity a for two trains at the same station, $a \in A_{safe}$, $h_a \in \mathbb{N}$
f_a	Frequency of the line to which activity a belongs, $a \in A_{regular}$, $f_a \in \mathbb{N}^+$
δ	A nonnegative integer describing the relaxation level of regularity activity constraints, $\delta \in \mathbb{N}$
DE_a	Deceleration time loss of activity a, $a \in A_{run}$
AC_a	Acceleration time loss of activity a, $a \in A_{run}$
O	An index of overtaking, equals to 0 when overtakings are prevented, and a (very large) constant when overtakings are allowed
P^i	The maximum number of stops of train i
M_{order}	Sequence matrix of (some) departure events of trains at their origin stations, e.g. $M_{order} = [\pi_{16}, \pi_{10}, \dots]$
T_{min}	Minimum value of the cycle time T, $T_{min} \in \mathbb{N}$
T_{max}	Maximum value of the cycle time T, $T_{max} \in \mathbb{N}$

Table 3: Decision variables in the CTP-SP model

Symbol	Definition
T	The cycle time, $T \in \mathbb{N}$
π_i	Planned time for event i, $0 \leq \pi_i \leq T-1, \pi_i \in \mathbb{N}$
x_a	Planned duration for activity a, $a \in A$, $0 \leq x_a \leq T-1, x_a \in \mathbb{N}$
z_a	A binary variable that is equal to 1 when $\pi_i > \pi_j$ and equal to 0 otherwise (i.e., the modulo variable of activity a)
y_a	The value of $z_a \times T$ for activity a, equals to T when $\pi_i > \pi_j$ and equals to 0 otherwise, $a \in A$
p_a	A binary variable that is equal to 1 when the trains of activity a stops at the station, and equal to 0 otherwise, $a \in A_{dwell} \cup A_{pass}$
$w_{ii'jj'}$	An auxiliary integer variable which takes values of 0, 1 or 2 for activities a_{ij}, $a_{i'j'}$, $a_{ii'}$, and $a_{jj'}$. a_{ij}, $a_{i'j'} \in A_{run}$ and belong to the same section, $a_{ii'}$, $a_{jj'} \in A_{safe}$ (see Zhang and Nie (2016), Yan and Goverde (2018) for further explanations)
$o_{ii'jj'}$	A binary variable which takes the value of 0 when overtakings are prevented at the stations, and equal to 1 otherwise. a_{ij}, $a_{i'j'} \in A_{dwell} \cup A_{pass}$ and belong to the station, $a_{ii'}$, $a_{jj'} \in A_{safe}$(see Yan and Goverde (2018) for further explanations)
$v_{ii'jj'}$	An auxiliary integer variable which takes values of 0, 1 or 2 for activities a_{ij}, $a_{i'j'}$, $a_{ii'}$, and $a_{jj'}$. a_{ij}, $a_{i'j'} \in A_{dwell} \cup A_{pass}$ and belong to the same station, $a_{ii'}$, $a_{jj'} \in A_{safe}$ (see Yan and Goverde (2018) for further explanations)

Decision variables in our problems are defined as integers measured in minutes. In fact, this assumption is based on common operating parameters, and integers measured in seconds are also feasible for our model which may increase the computation time.

4 Mathematical Formulation of the Cyclic Timetabling Model

In this section, the mathematical formulations of the CTP-SP model are presented, and the objectives and the constraints are explained in detail.

(1) Objective functions:

O1: Minimize T, (1)

O2: Minimize $\sum p_a$, $\quad a \in A_{dwell-a}$, (2)

O3: Minimize $\sum x_a$, $\quad a \in A_{run} \cup A_{dwell}$. (3)

Objective function (1) strives to minimize the cycle time. The number of the new train stops and the total train travel time are minimized in Objective function (2) and (3). Each objective function can be calculated individually and iteratively, such that the model can be considered as single-objective and easier to be calculated.

(2) Constraints of events and activities:

$$x_a = \pi_j - \pi_i + y_a, \quad \forall a \in A_{run} \cup A_{dwell} \cup A_{pass} \cup A_{safe} \cup A_{regular}, \tag{4}$$

$$l_{a_{ij}} + p_{a_{jj'}} \times AC_{a_{ij}} \leq x_{a_{ij}} \leq u_{a_{ij}} + p_{a_{jj'}} \times AC_{a_{ij}}, \quad \forall a_{ij} \in A_{run-o}, a_{jj'} \in A_{dwell} \cup A_{pass}, \tag{5}$$

$$l_{a_{ij}} + p_{a_{i'i}} \times DE_{a_{ij}} \leq x_{a_{ij}} \leq u_{a_{ij}} + p_{a_{i'i}} \times DE_{a_{ij}}, \quad \forall a_{ij} \in A_{run-d}, a_{i'i} \in A_{dwell} \cup A_{pass}, \tag{6}$$

$$l_{a_{ij}} + (p_{a_{i'i}} \times DE_{a_{ij}} + p_{a_{jj'}} \times AC_{a_{ij}}) \le x_{a_{ij}} \le u_{a_{ij}} + (p_{a_{i'i}} \times DE_{a_{ij}} + p_{a_{jj'}} \times AC_{a_{ij}}),$$
$$\forall a_{ij} \in A_{run-n}, a_{i'i}, a_{jj'} \in A_{dwell} \cup A_{pass}, \quad (7)$$
$$l_a \le x_a \le u_a, \quad \forall a \in A_{run} \backslash (A_{run-o} \cup A_{run-d} \cup A_{run-n}), \quad (8)$$
$$h_a \le x_a \le T - h'_a, \quad \forall a \in A_{safe}, \quad (9)$$
$$l_a \times p_a \le x_a \le u_a \times p_a, \forall a \in A_{dwell} \cup A_{pass}, \quad (10)$$
$$\frac{T}{f_a} - \delta \le x_a \le \frac{T}{f_a} + \delta, \forall a \in A_{regular}, \quad (11)$$
$$z_{a_{ij}} + z_{a_{i'j'}} + z_{a_{ii'}} + z_{a_{jj'}} = 2 \times w_{ii'jj'}, \forall a_{ij}, a_{i'j'} \in A_{run}, a_{ii'}, a_{jj'} \in A_{safe}. \quad (12)$$

Constraint (4) defines the relationship between event times and activity durations. In the original PESP model, Constraints (4) are typically formulated as $x_{ij} = \pi_j - \pi_i + z_{ij} \times T$. However, T and z_{ij} are decision variables in our model, and the use of this equation will cause the model to be non-linear. To prevent the model from violating linear programming conditions, the new variables $y_{ij} = z_{ij} \times T$ are proposed by Sparing and Goverde (2013, 2017). The usage of the new variable requires that $0 \le u_a \le T - 1$, which is different from the traditional PESP models. Constraints (5)-(8) describe the lower and upper bounds of running activities and the relationship between the planned duration of activities and the variables of stops. A binary variable p_a is generated for each dwelling and passing activity since it will be easier to build these constraints. It is clear that one train needs time to decelerate and accelerate when it plans to stop at one station and the related constraints of running time should be modified. Safety operation of two trains using the same infrastructure (station) is guaranteed in Constraint (9). In Constraint (10) and (11), bounds of dwelling, passing and regularity activities are restricted. And Constraint (12) can prevent illegal overtakings between two successive stations in sections (see Zhang and Nie (2016), Yan and Goverde (2018) for further explanations).

(3) Constraints of the timetabling strategies:

$$z_{a_{ij}} + z_{a_{i'j'}} + z_{a_{ii'}} + z_{a_{jj'}} = 2 \times v_{ii'jj'} + o_{ii'jj'},$$
$$\forall a_{ij}, a_{i'j'} \in A_{dwell} \cup A_{pass}, a_{ii'}, a_{jj'} \in A_{safe}, \quad (13)$$
$$\sum o_{ii'jj'} \le O, \quad a_{ij}, a_{i'j'} \in A_{dwell} \cup A_{pass}, a_{ii'}, a_{jj'} \in A_{safe}, \quad (14)$$
$$\pi_{m_i} \le \pi_{m_j}, \quad \forall m_i, m_j \in M_{order}, i < j, \quad (15)$$
$$\sum_{a \in A^r_{dwell}} p_a \le P^r, \quad \forall A^r_{dwell} \in SA_{dwell}. \quad (16)$$

Overtakings at stations can be described in Constraints (13) and (14) by changing the value of parameter O, i.e. O equals zero when overtakings are prevented, and a very large constant when overtakings are allowed (see Yan et al. (2018) for further explanations). In fact, these constraints can be used to restrict the number of overtakings, but we will not extend this topic in this paper. In Constraint (15), the departure sequence of trains at the origin stations can be restricted. Clearly, it is possible that M_{order} is an empty set, such that the order of trains at the origin stations is flexible. As mentioned, train stop plans can be only modified by adding a limited number of stops of trains in this study. Therefore, the maximum number of stops of each train is restricted in Constraint (16).

(3) Logic constraints:

$$y_a \le T_{max} \times z_a, \quad \forall a \in A_{run} \cup A_{dwell} \cup A_{safe} \cup A_{regular}, \quad (17)$$
$$y_a \le T, \quad \forall a \in A_{run} \cup A_{dwell} \cup A_{safe} \cup A_{regular}, \quad (18)$$
$$y_a \ge T - T_{max} \times (1 - z_a),$$
$$\forall a \in A_{run} \cup A_{dwell} \cup A_{safe} \cup A_{regular}, \quad (19)$$
$$y_a \ge 0, \quad \forall a \in A_{run} \cup A_{dwell} \cup A_{safe} \cup A_{regular}, \quad (20)$$
$$x_a = 0, \quad \forall a \in A_{pass}, \quad (21)$$

$$p_a = 0, \qquad \forall a \in A_{pass}, \tag{22}$$
$$p_a = 1, \qquad \forall a \in A_{dwell-c}, \tag{23}$$
$$T_{min} \le T \le T_{max}, \tag{24}$$
$$z_a \in \{0,1\}, \qquad \forall a \in A_{run} \cup A_{dwell} \cup A_{safe} \cup A_{regular}, \tag{25}$$
$$p_a \in \{0,1\}, \qquad \forall a \in A_{dwell} \cup A_{pass}, \tag{26}$$
$$0 \le \pi_i \le T - 1, \qquad \forall i \in \varepsilon, \tag{27}$$
$$0 \le x_a \le T - 1, \qquad \forall a \in A. \tag{28}$$

Constraints (17)-(28) are logic constraints. Constraints (17)-(20) are used to linearize the model (see Sparing and Goverde (2013) for more details). In Constraint (24), it will be better if T_{min} is known since this parameter can reduce the solution space of the model. Otherwise, $T_{min} = 0$ can be accepted.

5 Experiments and Results

In this section, the comparison results of the three timetabling strategies are presented based on a series of experiments of the Beijing-Shanghai High-speed Railway in China (see Figure 2). There are 23 stations in the rail corridor. All trains in the experiments are chosen from one practical line plan, which run from Beijing South station to Shanghai Hongqiao station (see Table 4, Figure 3 and Table 5). Parameters including minimum headway time at stations, accelerating and decelerating time loss of trains refer to the practical data. When using the strategy of adding new stops, it is assumed that one new stop can be added for each train at most. Trains of type A are not allowed to be added new stops, except for the experiments in Section 5.3. Due to the requirements of service, trains of type B in Case2240 and Case2204 depart from their origin stations exactly every *T*/2, i.e. half of the minimum cycle time. Trains of type A run at speed of 350km/h, and trains of other types run at speed of 300km/h. In our opinion, Case0008 has higher train homogeneity compared to the other two cases since the trains have the same train speed and the number of stops at least. The model was coded by MATLAB R2012a and solved by Cplex 12.5. The calculations were performed on a PC with an Intel E7 2.0-GHz processor, 28 CPU cores and 256 GB of RAM. In general, the computation time is always about several seconds/minutes (average computation time of all presented cases is 47 minutes). Nevertheless, the computation times of those cases with "more flexible" strategies will be much longer, i.e. may cost several hours (12 hours at most). Some iterative ideas are used in our experiments to reduce the computation time (e.g. the method in Zhang and Nie (2016)). All of the solutions are optimal.

For all cases, Objective (1) will be used first (O1), then the value of the minimum cycle time is transformed into a constraint (i.e. to guarantee that the *T* equals to the minimum cycle time) and Objective (2) will be used ("O1"+O2). After that, both of the values of the minimum cycle time and the minimum number of train stops are transformed into constraints, and Objective (3) will be calculated ("O1+O2"+O3).

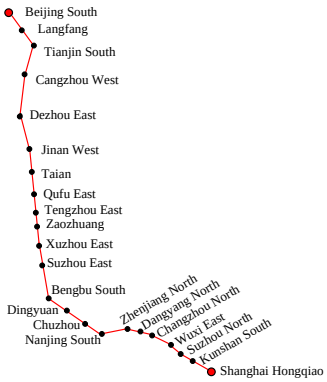


Figure 2: Schematic map of the Beijing-Shanghai high-speed railway (1318km)

Table 4: Number of trains in the cases*

Number of trains	**Case 2240**	**Case 2204**	**Case 0008**
type **A**	2	2	0
type **B**	2	2	0
type **C**	4	0	0
type **D**	0	4	8
Total	8	8	8

Notice*: names of the cases represent the number of different train types.

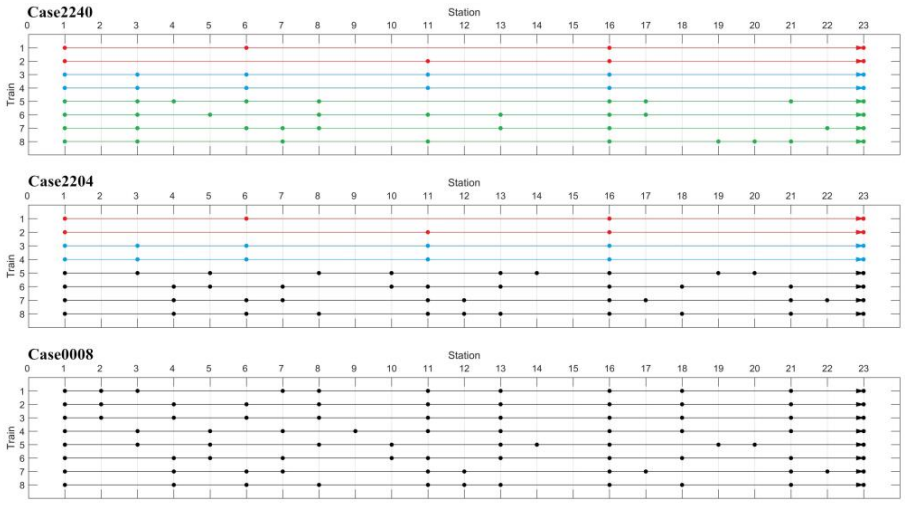


Figure 3: Stop plans of trains of type A, B, C and D in the cases
(red, blue, green and black lines for each train types, respectively)

Table 5: Abbreviation of the three strategies in the experiments

Abbreviation	Strategies
FT/IT	**F**lexible/g**i**ven **T**rain departure sequence at the origin stations
FO/IO	**F**lexible/forb**i**dden **O**vertakings at stations
FS/IS	**F**lexible/forb**i**dden new train **S**tops compared to the original line plan

5.1 Train sequence at origin stations and overtakings

In general, the strategies of train departure sequence at the origin stations and overtakings are widely used in timetabling, and both of them will not change the original line plan. Based on the given line plan, the impact of these two strategies on the capacity utilization and total train travel time are presented in Table 6. As expected, flexible train departure sequence and train overtakings lead to higher capacity utilization when train homogeneity is lower (i.e. in Case2240 and Case2204). However, too many overtakings always cause longer train travel time since trains have to wait at stations, and decrease the robustness of timetables because of the closer relationship of trains. If train homogeneity is high, i.e. in Case0008, train departure sequence will play a more important role than overtaking. In our opinion, finding a "good" train departure sequence at the origin stations is an important way to optimize capacity utilization of rail corridors, and corresponding "sacrifices" can be small. Hence, good train departure sequence and appropriate overtakings can be jointly considered since these two strategies can guarantee the quality of timetables with good capacity utilization, i.e. balance demand and supply.

Table 6: Experimental results of different strategies: train sequence and overtakings

No.	Train sequence	Over-taking	New stops	O1 (min)	"O1" +O3 (min)	Number of overtakings
Case 2240	FT	FO	IS	74	2750	12
	FT	IO	IS	110	2711	/
	IT	FO	IS	100	2734	8
	IT	IO	IS	194	2690	/
Case 2204	FT	FO	IS	86	2815	13
	FT	IO	IS	122	2763	/
	IT	FO	IS	104	2803	9
	IT	IO	IS	208	2752	/
Case 0008	FT	FO	IS	51	2990	0
	FT	IO	IS	51	2990	/
	IT	FO	IS	68	2988	0
	IT	IO	IS	68	2988	/

5.2 Overtakings and adding new stops

In practice, trains may have their "ideal departure time window" according to passenger demand or operation requirements. And new stops will be added at one station when one overtaking is needed at the station in practice sometimes. Therefore, it is necessary to analyse the impact of overtakings and new stops with given/fixed train departure sequence (see Table 7). In this section, train departure sequences are given beforehand and different from the results of the "flexible" sequence strategy. It is obvious that overtakings have more impact on the capacity utilization compared to adding new stops when train homogeneity is lower (i.e. in Case2240 and Case2204). And when overtakings are allowed, adding new stops will be better for the capacity utilization compared to the results with no overtakings. Further discussions of adding new stops are presented in Section 5.3. When train homogeneity is higher (i.e. in Case0008), the impact of overtakings are weaker, while the capacity utilization can be higher by adding new stops with longer train travel time. In our opinion, this may be the result of "balanced" stops of trains after adding new stops. For example, one train can be more "similar" (i.e. have the same stops at stations) to the neighbouring trains by adding new stops (e.g. in Figure 4).

Table 7: Experimental results of different strategies: overtakings and new stops

No.	Train sequence	Over-taking	New stops	O1 (min)	"O1" +O2 (min)	"O1+O2" +O3 (min)
Case 2240	IT	FO	FS	76	4	2793
	IT	FO	IS	100	/	2734
	IT	IO	FS	194	0	2690
	IT	IO	IS	194	/	2690
Case 2204	IT	FO	FS	74	5	2881
	IT	FO	IS	104	/	2803
	IT	IO	FS	208	0	2752
	IT	IO	IS	208	/	2752
Case 0008	IT	FO	FS	49	7	3103
	IT	FO	IS	68	/	2988
	IT	IO	FS	58	7	3101
	IT	IO	IS	68	/	2988

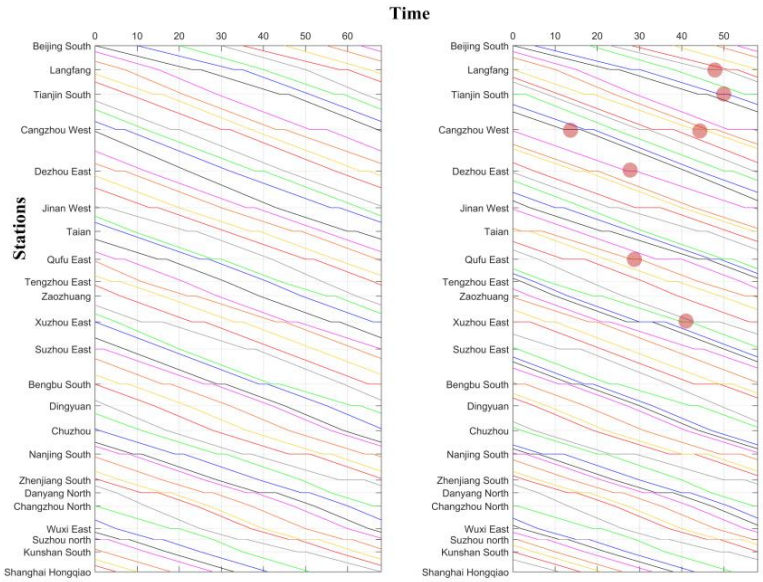


Figure 4: An example of timetables of adding new stops
(circles show the locations of new stops compared to the original timetable (left), different trains have different colors)

5.3 Trains offering the "fastest" transport service

In Section 5.2, we conclude that adding new stops have little impact on the capacity utilization when train homogeneity is lower. However, trains of type A are not allowed to be added new stops in the above cases since this kind of trains offer the "fastest" transport service (i.e. *highest train technical speed and least number of stops*). In this section, we relax this assumption in Case2240 to present the impressive impact of adding new stops for the "fastest" trains (see Table 8 and Figure 5). "*" means trains of type A are allowed to be added new stops and each train can be added at most one new stop. Obviously, capacity utilization can be higher if new stops of the "fastest" trains are allowed (i.e. FT-IO-FS* versus FT-IO-FS (yellow lines), and IT-IO-IS* versus IT-IO-FI (pink lines) in Figure 5), and better effectiveness of new stops are further presented. In order to obtain higher capacity utilization, new stops prefer to be added to the "fastest trains", i.e. trains of type A (the last volume in Table 8). When overtakings are allowed, the total number of overtakings of case* are less than that of the original case (i.e. FT-FO-FS* versus FT-FO-FS (purple nodes), and IT-FO-FS* versus IT-FO-FS (blue nodes)). In other words, overtakings can be more useful with adding new stops for the "fastest" trains.

Table 8: Experimental results of the "fastest" trains in Case2240

Train sequence	Over-taking	New stops	O1 (min)	"O1" +O2 (min)	"O1+O2" +O3 (min)	Number of new stops of different types of trains			
						A	B	C	D
FT	FO	IS	74	/	2750	/	/	/	/
FT	FO	FS	68	5	2798	0	1	4	0
FT	FO	FS*	60	6	2810	2	2	2	0
FT	IO	IS	110	/	2711	/	/	/	/
FT	IO	FS	108	2	2716	0	0	2	0
FT	IO	FS*	94	5	2756	2	2	1	0
IT	FO	IS	100	/	2734	/	/	/	/
IT	FO	FS	76	4	2793	0	2	2	0
IT	FO	FS*	70	6	2808	2	2	2	0
IT	IO	IS	194	/	2690	/	/	/	/
IT	IO	FS	194	0	2690	0	0	0	0
IT	IO	FS*	170	1	2708	1	0	0	0

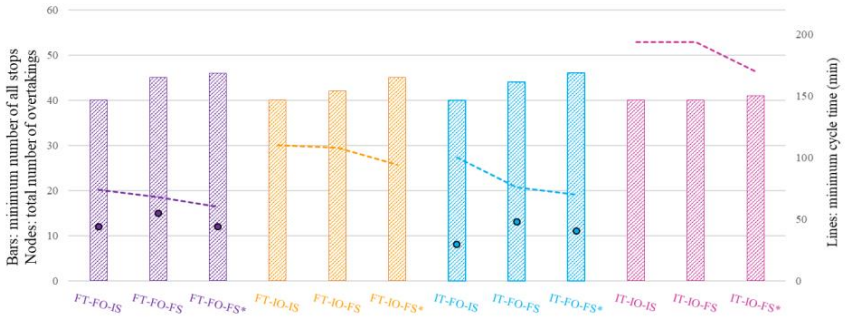


Figure 5: Impact of the "fastest" trains in Case2240

In sum, with flexible train departure sequence at the origin stations, the possibility of acquiring good capacity utilization can be higher and the impact on the quality of timetables can be little. Overtakings are very beneficial to the capacity utilization when train homogeneity is low, but train travel time and the robustness of timetables will be affected. Adding new stops changes the original line plans, and the impact on the capacity utilization depends on the usage of this strategy, i.e. which train is allowed to be added new stops and which stations can be selected to stop at. In our opinion, the capacity utilization and the service level of transportation should be balanced and jointly optimized by using the three timetabling strategies properly according to the characteristic of trains and passenger demand.

6 Conclusion

In this study, we propose a cyclic timetabling model based on the PESP with new variables which describe whether trains stop at the intermediate stations, and analyse the impact of the three timetabling strategies (i.e. train departure sequence at the origin stations, overtakings at stations and new train stops) on the capacity utilization by a series of experimental results. Flexible train departure sequence at the origin stations leads to higher possibility of acquiring good capacity utilization, and requires small sacrifices of the quality of timetables. For trains of low homogeneity, overtakings are also very beneficial to good capacity utilization. However, train travel time is always long and the robustness of timetables will be affected. The effectiveness of adding new stops depends on the ways of adding stops. Trains with higher technical speed and few stops should be mainly focused on, and integrating overtakings with new stops can be beneficial to the capacity utilization. Further research includes the analyses of the impact of the number of overtakings on the minimum cycle time.

Acknowledgments

This work was supported by the National Natural Science Foundation of China [No. U1434207], [No. 61703030], and the Ministry of science and technology of China [No. 2018YFB1201402]

References

Abril, M., Barber, F., Ingolotti, L., Salido, M.A., Tormos P., Lova, A., 2008. “An assessment of railway capacity”, *Transportation Research Part E: Logistics and Transportation Review*, vol. 44, pp. 774-806.

Abril, M., Salido, M.A., Barber, F., Ingolotti, L., Lova, A., Tormos, P., 2005. “A heuristic technique for the capacity assessment of periodic trains”, *Frontiers in Artificial Intelligence and Applications*, vol. 131, pp. 339–346.

Bergmann, D.R., 1975. “Integer programming formulation for deriving minimum dispatch intervals on a guideway accommodating through and local public transportation services”, *Transportation Planning Technology*, vol. 3, pp. 27–30.

Burdett, R.L., Kozan, E., 2006. “Techniques for absolute capacity determination in railways”, *Transportation Research Part B: Methodological*, vol. 40, pp. 616-632.

Cacchiani, V., Toth, P., 2012. “Nominal and robust train timetabling problems”, *European Journal of Operational Research*, vol. 219, pp. 727–737.
Caimi, G., Fuchsberger, M., Laumanns, M., Schüpbach, K., 2011. “Periodic railway timetabling with event flexibility”, *Networks*, vol. 57, pp. 3-18.
Caprara, A., Monaci, M., Toth, P., Guida, P.L., 2006. “A Lagrangian heuristic algorithm for a real-world train timetabling problem”, *Discrete Applied Mathematics*, vol. 154, pp. 738–753.
Cordone, R., Redaelli, F., 2011. “Optimizing the demand captured by a railway system with a regular timetable”, *Transportation Research Part B: Methodological*, vol. 45, pp. 430-446.
Dicembre, A., Ricci, S., 2011. “Railway traffic on high density urban corridors_ Capacity, signalling and timetable”, *Journal of Rail Transport Planning & Management*, vol. 1, pp. 59-68.
Goerigk, M., Schöbel, A., 2013. “Improving the modulo simplex algorithm for large-scale periodic timetabling”, *Computers & Operations Research*, vol. 40, pp.1363-1370.
Goverde, R.M.P., 2010. “A delay propagation algorithm for large-scale railway traffic networks”, *Transportation Research Part C: Emerging Technologies*, vol. 18, pp. 269-287.
Hansen, I., Pachl, J., 2008. *Railway Timetable & Traffic: Analysis-Modelling-Simulation*, Eurailpress, Hamburg, Germany.
Harrod, S.S., 2012. “A tutorial on fundamental model structures for railway timetable optimization”, *Surveys in Operations Research and Management Science*, vol. 17, pp. 85-96.
Heydar, M., Petering, M.E.H., Bergmann, D.R., 2013. “Mixed integer programming for minimizing the period of a cyclic railway timetable for a single track with two train types”, *Computers & Industrial Engineering*, vol. 66, pp. 171-185.
Kinder, M., 2008. *Models for periodic timetabling*, TechnischeUniversität, Berlin.
Kroon, L.G., Peeters, L.W.P., 2003. “A variable trip time model for cyclic railway timetabling”, *Transportation Science*, vol. 37, pp. 198-212.
Kroon, L.G., Peeters, L.W.P., Wagenaar, J.C., Zuidwijk, R., 2013. “Flexible connections in PESP models for cyclic passenger railway timetabling”, *Transportation Science*, vol. 48, pp. 136-154.
Lai, Y.C., Wang, S.H., 2012. “Development of analytical capacity models for conventional railways with advanced signaling systems”, *Journal of Transportation Engineering*, vol. 138, pp. 961-974.
Landex, A., 2008. *Methods to estimate railway capacity and passenger delays*, Technical University of Denmark (DTU), Denmark.
Liebchen, C., 2004. “Symmetry for periodic railway timetables”, *Electronic Notes in Theoretical Computer Science*, vol. 92, pp. 34-51.
Liebchen, C., 2006. “Periodic Timetable Optimization in Public Transport”. dissertation.de – Verlag im Internet, Berlin.
Liebchen, C., Möhring, R.H., 2007. “The Modeling Power of the Periodic Event Scheduling Problem: Railway Timetables - and Beyond”, In: *Algorithmic Approaches for Transportation Modeling, Optimization, and Systems*, pp. 3-40.
Liebchen, C., Schachtebeck, M., Schöbel, A., Stiller, S., Prigge, A., 2010. “Computing delay resistant railway timetables”, *Computers & Operations Research*, vol. 37, pp. 857-868.
Lindfeldt, O., 2011. “An analysis of double-track railway line capacity”, *Transportation Planning and Technology*, vol. 34, pp. 301-322.

Nachtigall, K., Opitz, J., 2008. “Solving periodic timetable optimisation problems by modulo simplex calculations”. In: M. Fischetti, & P. Widmayer (Eds.), *8th Workshop on Algorithmic Approaches for Transportation Modeling, Optimization, and Systems (ATMOS'08)*, vol. 9.

Peeters, L., 2003. *Cyclic railway timetable optimization*, Erasmus Research Institute of Management, Rotterdam.

Petering, M.E., Heydar, M., Bergmann, D.R., 2015. “Mixed-integer programming for railway capacity analysis and cyclic, combined train timetabling and platforming”, *Transportation Science*, vol. 50, pp. 892-909.

Salido, M.A., Barber, F., 2009. “Mathematical Solutions for Solving Periodic Railway Transportation”, *Mathematical Problems in Engineering*, pp. 1-19.

Schachtebeck, M., Schöbel, A., 2010. “To wait or not to wait-and who goes first? delay management with priority decisions”, *Transportation Science*, vol. 44, pp. 307-321.

Schmidt, M., Schöbel, A., 2015. “Timetabling with passenger routing”, *OR Spectrum*, vol. 37, pp. 75-97.

Schöbel, A., 2007. “Integer programming approaches for solving the delay management problem”, *Algorithmic Methods for Railway Optimization*, pp. 145-170.

Serafini, P., Ukovich, W., 1989. “A mathematical model for periodic scheduling problems”, *SIAM Journal on Discrete Mathematics*, vol. 2, pp. 550-581. http://dx.doi.org/10.1137/0402049.

Siebert, M., Goerigk, M., 2013. “An experimental comparison of periodic timetabling models”, *Computers & Operations Research*, vol. 40, pp. 2251-2259.

Sparing, D., Goverde, R., 2013. “An optimization model for periodic timetable generation with dynamic frequencies”, In: *IEEE16th International Conference on Intelligent Transportation Systems (ITSC)*, Netherlands, pp. 785-790.

Sparing, D., Goverde, R., 2017. “A cycle time optimization model for generating stable periodic railway timetables”, *Transportation Research Part B: Methodological*, vol. 98, pp. 198-223.

Xie, M.Q., Nie, L., 2009. “Model of cyclic train timetable”, *Journal of the China railway Society*, vol. 31, pp. 7-13.

Yan, F., Bešinović, N., Goverde, R., 2018. “Multi-objective periodic railway timetabling with overtaking optimization”, RAS 2016 Problem Solving Competition (PSC) Final Report, In: *INFORMS Annual Meeting 2018, Railway Application Section, Problem Solving Competition*, Nashville, USA.

Zhang, X., Nie, L., 2016. “Integrating capacity analysis with high-speed railway timetabling: A minimum cycle time calculation model with flexible overtaking constraints and intelligent enumeration”, *Transportation Research Part C: Emerging Technologies*, vol. 68, pp. 509-531.

Zhou, X.S., Zhong, M., 2007. “Single-track train timetabling with guaranteed optimality: Branch-and-bound algorithms with enhanced lower bounds”, *Transportation Research Part B: Methodological*, vol. 41, pp. 320-341.

Zhu, J.H., Liu, X., Zhang, W., 2009. “Effect of train stops on the hourly carrying capacity of intercity railway”, *China Railway Science*, pp. 108-112.

An Optimization Model for Rescheduling Trains to Serve Unpredicted Large Passenger Flow

Junduo Zhao[a,b,1], Haiying Li[a,2], Lingyun Meng[b,3], Francesco Corman[c]
[a] State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University
No.3 ShangYuanCun, HaiDian District, Beijing 100044, China
[b] School of Traffic and Transportation, Beijing Jiaotong University
No.3 ShangYuanCun, HaiDian District, Beijing 100044, China
[c] Institute for Transport Planning and Systems (IVT), ETH Zurich
Stefano-Franscini-Platz 5, 8093 Zurich, Switzerland
[1] Email: zhaojunduo@bjtu.edu.cn, Phone: (86)-13401153214
[2] Email: hyli@bjtu.edu.cn, Phone: (86)-10-51688063
[3] Email: lymeng@bjtu.edu.cn, Phone: (86)-10-51688520

Abstract
As the separation of vertically-integrated organizations in railway transportation, not only the competitive but also the collaboration between different operating companies and different modes should be considered emphatically in the rapidly changing multimodal transportation market. This paper tries to solve the Train Timetable Problem for serving Unpredicted Large Passenger Flow causing by the stop of air traffic in collaborating with air transportation companies. We address the Unpredicted Large Passenger Flow as a perturbation in normal train dispatching and solve this problem through an optimization approach. Two strategies of reassigning remaining seats and inserting new trains are adopted to establish integer programming model in dispatching to evacuate unpredicted passengers. The proposed model is solved by a standard CPLEX solver and test through a study case. The effectiveness of the proposed model is demonstrated in the study case and both two strategies take part in serving ULPF.

Keywords
Train rescheduling, Collaboration, Unpredicted large passenger flow, Inserting new trains

1 Introduction

Railway, with its capacity of transporting large passenger flow, plays an important role in the rapidly changing multimodal transportation market. However, the competitiveness of railway is receded sharply over the years. How to maintain and further improve the competitiveness are of great importance for railway companies and their operators. As the separation of vertically-integrated organizations of railway, train operating companies always concentrate on the competitiveness with others and other transportation modes. Nevertheless, not only the competitiveness but also the collaboration (i.e., complementation and connection) between different companies and transportation modes should be considered emphatically. While unavoidable perturbations (e.g. bad weather) disrupt airport causing stop of air transportation, a large amount of passengers are remained causing unpredicted transporting demand, which is called Unpredicted Large Passenger Flow (simply for ULPF) in this paper. The problem encountered by dispatchers of railway is how to rescheduled train timetables in collaborating with air transportation,

for the purpose of win-win situation.

Traditionally, a sequential process consisting of line planning, train timetabling, rolling stock and crew scheduling is used for planning train operations. The outcome of each stage is used as an input of the following stage (Desaulniers and Hickman, 2007). While a planned timetable is put into operation, unavoidable stochastic perturbations (e.g., bad weather, large passenger flow, capacity breakdowns) may influence the scheduled train running and dwelling times causing delays, thus the timetables need to be rescheduled to recover common(Luan et al., 2017). Always, passenger demand is an input of a line plan rather than other stages including rescheduling.

In this paper, we focus on generating an optimal dispatching solution for serving ULPF. We solve the Train Timetables Problem for serving Unpredicted Large Passenger Flow (TTP-ULPF) through an optimization approach to explicitly consider the characteristics of passengers from the stop of air transportation. We considered ULPF as a stochastic perturbation in the normal rescheduling, and two strategies of organizing remained seats and inserting new trains, are adopted to serve ULPF. The proposed integer programming model for formulating the TTP-ULPF problem is solving by a standard CPLEX solver.

The remainder of this paper is organized as follows. Section 2 provides a detailed literature review on relevant studies. In Section 3, the ULPF is described visually. In section 4, a mathematical model is proposed to reschedule timetables for serving ULPF with the statement of rail network and model assumption, followed by a case study in Section 5, which quantify the trade-off between the delay cost of existing passengers and the revenue of increasing new passengers. Finally, conclusions and further research are given in Section 6.

2 Literature Review

2.1 Train rescheduling

The train rescheduling problem has been studied in the past few decades. Carey and Lockwood (1995) presented a mixed integer programming model and solution algorithms for the train timetabling problem on a double-track rail line. Carey (1994a) further developed an extended model to consider more general and more complex rail networks with possible choices of lines and station platforms. A companion paper by Carey (1994b) proposed an extension from one-way to two-way rail lines. Caprara et al. (2002) proposed a graph-theoretic formulation for the periodic-timetabling problem using a directed multi-graph by incompatible arcs and forbid the simultaneous selection of such arcs through a novel concept of clique constraints. This formulation is used to derive an integer linear programming model that is relaxed in a Lagrangian way, which embedded within a heuristic algorithm that makes extensive use of the dual information associated with the Lagrangian multipliers. Depending on the basic problem of TTP, Caprara et al. (2006) proposed a mathematical model incorporating several additional constraints (e.g., Manual block signalling for managing, station capacities, prescribed timetable for a subset of the trains and Maintenance operations). Meng and Zhou (2014) develop an Integer Programming model for the problem of train dispatching on an N-track network by means of simultaneously rerouting and rescheduling trains. A vector of cumulative flow variables was introduced by them to reformulate the track occupancy so that they can decompose the original complex rerouting and rescheduling problem efficiently into a sequence of

single train optimization sub-problems. The decompose mechanism provide us a method to deal with large-scale optimal problems of train dispatching.

On the other hand, inserting new trains into existing timetables is a critical manner in rescheduling. Cacchiani et al. (2010) describe a problem for inserting new freight trains, which send requests for infrastructure usage, to existing passenger trains timetables. An Integer Linear Programming (ILP) model with the objective of total deviation between the actual timetable and the ideal one of all the freight trains is proposed, and solved by Lagrangian heuristic solution. It is a large-scale dispatching problem, since timetables should be rescheduled associating with new trains added. However, inserting new trains into existing timetables was used by Cacchiani et al. (2010) in the offline scheduling, while the capacities of network have not been used completely. The main goal of the study is to schedule the timetables of inserting train more close to the ideal ones, with the existing trains fixed. If we used this method to serve ULPF, it is an online scheduling, as all train timetables are on duty, and no train was fixed or has priority than others.

2.2 Railway transportation in multimodal market

Recently, the issue of competition between different operating companies received much attention in multimodal transportation market. Directive 91/440/EC (Commission of the European Communities, 1991) introduced separation of concerns between IM and TOCs. The IM holds a monopoly in the supply of access to its network and has the duty of providing fair and non-discriminatory access to the available infrastructure capacity. The TOCs are companies that compete to offer services to customers. Luan et al. (2017) focus on competition between different train operating companies. A Mixed Integer Linear Programming (MILP) model is proposed by Luan et al. (2017) to describe the trade-off between equity and delays in non-discriminatory train dispatching in multimodal transportation market. However, not only competition between different operating companies exists in the multimodal transportation market, but also the collaboration. Researchers pay more attention on competition, but less on collaboration, which reflect abilities (include stabilities and reliabilities) for the enhancement of competitiveness, as the research by Luan et al. (2017).

2.3 Paper contributions

There are three major contributions in this paper as followed:

(1) This paper focus on the train rescheduling problem with consideration of collaboration with air transportation, which is not found in previous studies to our best knowledge. It makes a step forward to perfect rescheduling trains in multimodal transportation market. It provides a model for cooperation between different transportation modes.

(2) This paper develops an ILP model considering jointly the balance of delays of existing passengers and revenue of unpredicted new passengers in the emergency situation, which are studied separately in previous researches. Thus, the trade-off between the above two represents one important contribution of this paper.

(3) In addition, the rescheduling planning generated by the model proposed in this paper, can give a supplement for existing frame of research.

3 Problem Description

Before formulating the TTP-ULPF problem, we first explain the terms used in describing the ULPF in the following formulations.

In this paper, we address the optimization problem of rescheduling trains to serve ULPF, which comes into being with the characteristics of 1) nonstop between original and destination metropolises 2) having willing to pay high cost for short travel time. Therefore, high-speed railway is first and foremost considered in this paper to serve ULPF.

As the transportation mode (e.g. travel time, stop manners, etc.) of railway has significantly different from air traffic, not all the passengers from disrupted air transportation have willing to transfer to railway. In order to contact the willing of ULPF and dispatching manners, a concept of time interval is introduced to depict the relationship. The time interval in this paper is the gap between expected arrival time of ULPF at its destination and the actual arrival time. The relationship of time interval and passengers' willing to transfer can be observed through investigation, and is regard as a linear function for assumption in this paper. Table 1 list the relationship between the volume of passenger willing to transfer and time interval at its destination (maximum volume: 100). We assigned that all the passengers have willing to transfer while the train to serve them departure from origin at the time that passengers generated and do not stop at any intermediate stations. And the volume reduced with the addition of time interval linearly by 5% per minute as shown in Table 1.

Table 1: Relationship between passengers volume and time interval

Time interval (min)	Passengers volume
0	100
1	95
2	90
3	85
4	80
5	75

Two strategies can used to serve ULPF transferred from air transportation: 1) organizing the seats remained in the planned trains; 2) inserting a new train. Obviously, inserting a new train is not a feasible manner to serve ULPF in congested timetables. But while the ULPF generated, the time is too close for existing trains to have enough remained seats for serving ULPF. Therefore, both of the two strategies should be used to realize the goal in this paper. It is hard to insert a new train in an existing timetable, since the timetables of some lines are too dense that there is no interspace between any of two trains to insert without changing their prescribed arrival/departure time.

The solution of inserting a new train in the congested timetables is to use the recovery time in the running and dwelling time of a train and the buffer time between two trains in the existing timetables. Fig.1 depicts a simple timetable with 3 stations and 2 segments. Three trains operate from station A to station C in the existing timetables in the Fig.1(a). It is easy to see that, both the running and dwelling time of existing trains and the headway between any of two consecutive trains are reduced to the limited value (e.g. 5min, 1min and 3min) to obtain time gap to insert a new train as illustrate in Fig.1(b). This strategy explores the trade-off between the revenue of inserting a new train and delay cost of existing trains at a part of the intermediate stations.

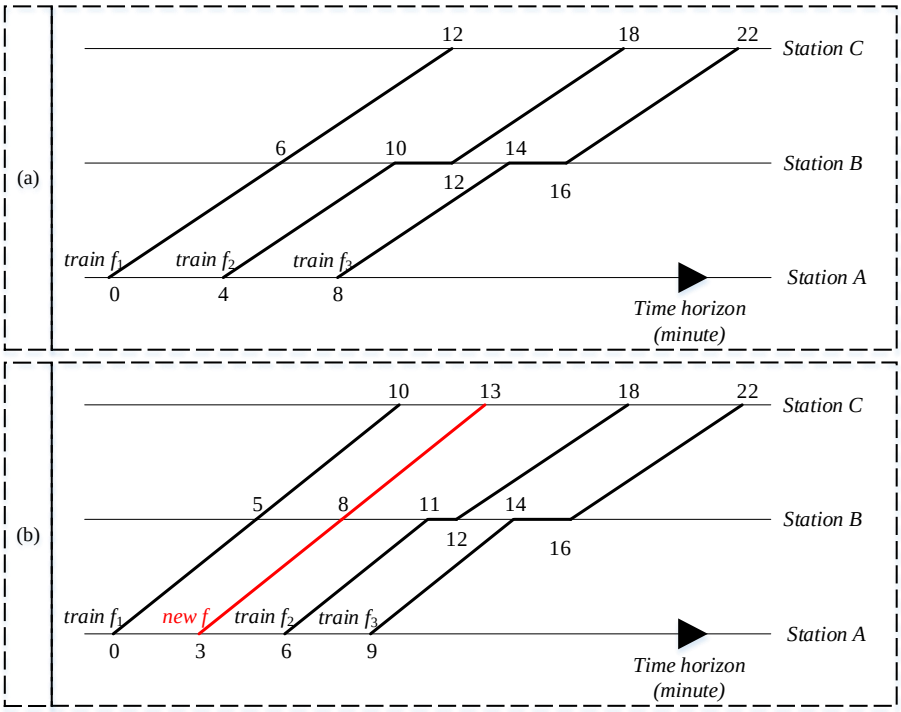


Figure 1: A sample of timetables

4 Mathematical Formulation

4.1 Description of railway network

In this paper, we focus on a simple railway network with only one line that consist of a sequence of station and double track segments between two consecutive station. Fig.2 and Fig.3 illustrates two networks for instance at microscopic level and modelling level considered in this paper respectively. In Fig.2, the railway network is consist of double track, signal and platform. The segment between two stations is divided into several block sections for the purpose of train safety. The station is also regard as two or more block sections according to the numbers of siding tracks.

The network in Fig.2 can be further simplified as shown in Fig.3, which the railway network is described as $G = (N, E)$ with a set of nodes N and a set of cells E. In order to explain the space-time network, two concepts should be introduced in this paper, i.e. node and cell. A cell represents a block section, and a node represents a beginning/ending point of block section. A station is regard as a node for simplicity, since the routing in the station make no difference to the objective and the capacity of station is assumed as sufficient in this paper. Therefore, two set of nodes are defined in our problem: a station node represents a station in physical network where trains can stop for loading/unloading and crossing which is shown as big dot in Fig.3; a segment node represents the point between two adjacent block sections where trains cannot stop which is shown as little dot in Fig.3. A cell is a vector directed from a starting node i to an ending node j, as well as

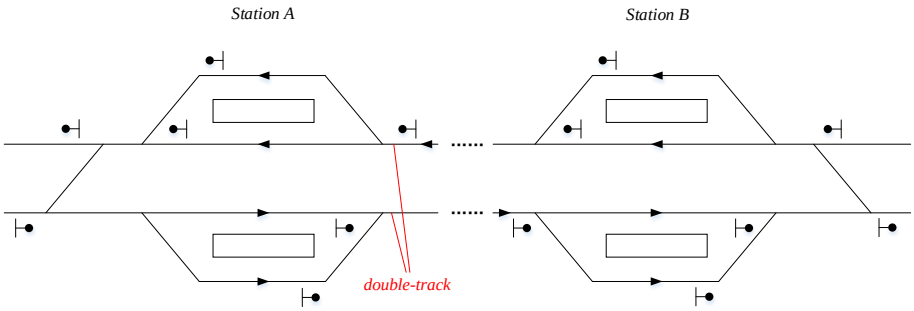


Figure 2: Railway physical network

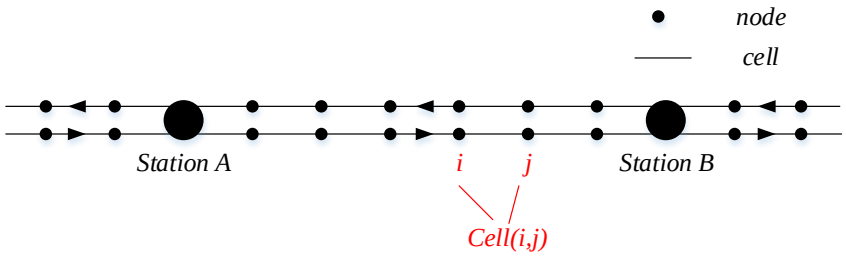


Figure 3: Modelling network

the minimum running unit for a train. The default of cell capacity in this paper is one at any given time, so that any of two trains cannot occupy one cell simultaneously.

4.2 Problem statement

In this optimization problem, the external inputs include:

(1) A high-speed railway (HSR) line given with stations and segments. Stations are simplified to a number of nodes, and the double-track segments are modelled as a sequence of directional cells, as illustrate in Fig.3.

(2) A set of existing trains with their origins, destinations, prescribed arrival and departure time at each cells, free flow running time at each segments, minimum dwelling time at each stations, loading quantity of passengers at each stations, and remaining seats between different origin and destination (OD) pairs.

(3) A set of candidate trains for inserting with their origins, destinations, earliest departure time at original station, free flow running time at each cells, minimum dwelling time at each stations, and capacity for transporting passengers.

(4) A set of ULPF with their origin and destination (OD), expected departure and arrival time at OD stations, and quantity of passengers.

The models proposed in this paper result in determining the arrival/departure time and train orders at each cell of all the trains, include new inserting train. Note that the granularity of time is one minute.

Six major assumptions are considered in the following formulations:

(1) A station is assumed as a node in this paper, since the routing and capacity of the station is not considered.

(2) The length of a train is assumed to be zero.

(3) Passengers' transfer in the intermediate station is not considered in this paper,

which means passengers can only take direct trains from origin to destinations.

(4) The value of 1)volume of ULPF 2)remaining seats in the existing trains 3)numbers of loading passengers at each station are all known before rescheduled.

(5) In the process of serving ULPF, other disruptions are not occurred for simplicity.

(6) We assumed that all the ULPF have the same origin and destination (OD), and cannot be divided furthermore.

4.3 Notation

Table 2-4 list the subscripts, input parameters and decision variables respectively.

Table 2: Subscripts

Symbol	Description
i, j, k	Node index, $i, j, k \in N$, N is the set of nodes, $N = N_s \cup N_r$, N_s is the set of station nodes and N_r is the set of segment nodes
e	Cell index, generated by two adjacent nodes i and j, $e = (i, j) \in E$, E is the set of cells
f	Train index, $f \in F$, F is the set of trains, $F = F_1 \cup F_2$, F_1 is the set of existing trains and F_2 is the set of candidate inserting trains
M	A sufficiently large positive number

Table 3 Input parameters

Symbol	Description
N_f	Set of station nodes train f need to stop for loading/unloading, $N_f \in N_s \in N$
E_f	Set of cells train f may use, $E_f \in E$
$w_f^{min}(i)$	Minimum dwell time for train f at station node i
$\vartheta_f(i,j)$	Free flow running time for train f to drive through the cell (i,j)
o_f	Origin node of train f
s_f	Destination node of train f
ε_f	Earliest departure time of train f from its origin node
ϵ_f	Latest arrival time of train f at its destination node
$\bar{a}_f(i,j)$	Predetermined arrival time of existing train f on cell (i,j), $f \in F_1$
$\bar{d}_f(i,j)$	Predetermined departure time of existing train f on cell (i,j), $f \in F_1$
o_p	Origin node of ULPF
s_p	Destination node of ULPF
ψ_o	Ideal departure time of ULPF from its origin node
ψ_d	Ideal arrival time of ULPF at its destination node
p	Passenger number of ULPF
$p_f(i,j)$	Remaining seats of train f between nodes i and j
$\overline{p}_f(i)$	Number of loading passengers on train f at node i
λ_a	Delay cost of each passenger on existing trains
λ_b	Loss cost of each passenger of ULPF failure to transfer to railway

Table 4 Decision variables

Symbol	Description
$x_f(i,j)$	0-1 binary routing variables, $x_f(i,j)=1$, if train f used cell (i,j) at some time, and otherwise $x_f(i,j)=0$
$a_f(i,j)$	Arrival time of train f on cell (i,j)
$d_f(i,j)$	Departure time of train f on cell (i,j)
$\theta(f,f',i,j)$	0-1 binary train ordering variables, $\theta(f,f',i,j)=1$, if train f' arrive at cell (i,j) after train f, and otherwise $\theta(f,f',i,j)=0$
$T_f(i,j)$	Running time for train f on cell (i,j)
$TT_f(i)$	Delay time of train f at station node i
$\delta_f(i,j)$	Passenger number of ULPF transport from origin node i to destination node j through train f

4.4 Mathematical model

A mathematical model, which formulizes inserting trains to serve ULPF by a set of constraints, is first presented. The objective is 1) to minimize the total delay costs of passengers in existing trains 2) and to maximize the revenues of increasing passengers transferred from ULPF simultaneously. Since the two objective is on the contrary, we transfer the revenues of increasing passengers transferred from ULPF to the loss costs of passengers failure to transfer from ULPF, as formulated in Eq.(1).

$$Min\ C=\sum_{f\in F_1}\sum_{i\in N_f\backslash\{o_f\}}\lambda_a\times\overline{p}_f(i)\times TT_f(i)+\sum_{f\in F}\lambda_b\times\left(p-\delta_f(o_p,s_p)\right) \quad (1)$$

Subject to

Group I: Flow balance constraints

Flow balance constraints at origin node:

$$\sum_{j:(o_f,j)\in E_f}x_f(o_f,j)=1,\forall f\in F_1 \quad (2)$$

$$\sum_{j:(o_f,j)\in E_f}x_f(o_f,j)\le 1,\forall f\in F_2 \quad (3)$$

Flow balance constraints at intermediate nodes:

$$\sum_{i:(i,j)\in E_f}x_f(i,j)=\sum_{k:(j,k)\in E_f}x_f(j,k),\forall f\in F,j\in N\backslash\{o_f,s_f\} \quad (4)$$

Flow balance constraints at destination node:

$$\sum_{i:(i,s_f)\in E_f}x_f(i,s_f)=1,\forall f\in F_1 \quad (5)$$

$$\sum_{i:(i,s_f)\in E_f}x_f(i,s_f)\le 1,\forall f\in F_2 \quad (6)$$

Group II: Running and dwelling time constraints

Running time constraints:

$$T_f(i,j)=d_f(i,j)-a_f(i,j),\forall f\in F,(i,j)\in E_f \quad (7)$$

Minimum running time constraints:

$$T_f(i,j)\ge x_f(i,j)\times\vartheta_f(i,j),\forall f\in F,(i,j)\in E_f \quad (8)$$

Minimum dwelling time constraints:

$$d_f(i,j) + w_f^{min}(j) \le a_f(j,k), \forall f \in F, j \in N_f \backslash \{s_f\}, (i,j) \in E_f, (j,k) \in E_f \quad (9)$$

Group III: Time-space network constraints

Starting time constraints at origin node:

$$a_f(o_f,j) + \left(1 - x_f(o_f,j)\right) \times M \ge \varepsilon_f, \forall f \in F, (o_f,j) \in E_f \quad (10)$$

Ending time constraints at destination node:

$$d_f(i,s_f) + \left(1 - x_f(i,s_f)\right) \times M \le \epsilon_f, \forall f \in F, (i,s_f) \in E_f \quad (11)$$

Departure time constraints at intermediate node:

$$a_f(i,j) \ge \overline{a}_f(i,j), \forall f \in F_1, i \in N_f, (i,j) \in E_f \quad (12)$$

Cell transition constraints:

$$d_f(i,j) \ge a_f(i,j), \forall f \in F, (i,j) \in E_f \quad (13)$$

Cell-to-cell transition constraints at station nodes:

$$\sum_{i:(i,j)\in E_f} d_f(i,j) \le \sum_{k:(j,k)\in E_f} a_f(j,k), \forall f \in F, j \in N_f \quad (14)$$

Cell-to-cell transition constraints at segment nodes:

$$\sum_{i:(i,j)\in E_f} d_f(i,j) = \sum_{k:(j,k)\in E_f} a_f(j,k), \forall f \in F, j \in N \backslash N_f \quad (15)$$

Mapping constraints between time-space network and physical network:

$$x_f(i,j) - 1 \le a_f(i,j) \le x_f(i,j) \times M, \forall f \in F, (i,j) \in E_f \quad (16)$$

$$x_f(i,j) - 1 \le d_f(i,j) \le x_f(i,j) \times M, \forall f \in F, (i,j) \in E_f \quad (17)$$

Group IV: Inserting trains constraints

Starting time constraints of inserting trains to serve ULPF:

$$a_f(o_f,j) + \left(1 - x_f(o_f,j)\right) \times M \ge \psi_o, \forall f \in F_2, (o_f,j) \in E_f \quad (18)$$

Number constraints of inserting trains

$$\sum_{f\in F_2} \sum_{j:(o_f,j)\in E_f} x_f(o_f,j) \le 1 \quad (19)$$

Group V: Mapping constraints between two types of decision variables

Mapping constraints between train orders and cell usage:

$$x_f(i,j) + x_{f'}(i,j) - 1 \le \theta(f,f',i,j) + \theta(f',f,i,j) \le 3 - x_f(i,j) - x_{f'}(i,j), \forall f \in F, f' \in F, f \ne f', (i,j) \in E_f \cap E_{f'} \quad (20)$$

$$\theta(f,f',i,j) \le x_f(i,j), \forall f \in F, f' \in F, f \ne f', (i,j) \in E_f \cap E_{f'} \quad (21)$$

$$\theta(f,f',i,j) \le x_{f'}(i,j), \forall f \in F, f' \in F, f \ne f', (i,j) \in E_f \cap E_{f'} \quad (22)$$

Mapping constraints between passengers transportation and cell usage

$$x_f(o_f,j) - 1 \le \delta_f(o_p,s_p) \le x_f(o_f,j) \times M, \forall f \in F_2, o_p \in N_f, s_p \in N_f, (o_f,j) \in E_f \quad (23)$$

Group VI: Capacity constraints on the same cell

$$a_{f'}(i,j) + \left(3 - x_f(i,j) - x_{f'}(i,j) - \theta(f,f',i,j)\right) \times M \ge d_f(i,j), \forall f \in F, f' \in F, f \ne f', (i,j) \in E_f \cap E_{f'} \quad (24)$$

Group VII: Delay time constraints

$$TT_f(j) \ge d_f(i,j) - \overline{d}_f(i,j), \forall f \in F, j \in N_f, (i,j) \in E_f \quad (25)$$

$$TT_f(j) \le \left|d_f(i,j) - \overline{d}_f(i,j)\right|, \forall f \in F, j \in N_f, (i,j) \in E_f \quad (26)$$

$$TT_f(j) \ge 0, \forall f \in F, j \in N_f, (i,j) \in E_f \quad (27)$$

Group VIII: ULPF constraints

Passenger volume constraints:

$$0 \le \delta_f(o_p,s_p) \le p, \forall f \in F, o_p \in N_f, s_p \in N_f \quad (28)$$

$$\delta_f(o_p,s_p) \le 0, \forall f \in F: \varepsilon_f < \psi_o, o_p \in N_f, s_p \in N_f \quad (29)$$

$$\delta_f(o_p, s_p) \le p \times \left(1 - 5\% \times \left(d_f(i, s_p) - \psi_d\right)\right), \forall f \in F, (i, s_p) \in E_f, o_p \in N_f, s_p \in N_f \quad (30)$$

$$\delta_f(o_p, s_p) \le p_f(o_p, s_p), \forall f \in F, o_p \in N_f, s_p \in N_f \quad (31)$$

$$\sum_{f \in F, o_p \in N_f, s_p \in N_f} \delta_f(o_p, s_p) \le p \quad (32)$$

In Group I, constraints (2)-(6) ensure the consistency of trains' movement in the network at their origin, intermediate and destination nodes respectively. Note that the flow of trains at their origin and destination nodes in Eq.(3) and Eq.(6) is not identical equal to one, as not all the trains in candidate inserting set need to put into operation necessarily.

In Group II, constraint (7) defines the required running time on cells. Constraints (8) and (9) force the minimum running time on cells and minimum dwelling time at station nodes respectively.

In Group III, constraints (10) and (11) guarantee that trains do not leave their origin nodes before earliest departure time and not reach their destination nodes after latest arrival time respectively. Constraint (12) make sure that existing trains do not leave intermediate station nodes before the prescribed departure time, so as the passengers predetermined can boarding successfully. Constraints (13) and (14) enforce the sequential time orders between departure time and arrival time on the cells and at the station nodes respectively. Constraint (15) further makes sure that all trains cannot stop at segment nodes. Constraints (16) and (17) are imposed to map the arrival and departure time in time-space network to the cell usage variables in physical network, so as to describe the relationship between cells selection of a train and its timetables.

In Group IV, constraint (18) further guarantees that the departure time of inserting trains cannot be early than ideal departure time of ULPF at their origin nodes, so as the strategy of inserting is effective for serving ULPF. Constraint (19) denotes the total quantity of inserting trains.

In Group V, constraints (20)-(22) link train orders variables and cell usage variables. Additionally, if and only if both train f and train f' use cell (i, j), the two trains have the sequential order $\theta(f, f', i, j) = 1$ or $\theta(f', f, i, j) = 1$, that is either train f arrives at cell after train f' or the opposite condition. If $x_f(i, j) = 0, x_{f'}(i, j) = 1$ or $x_f(i, j) = 1$, $x_{f'}(i, j) = 0$ or $x_f(i, j) = 0$, $x_{f'}(i, j) = 0$, constraints (20) reduce to non-active inequalities. Constrain (23) link passengers' transportation variables and cell usage variables. If and only if the inserting train f use cell (i, j), i.e., $x_f(i, j) = 1$, the number of passengers served by train f is greater than or equal to zero, else $\delta_f(o_p, s_p) = 0$.

In Group VI, constraint (24) explicitly makes sure that any of two trains cannot occupy the same cell simultaneously at any given time. Note that for train f and f' traversing on cell (i, j), i.e., $x_f(i, j) = x_{f'}(i, j) = 1$, constraint (24) can be reduced to common if-then conditions as follow: If train f arrives at cell after train f', i.e., $\theta(f, f', i, j) = 1$, then the arrival time of train f' should be no earlier than the departure time of train f on cell (i, j); else the constraint reduce to non-active inequality.

In Group VII, constraints (25)-(27) define the delay time of existing trains at each station nodes. If train f arrive at the station node j before its predetermined time point, i.e., $d_f(i, j) \le \overline{d}_f(i, j)$, then the delay time $TT_f(j) = 0$; else delay time $TT_f(j) = d_f(i, j) - \overline{d}_f(i, j)$.

In Group VIII, constraints (28) and (32) restrict range of the numbers of passengers from ULPF. Constraint (29) expresses that an existing train with its departure time earlier than the ideal departure time of ULPF cannot be used to serveULPF. Constraint (30)

indicate the function relationship between arrival time and numbers of served passengers of each train f, in which the reduction is assumed as 5% per minute for simplicity. Constrain (31) make sure that train f has enough seats for serving ULPF.

5 Numerical experiments

5.1 Experimental setup

The optimization model, which proposed for serving Unpredicted Large Passengers Flow (ULPF), is implemented as an integer programming model through a commercial solver ILOG CPLEX by IBM with version number 12.3. All the following experiments are performed on a Lenovo PC with 2.3GHz Intel i5-6200U CPU and 8 GB memory.

Due to the protection of commercial data, we couldn't get detailed block sections data of the Beijing-Shanghai high-speed railway (HSR) line. Therefore, we use an assumed line with the background of Beijing-Shanghai HSR line as the test bed. The total length of this line is 180 km with 4 stations and 30 block sections, as illustrate in Fig.4. There are totally 9 trains being dispatched in our case study, including 6 existing trains and 3 candidate inserting trains. The time horizon of this case is 52 minutes. The minimum running time in the segments and the minimum dwelling time at stations are 12 min and 1 min respectively. The OD station of ULPF are Sta_A and Sta_D respectively, and the quantity of ULPF is assumed as 1000 in this paper, which is the same as the capacity of a train.

Table 5: Timetables of existing trains (unit: min)

Train ID	Station A	Station B		Station C		Station D
	Departure time	Arrival time	Departure time	Arrival time	Departure time	Arrival time
1	0	12	14	26	28	41
2	4	16	16	28	30	43
3	6	18	20	32	32	45
4	10	22	22	34	34	47
5	12	24	24	36	36	49
6	14	26	26	38	38	51

Table 6 Loading volume of existing passengers

Train ID	Station A	Station B	Station C	Station D
1	——	200	400	900
2	——	——	400	900
3	——	400	——	900
4	——	——	——	900
5	——	——	——	900
6	——	——	——	900

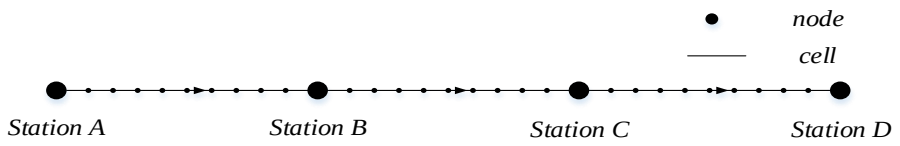


Figure 4: A sample network for case study

Table 7 Remaining seats between each OD pair

Train ID		Station B	Station C	Station D
1	Station A	50	50	100
	Station B	——	50	50
	Station C	——	——	50
2	Station A	——	50	100
	Station B	——	——	——
	Station C	——	——	50
3	Station A	50	——	100
	Station B	——	——	50
	Station C	——	——	——
4	Station A	——	——	100
	Station B	——	——	——
	Station C	——	——	——
5	Station A	——	——	100
	Station B	——	——	——
	Station C	——	——	——
6	Station A	——	——	100
	Station B	——	——	——
	Station C	——	——	——

The detailed information of arrival/departure time of existing trains is shown in Table 5, and the detailed information of loading passenger number at each station and remaining seats between each OD pair are illustrated in Table 6 and 7.

5.2 Experimental results

There are 6 existing trains and 3 inserting trains calculating in the study case. The number of variables and constraints are 2199 and 6406 respectively. The computational time is about 5.87 seconds on the platform stated above. The result of this case study is illustrated in Table 8, Table 9 and Fig.5.

As list in Table 8, one of the trains in candidate set, i.e., train ID9, is inserted from station A to station D to serve ULPF with the minimum travel time of 36min and the mode of nonstop at intermediate stations. The timetables before and after inserting are depicted intuitively in Fig.5 for convenient exhibition.

Due to the inserting of new train ID9, all of the existing trains are affected at their intermediate stations and/or destination stations. Owing much to the recovery time predetermined in the running and dwelling time of existing trains, all the passengers on train 3-6 are not affected at their destinations. However, a part of passengers on train 1 and 2 are not so lucky due to the delay of trains with totally about 1600 person-time as a trade-off for inserting new train. Due to the shorter travel time of inserting train, all the

passengers causing by stop of air transportation are willing to transfer to this train to their destinations, as illustrated in Table 9. The reason why passengers do not take the existing trains is not uniform. Train 1 do not take part in serving ULPF only because of its earlier departure time than the ideal time of ULPF. And the reason of train 2-6 is that their arrival time at destination are later than the inserting new train.

Table 8 Computational result of all trains (unit: min)

Train ID	Inserting or not (I/N)	Station A	Station B		Station C		Station D
		Departure time	Arrival time	Departure time	Arrival time	Departure time	Arrival time
1	——	0	12	16	28	29	41
2	——	4	16	18	30	31	43
3	——	6	18	20	32	33	45
4	——	10	22	22	34	35	47
5	——	12	24	24	36	37	49
6	——	14	26	26	38	39	51
7	N	——	——	——	——	——	——
8	N	——	——	——	——	——	——
9	Y	2	14	14	26	26	38

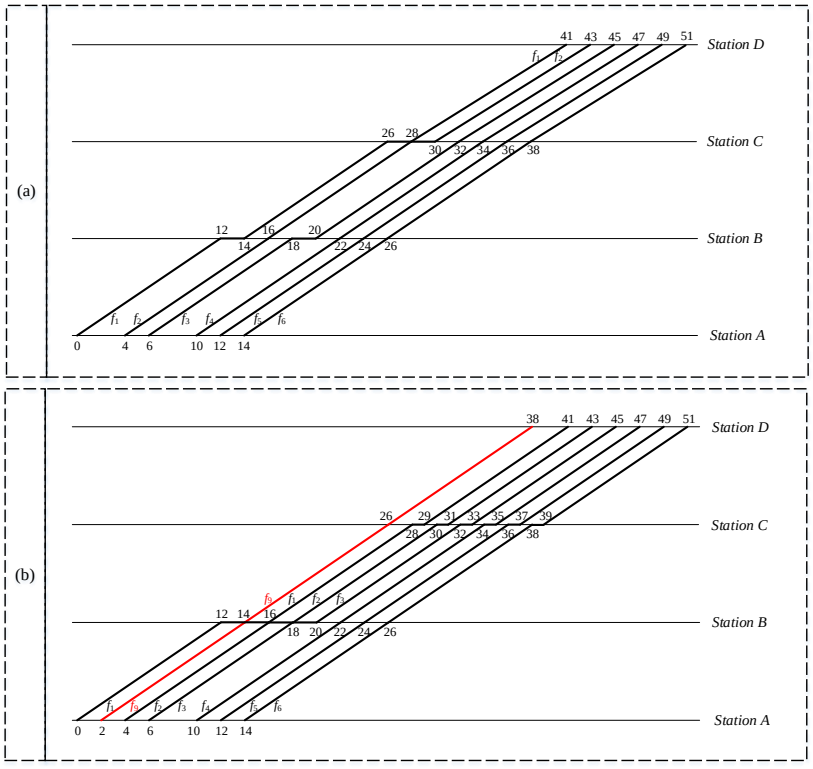


Figure 5: Timetables before and after inserting

Table 9 Computational result of passenger from ULPF transported in each train

Train ID	1	2	3	4	5	6	7	8	9
Passengers volume	0	0	0	0	0	0	—	—	1000

6 Conclusion and Future Research

This paper concentrates on solving the TTP for serving ULPF causing by the stop of air traffic in collaborating with air transportation companies. We address this problem through an optimization approach to explicitly consider ULPF as a stochastic perturbation in normal dispatching. Two strategies of organizing remaining seats and inserting new trains are adopted to formulate the integer programming model to serve ULPF. The proposed model is solved by a standard CPLEX solver and test through a study case. The effectiveness of the proposed model is demonstrated in the study case and both two strategies take part in serving ULPF.

Our future research would address the following main extensions.

(1) Station capacity and routing are not considered in this paper. Our future research is to develop a model incorporating these constraints.

(2) We assume that passengers take a direct train to arrive at destinations in the proposed model. The next step, we are going to relax this assumption and to consider the transfer of passengers.

(3) In this paper, the schedule of rolling stock is ignored for simplicity. One may take rolling stock schedule, even the crew schedule into account to better represent the realistic conditions.

(4) In the computational experiments, we use a line with only 4 station and 3 segments. In the future study, one can enlarge the scale of the network and solve the model using heuristic algorithm.

(5) The weight of delay cost of existing passengers and increased revenue of ULPF is assumed in this paper. One can concentrate on the study of influence factors of the weight and calculate the value more close to realistic condition for the further study.

Acknowledgements

This work is supported by National Key R&D Plan (No. 2018YFB1201403).

References

Cacchiani, V., Caprara, A., Toth, P. 2010. *Scheduling extra freight trains on railway networks*. Transportation research Part B: Methodological, 44, 215-231.

Caimi. G., Kroon. L., Liebchen. C., 2017. *Models for railway timetable optimization: Applicability and applications in practice*. Journal of Rail Transport Planning & Management. 6, 285-312.

Caprara, A., Fischetti, M., Toth, P., 2002. *Modeling and solving the train timetabling problem*. Oper. Res. 50 (5), 851–861.

Caprara, A., Monaci, M., Toth, P., Guida, P.L., 2006. *A Lagrangian heuristic algorithm for a real-world train timetabling problem*. Discrete Appl. Math. 154 (5), 738–753.

Carey, M., Lockwood, D., 1995. *A model, algorithms and strategy for train pathing*. J. Oper. Res. Soc. 46 (8), 988–1005.

Carey, M., 1994a. *A model and strategy for train pathing with choice of lines, platforms, and routes*. Transp. Res. Part B 28 (5), 333–353.

Carey, M., 1994b. *Extending a train pathing model from one-way to two-way track*. Transp. Res. Part B 28 (5), 395–400.

Commission of the European Communities, 1991. *Council Directive 91/440/EEC on the development of the Community's railways*.

Desaulniers, G., Hickman, M., 2007. *Public transit. In: Laporte, G., Barnhart, C., (eds) Transportation*. Handbooks in operations research and management science, 14: 69-127.

Klabes, S.G., 2010. *Algorithmic railway capacity allocation in a competitive European railway market. Fakultat fur Bauingenieurwesen der Rheinisch-Westfalischen Technischen Hochschule Aachen zur Erlangung des akademischen Grades eines Doktors der Ingenieurwissenschaften genehmigte Dissertation*. Manuscript. 2010. 209 p.

Kurosaki, F., 2008. *An analysis of vertical separation of railways Doctoral dissertation*. University of Leeds, NY.

Luan, X., Corman, F., Meng, L. 2017. *Non-discriminatory train dispatching in a rail transport market with multiple competing and collaborative train operating companies*. Transportation Research Part C: Emerging Technology, 80, 148-174.

Meng, L., Zhou, X. 2014. *Simultaneous train rerouting and rescheduling on an N-track network: A model reformulation with network-based cumulative flow variables*. Transportation research Part B: Methodological, 67, 208-234.

Long Short-term Memory Neural Network for Short-term High-speed Rail Passenger Flow Forecasting

Yangyang Zhao [a], Xinguo Jiang [a,1]
[a] Department of transportation and logistics, Southwest Jiaotong University
National United Engineering Laboratory of Integrated and Intelligent Transportation, West Park, High-Tech District, Chengdu, China 611756
[1] E-mail: ejiang@gmail.com, Phone: +86 15229308543

Abstract:
The uncertainty of estimating the railway passenger flow in advance may disrupt the passenger operation and management (e.g., passenger evacuation planning, seat allocation, and train timetable programming). In order to proactively improve the service quality and efficiency of the railway system, the short-term passenger flow prediction technique is vital in the field of operation and management system. Utilizing the deep learning library-Keras, the study develops a long short-term memory neural network (LSTM NN) to predict the short-term high-speed rail (HSR) passenger flow. Processing the raw data, we first construct the passenger flow sequences as the input (output) variables. Then the gird search and cross validation techniques are applied to optimize the LSTM NN parameters. At last we utilize the data provided by Shanghai railway administration of China as the case study. Through a comparison with other representative methods, including Auto-Regressive Integrated Moving Average (ARIMA), Back Propagation Neural Network (BPNN), and Support Vector Machine Regression (SVR), results suggest that the proposed LSTM NN can generate great potentials for accurate passenger flow predictions.

Keywords:
Short-term passenger flow prediction, High-speed rail, Long short-term memory neural network, Grid search, Cross validation

1 Introduction

High-speed rail (HSR), as a high-quality inter-city transportation mode, is developing rapidly in many countries. For example, in China HSR has become more and more popular with travelers and can effectively relieve the pressure of transporting passengers among the major metropolises. For railway operators, short-term forecasting is closely related to revenue management and service quality. Demand information provided by short-term forecasting can be used as inputs for other systems such as passenger evacuation, seat location, pricing, and train timetable programming. Thus, accurate prediction of short-term passenger flow is significant in the railway operational decision-making and dynamic operation adjustment. In the past decades, numerous short-term traffic flow prediction models have been proposed in the

transportation systems (e.g., freeway, railway, bus, and metro). These models can be generally categorized into parametric and nonparametric ones.

Parametric models, in particular smoothing techniques (Williams, 1998), grey forecasting model (Hsu and Wen, 1998; Fang and Wu, 2006), state space model (Anthony and Karlaftis, 2003), and autoregressive integrated moving average (ARIMA) (Hamed et al., 1995; Lee and Fambro, 1999) have been used extensively. Especially, ARIMA has aroused wide interest since 1970s due to its effectiveness in modeling linear and stationary time series. Utilizing the 20-sec (30-sec and 60-sec) traffic flow data, Ahmed and Cook (1979) showed that ARIMA outperformed moving-average, double-exponential smoothing, and exponential smoothing with adaptive response. Lee and Fambro (1999) revealed that the subset ARIMA provided more stable and accurate predictions than full ARIMA through 5-min traffic flow prediction. What's more, with the 15-min traffic flow data, a comparison conducted by Williams and Hoel (2003) between the nearest-neighbor (neural network and historical average model) and seasonal ARIMA favored the seasonal ARIMA. However, ARIMA assumed a linear correlation among time series data and might not address the nonlinearity issue inherent in the traffic flow; comparatively, nonparametric techniques could deal with the nonlinearity and were expected to achieve more accurate predictions. Generally, ARIMA is compared as a benchmarking method to the newly proposed nonparametric models.

For the nonparametric models, much more work has been done such as Bayesian network (Zheng et al., 2006), Kalman filtering (Qkutani and Stephanedes, 1984), Support Vector Machine Regression (SVR) (Manoel et al., 2009), K-nearest Neighbor Model (Zhang et al., 2013), the probability tree (Leng et al., 2013), and the random forest (Kecman and Goverde, 2015). Regarding the short-term prediction of passenger flow only, Wei and Chen (2012) combined empirical mode decomposition (EMD) with back-propagation neural network (BPNN) to predict the 15-min metro passenger flow. Sun et al. (2015) integrated Wavelet with SVM to forecast Beijing subway ridership particularly in the rush hours. Under special events scenarios, Li et al. (2017) developed multiscale radial basis function networks to predict the 15-min metro passenger flow. Additionally, gradient boosting decision trees (Ding et al., 2016) was also applied to predict 15-min subway ridership and identify the relative influences of the independent predictor input variables. In addition, to predict railway passenger flow in a day, Tsai et al. (2009) constructed a multiple temporal unit neural network and a parallel ensemble neural network, and Jiang et al. (2014) devised a hybrid model which integrated ensemble EMD with grey support vector machine (GSVM).

Among the nonparametric models, neural networks have drawn the greatest attentions for its mapping capabilities. As a subset of neural network, recently deep learning has been applied with success in many fields, such as dimensionality reduction (Hinton and Salakhutdinov, 2006), natural language processing (Collobert and Weston, 2008), object detection (Goodfellow et al., 2013), and classification tasks (LeCun et al., 2015). Therefore, it inspires us to combine the short-term prediction with the deep architecture models. However, currently most researchers concentrated on road traffic (Liu and Chen, 2017; Bai et al., 2017; Polson and Sokolov, 2017; Mackenzie et al., 2018) and limited attention has been paid to railway passenger flow. The paper develops a LSTM NN to predict the short-term HSR passenger flow and the effectiveness of the proposed LSTM NN is validated through a comparison with ARIMA, BP NN, and SVR.

2 Methodology

2.1 Long short-term memory neural network

LSTM NN was originally introduced by Hochreiter and Schmidhuber (1970) and improved by Gers at al. (2000). The primary objective of LSTM NN is to model long-term dependencies. A LSTM NN is composed of one input layer, one recurrent hidden layer and one output layer. Different from the traditional NN, the basic unit of the hidden layer is memory block (Abigogun, 2005). The memory block contains a memory cell, an input gate, an output gate, and a forget gate. The cell is responsible for transporting values over arbitrary time intervals and memorizing the temporal state. Three gates can be treated as "conventional" artificial neurons, similar to those in a feedforward neural network (i.e., the input gate and output gate control the input and output activations into the block, the forget gate selects the partial output from the upper memory block to prevent the cell values growing without bound). Through the multiplicative gates, LSTM memory cells can store and access information during the long periods of time, thus mitigating the vanishing gradient problem. The above procedure is shown in Figure 1.

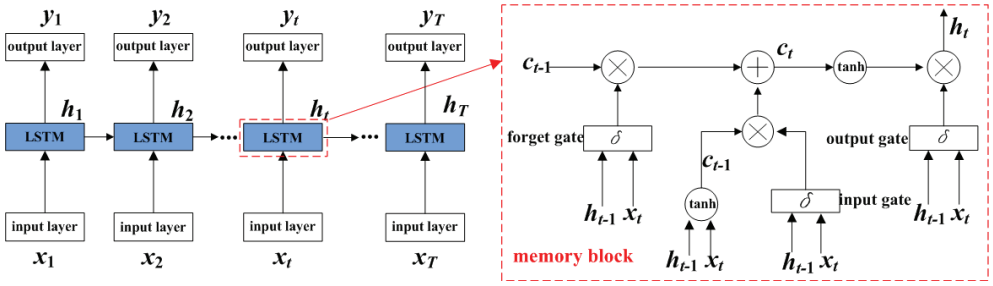


Figure 1: LSTM NN architecture

Given the model input $x = (x_1, x_2, \cdots, x_T)$, the output $y = (y_1, y_2, \cdots, y_T)$, and the hidden output $h = (h_1, h_2, \cdots, h_T)$, where T represents the prediction period. In the context of short-term passenger flow prediction, x can be considered as historical input data (e.g., time of day, weather condition, passenger flow), and y is the estimated passenger flow. The predicted passenger flow will be iteratively calculated by equations (1)-(8):

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i) \quad (1)$$
$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f) \quad (2)$$
$$c_t = f_t * c_{t-1} + i_t * \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \quad (3)$$
$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o) \quad (4)$$
$$h_t = o_t * \tanh(c_t) \quad (5)$$
$$y_t = \sigma(W_{hy}h_t + b_y) \quad (6)$$

where t is the order of observation time interval during T, W terms are weight matrices (e.g., W_{xi} is the matrix of weights from the input gate i to the input x), the b terms are bias vectors

(b_i is the input gate *i* bias vector), and *i*, *f*, *o*, *c* represent the input gate, forget gate, output gate, and cell activation vectors, respectively. $\delta(\cdot)$ is the standard logistic sigmoid function:

$$\sigma(x) = \frac{1}{1+e^{-x}} \tag{7}$$

The square error is used as the loss function as follows:

$$e(t) = \sum(y_t - \bar{y}_t)^2 \tag{8}$$

where y_t, $\bar{y}_t$ is the observed and predicted passenger flow, respectively.

The truncated Back Propagation Through Time (BPTT) is widely used to train LSTM NN. Due to extensive mathematical derivations, the detailed steps are not covered in this section and readers may refer to (Gers, 2001) for more information.

2.2 Other representative models

We employ three typical forecasting models: ARIMA, BPNN, and SVR to test the performance of the LSTM NN.

ARIMA

ARIMA models are linear estimators regress on past values of the modeled time series (the autoregressive terms) or past prediction errors (the moving average terms), and are also written as ARIMA (*p*, *d*, *q*), where *p* is the number of autoregressive terms, *d* is the number of order, and *q* is the moving average parameter.

BPNN

BPNN is a kind of artificial neural network which adopts a backpropagation algorithm to modify the weights of the neurons, and thus minimize the errors between the actual output values and the target output values. The basic structure of BPNN consists of an input layer, a hidden layer, and an output layer. Details of the algorithm is described in (Park and Rilett, 1999).

SVR

SVR is SVM for regression, it is based on the computation of a linear regression function in a high dimensional feature space, where the input data are mapped via a non-linear function. Several studies have shown the SVR effectiveness in forecasting traffic flow (Chen et al., 2011; Zhang and Liu, 2009; Zhang and Xie, 2007).

2.3 Grid search and cross validation

Grid search is the process of scanning the data to configure the optimal parameters for a given model. Considering the possible values of models, grid search will build a model on each parameter combination and iterate through each combination accordingly. With different model performance, it can eventually select the optimal parameters of the given model. It needs to note that grid search can be extremely computationally expensive when dealing with a high dimensional set of parameters. Generally, parameters in the LSTM NN consist of the training batch-size, the epoch, the activation function of each hidden layer, the number of hidden layers and the hidden neurons.

Cross validation is a class of model evaluation methods. The basic idea is that some of the data is removed from the entire data set before training, and the removed data can be used to test the performance of the learned model on "new" data. According to the different training set and testing sets, there are mainly three kinds of cross validation, including the holdout method, *K*-fold cross validation, and leave-one-out cross validation. In the paper *K*-fold cross validation is adopted and the process is listed in as follows:

Step1. Dividing the data set into *k* subsets.

Step2. Taking one of the *k* subsets as the test set and the other *k*-1 subsets as the training set, the test error is recorded.

Step3. Training and testing the model for *k* times to ensure every data point gets to be in the test sets exactly once and in the training sets *k*-1 times.

Step4. Evaluating the model by calculating the average test error.

Grid search and cross validation are usually integrated together to find the model's optimal parameters and evaluate its performance on "new" data simultaneously (Anguita et al., 2009; Krstajic et al., 2014).

3 Experimentation

The short-term passenger flow prediction problem can be stated as follows. Let $\{x_t^T\}$ denote the observed passenger flow in the t^{th} time interval of the T^{th} day during a period time. Given a sequence $\{x_t^T\}$ of observed passenger flow, T=T, T-1, …, T-m, the goal is to predict the $\{x_t^{T+N}\}$, where N represents the prediction horizon and m represents the length of observation time period.

3.1 Data description

The daily sale data of Shanghai-Beijing HSR line from July 1 (Saturday), 2017 to July 31 (Monday), 2017 are provided by the Shanghai railway administration of China. There are 24 stations along Beijing-Shanghai HSR line with the length of 1,318 km. Table 1 lists the related data fields.

Table 1: List of data fields

No	Field	Description
1	Transcation_id	Identifying a transcation
2	Transcation_timestamp	Time of day, day of week
3	Station_name	Origin station, destination station

Considering the dispatched and attracted HSR passenger flows at each station, three OD pairs with the highest passenger demand include Shanghai-Beijing, Shanghai-Nanjing, and Nanjing-Beijing during the survey period. Taking July 1 as an example, these OD pairs consist of 23.42% of the total Shanghai-Beijing HSR travel demand (10.04%, 8.27%, and 5.11%, respectively).

3.2 Model development

Training data

The training data consists of input data and output data. In addition to passenger flow $X_t^T = (x_t^T,\ x_t^{T-1}, \cdots, x_t^{T-m})$, the features of time of day *t*, day of week *T* are all used as inputs in the model training stage, and the output is x_t^{T+N}. The validation dataset has the same features as the training dataset, and the test dataset has the same features as the inputs of training dataset, the output is the target value for prediction. Table 2 describes the code values of each feature.

Table 2: Feature descriptions for training data and label data

No	Feature	Code values	Description
1	Time of day (*t*)	0-11	0-11 represent the 12 hours in 6:00-18:00
2	Day of week (*T*)	1-7	1-7 represent Monday-Sunday
3	Passenger flow	x_t^T	hourly aggregated passenger flow

Based on the collected daily sale data, normally we have to normalize the training data by equation (9) to improve the model efficiency:

$$\bar{x} = (x - \mathrm{a})/(\mathrm{c} - \mathrm{b}) \tag{9}$$

where *x* denotes the code value of a feature, a, b, c is the average, minimum and maximum value of a feature, respectively, the normalized X_t^T, x_t^T, *T*, *t* is marked as ($X_{t'}^{T'}$, $x_{t'}^{T'}$, T', t'). As a result, for the LSTM NN, the final input is a sequence of passenger flow $x_{t'}^{T'}$ with its corresponding temporal features T' and t', that is, vectors $X_{t'}^{T'}$, $\boldsymbol{T}'$and $\boldsymbol{t}'$, respectively, and the output is $x_{t'}^{T'+1}$.

Model structure

The first 28-day (July 1-28) data are used for training (90%) and validation (10%), and the last 3-day (July 29-31, Saturday-Monday) data for testing. The prediction time ranges from 06:00-18:00, time interval is one hour. One-step ahead prediction means prediction horizon *N* is 1. HSR passenger flow shows regular changes in weeks, thus the length of observation time period *m* is 7. Given the small datasets in the paper, we fuse the grid search and the *k*-fold cross validation to enhance the model stability and reliability. The *k* is 3~10 empirically due to massive calculation. Here we set *k* as 7, which means to train the model 7 times on different training and validation data sets. According to related studies(Liu and Chen, 2017; Lv et al., 2015), the hidden layer (LSTM layer) size ranges from 1 to 6, the number of hidden units $l_N \in$ {30,60,90,120,150,180}, the activation function of each layer is "tanh", the batch-size $l_S \in$ {1,2,3,4}, epoch value depends on the number of training parameters based on equations (1-7), at last a dense layer is added to the output layer. Figure 2 shows the prediction process.

The deep learning library-Keras is a high-level neural networks API, written in Python and capable of running on top of TensorFlow, CNTK, or Theano. It was developed with a focus on enabling fast experimentation. Utilizing the Keras, Table 3 shows the optimal parameters in the LSTM NN.

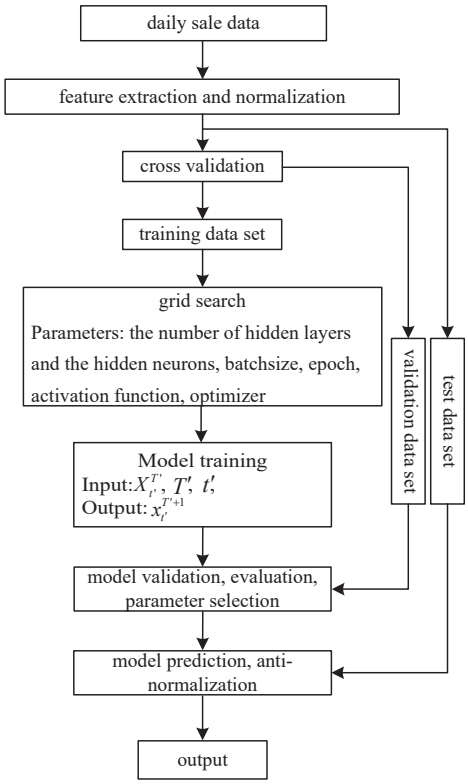


Figure 2: The training and forecasting process

Table 3: The key hyperparameters in the LSTM NN

Task	Hidden layers	Hidden units(bottom-top)
A day ahead passenger flow prediction (time interval : an hour)	2	[90,90]

For the ARIMA, the input is passenger flow vector X_t^T and the output is x_t^{T+1}, the best model ARIMA (5,0,2) selected by *auto.arima* function.

For the BPNN a, the input is $X_{t'}^{T'}$, $\boldsymbol{T'}, \boldsymbol{t'}$ and the output is $x_{t'}^{T'+1}$, the grid search and the cross-validation were utilized to build the optimal structure of the BP NN, while the hidden layer size was less than 3 caused by vanishing gradient.

For the SVR, the input is $X_{t'}^{T'}$, $\boldsymbol{T'}, \boldsymbol{t'}$ and the output is $x_{t'}^{T'+1}$, radial basis function (RBF) was used (Cherkassky and Ma, 2004) with three other parameters: cost C, width parameter g, and epsilon ε. Parameters were learned by parameter tuning function *tune* using gird search.

3.3 Model evaluation

The prediction accuracy is evaluated with the use of Mean Absolute Percentage Error (MAPE) and Root Mean Square error (RMSE). These measures are defined as follows:

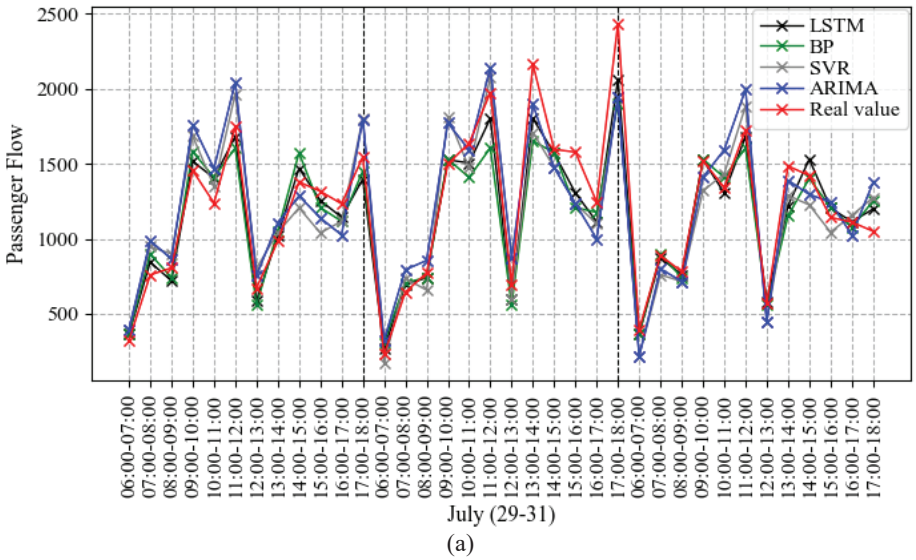
$$\mathrm{MAPE}(y,\hat{y}) = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i-\hat{y}_i|}{y_i} \tag{10}$$

$$\mathrm{RMSE}(y,\hat{y}) = \left[\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2\right]^{\frac{1}{2}} \tag{11}$$

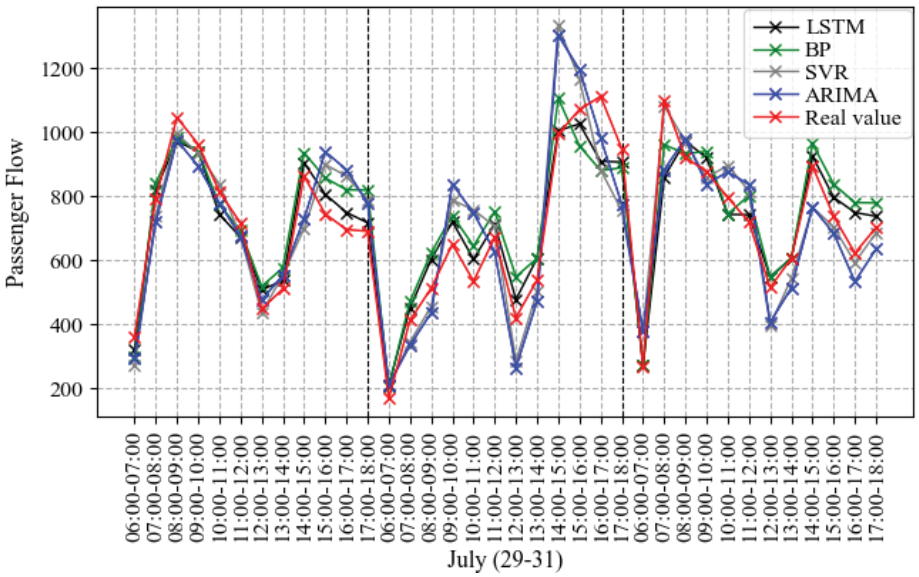
where y_i is the observed passenger flow, $\hat{y}_i$ is the predicted passenger flow, and n is the total number of predictions.

4 Results

The OD pairs (Shanghai-Beijing and Shanghai-Nanjing) with the highest passenger demand belong to two typical categories (i.e., long distance (more than 800 km) and middle distance (200-800 km)) in the railway operations, which are selected for the prediction. Figure 3 shows the outputs of ARIMA, BPNN, SVR, and LSTM NN. Table 4 presents the forecasting errors.



(a)



(b)

Figure 3: Prediction results from July 29 to 31 (a: Shanghai-Beijing; b: Shanghai-Nanjing)

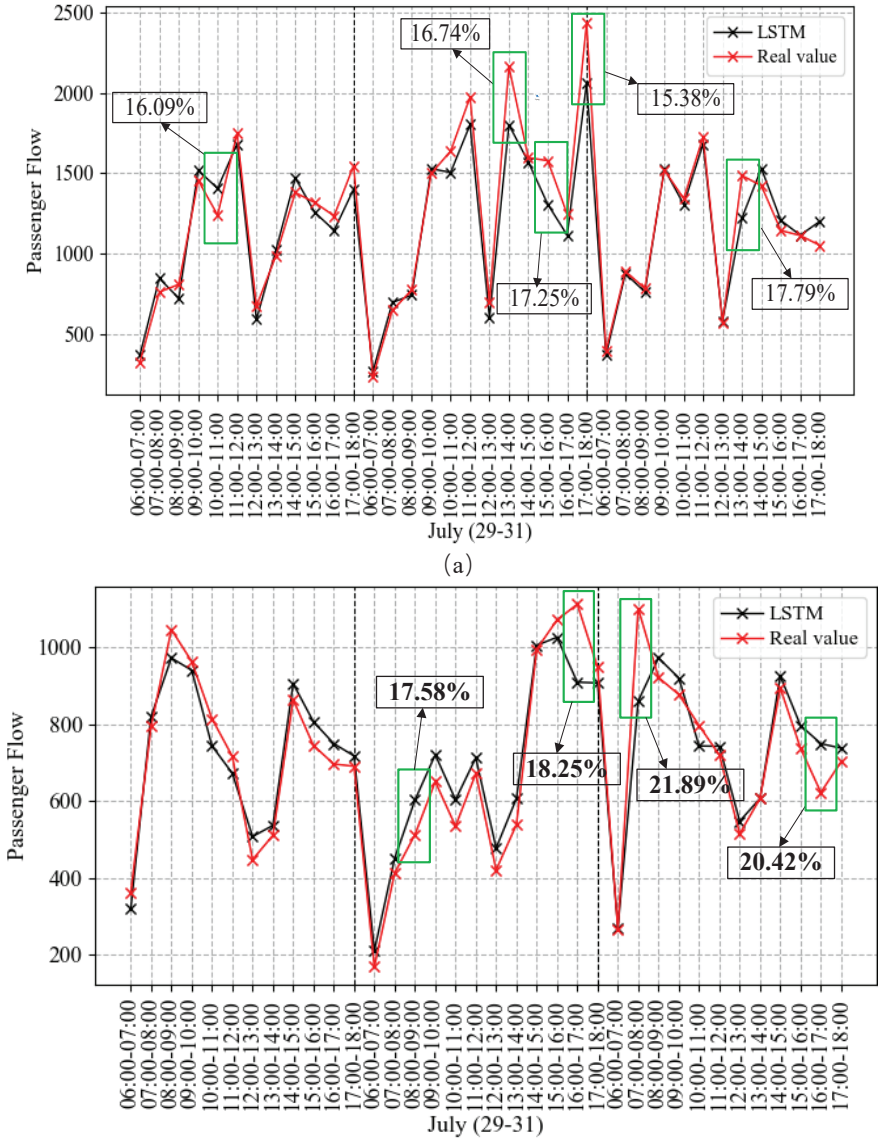
Table 4: Prediction accuracy comparison

OD	Model	MAPE	RMSE	OD	Model	MAPE	RMSE
Shanghai-Beijing	LSTM	8.01%	133	Shanghai-Nanjing	LSTM	8.47%	73
	BP	11.79%	187		BP	11.96%	91
	SVR	15.17%	198		SVR	14.39%	117
	ARIMA	16.49%	204		ARIMA	16.25%	123

According to the MAPE and RMSE values in Table 4, the LSTM NN is superior to the BPNN, SVR, and ARIMA for the short-term HSR passenger flow prediction, while the BPNN outperforms SVR, and SVR outperforms ARIMA. The results strongly indicate that the relationship between the historical and forecasting passenger flow are non-linearly correlated.

Also, Figure3 shows that passenger flow time distribution is significantly different in the two chosen OD pairs. Additionally, the test data include passenger flow of both working and non-working days (Saturday-Monday) between which passenger flow time distribution also shows a weak regularity for both OD pairs, however, the prediction accuracy of four models shows slight fluctuations. The observations can be attributed to two possible reasons. On the one hand, the adopted models are efficient enough to forecast passenger flow. On the other hand, given the chosen OD pairs, the mathematical relationship between the input and output is similar, which leads to the stable model performance.

Nonetheless, the LSTM NN does not seem to perform well in several time intervals (e.g., 11:00-12:00 on July 29 and 15:00-16:00 on July 30 (Shanghai-Beijing);16:00-17:00 on July 30 and July 31 (Shanghai-Nanjing)) in Figure 4.



(a)

Figure 4: Prediction errors analysis (Shanghai-Beijing, Shanghai-Nanjing)

The phenomenon may be explained that the model input features (i.e., time of day, day of week, and passenger flow volume) are not sufficient for prediction under special scenarios. For instance, passenger flow with a sharp increase (e.g., 13:00-14:00 on July 30(Shanghai-Beijing)) or decrease (e.g., 16:00-17:00 on July 31(Shanghai-Nanjing)) was easily affected by other factors (e.g., weather and an emergency), we have not considered these factors yet and consequently degraded the model prediction performance. More features need to be assessed and added to enhance the model reliability and stability. Besides, to analyze the LSTM NN performance with different time intervals, x_t^T is also aggregated by 2-hour (3-hour, 4-hour, and 5-hour), Figure 5 shows the forecasting errors.

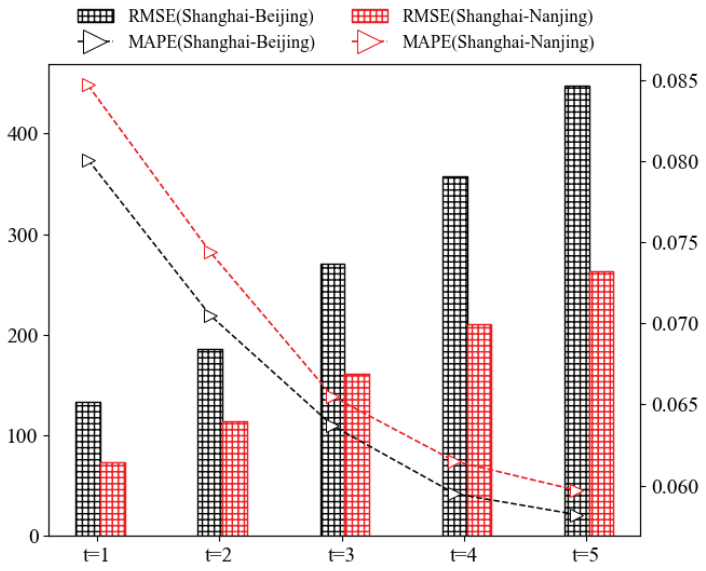


Figure 5: Prediction errors with different time intervals

With larger prediction time interval, the forecasting errors of both OD pairs become smaller in Figure 5. The underlying reason is that smaller time interval increases the fluctuation of passenger flow data, while greater interval contributes to more stability and thus passenger flow prediction is relatively easier. The conclusion is consistent with the variability in regularity of metro passenger flow (Zhong et al., 2016), it points out that dramatically increased invariability may occur up to the temporal scale of about 15minutes according to the cases of Beijing, Singapore, and London, implying that time interval limits exist when we attempt to forecast the short-term passenger flow.

Additionally, we evaluate the forecasting performance in up to 7-step ahead, that is, to predict x_t^{T+1},x_t^{T+2}, x_t^{T+3}, x_t^{T+4}, x_t^{T+5}, x_t^{T+6}, x_t^{T+7}, respectively. We first forecast the first step, then the forecasted value is applied as a part of the input variables for the next step prediction. Thus, the entire horizon can be repeatedly predicted step by step. Figure 6 shows the forecasting errors.

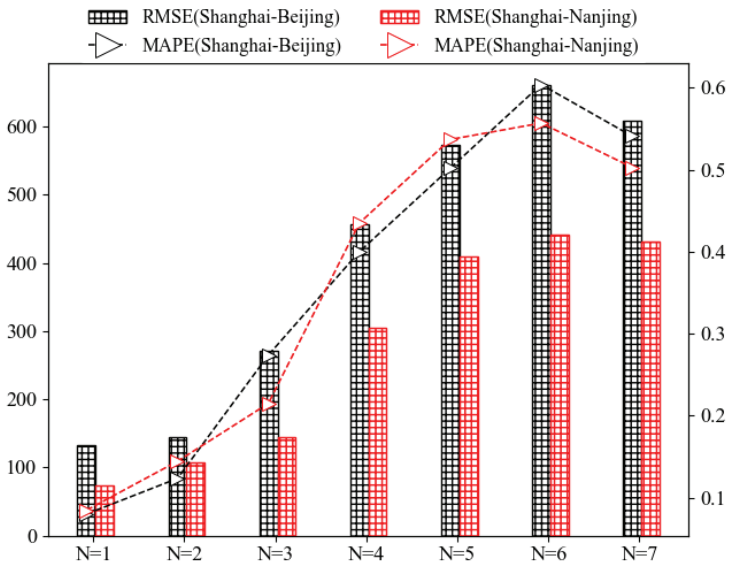


Figure 6: Prediction errors with different forecasting steps

As the forecasting step increases, the prediction accuracy also decreases in Figure 6. When N>3, the performance of the LSTM NN degrades significantly. The reason could be that the cumulative forecasting error causes the input data to be much less efficient, and x_t^{T+N} is less predictable.

5 Conclusions

The main contribution of the paper is to develop a LSTM model to predict short-term HSR passenger flow. Preprocessing the raw data, we construct the input and output variables, and get the optimal LSTM parameters through grid search and cross-validation. Compared with the traditional forecasting methods (ARIMA, BPNN, and SVR), the proposed model poses a great potential for short-term passenger flow prediction. In addition, the LSTM NN performed better with larger time intervals, but performance degrades significantly with the increase of forecasting step.

One of the drawbacks of deep learning models is the low explanatory power (Polson and Sokolov, 2017)). In a recent review of short-term forecasting technique (Vlahogianni et al., 2014), model interpretability is mentioned as one of the barriers in implementing more sophisticated machine learning models in practice. Liu and Chen (2017) tried to explain that the deep architecture could extract the deep features and perform well in BRT passenger flow prediction, while the passenger travel behavior of BRT is different from HSR. The issue needs to be studied further.

Acknowledgements

Shanghai railway administration of China is acknowledged for providing the data, and they own the data used in the present paper. The authors would like to acknowledge the support from National Natural Science Foundation of China (No.71771191).

References

Abigogun, O.A., 2005. Data mining, fraud detection and mobile telecommunications: Call pattern analysis with unsupervised neural networks, Master Thesis, University of the Western Cape.

Ahmed, M.S., and Cook, A.R., 1979. "Analysis of freeway traffic time-series data by using Box-Jenkins techniques", *Transportation Research Record: Journal of the Transportation Research Board*, vol. 773, pp. 1-9.

Anguita, D., Ghio, A., Ridella,S., Sterpi, D., 2009. "K-fold cross validation for error rate estimate in support vector machines", In: *Proceedings of The International Conference on Data Mining*, Austin, United States.

Anthony, S., Karlaftis, M.G., 2003. "A multivariate state space approach for urban traffic flow modeling and prediction", *Transportation Research Part C: Emerging Technologies*, vol. 11, pp. 121-135.

Bai, Y., Sun, Z., Zeng, B., Deng, J., Li, C, 2017. "A multi-pattern deep fusion model for short-term bus passenger flow forecasting", *Applied Soft Computing*, vol. 58, pp. 669-680.

Chen, Q., Li, W., Zhao, J., 2011. "The use of LS-SVM for short-term passenger flow prediction", *Transport*, vol. 26, pp. 5-10.

Cherkassky,V., Ma, Y. 2004. "Practical selection of svm parameters and noise estimation for svm regression", *Neural networks*, vol. 17, pp. 113-126.

Collobert, R., and J. Weston., 2008. "A unified architecture for natural language processing: Deep neural networks with multi-task learning", In: *Proceedings of The 25th International Conference on Machine Learning*, Helsinki, Finland.

Ding, C., Wang, D., Ma, X., Li. H., 2016. "Predicting short-term Subway ridership and prioritizing its influential factors using gradient boosting decision trees", *Sustainability*, vol. 8, pp. 1100.

Fang, L.J., Wu, Z., 2006. "Application of GM (1, 3) in highway passenger capacity forecast of transportation system", *Journal of Highway & Transportation Research & Developmen*t, vol. 26, pp. 163-166.

Gers, F., 2001. Long Short-term Memory in Recurrent Neural Networks, Ph.D. Dissertation, École Polytechnique Fédérale De Lausanne.
Gers, F. A., Schmidhuber, J., Cummins, F., 2000. "Learning to forget: Continual prediction with LSTM", *Neural Computation*, vol. 12, pp. 2451-2471.
Goodfellow, I.J., Bulatov, Y., Ibarz, J., Arnoud, S., Shet, V., 2013. "Multi-digit number recognition from street view imagery using deep convolutional neural networks", ArXiv Preprint arXiv:1312.6082v4.
Hamed, M.M., AI-Masaeid, H.R., Bani Said, Z.M., 1995. "Short-term prediction of traffic volume in urban arterials", *Journal of Transportation Engineering*, vol. 121, pp. 249-254.
Hinton, G.E., Salakhutdinov, R.R., 2006. "Reducing the dimensionality of data with neural networks", *Science*, vol.313, pp. 504-507.
Hochreiter, S., Schmidhuber, J., 1997. "Long short-term memory", *Neural Computation*, vol. 9, pp. 1735-1780.
Hsu, C.I., Wen, Y.H., 1998. "Improved grey prediction models for the trans-pacific air passenger market", *Transportation Planning & Technology*, vol 22, pp. 87-107.
Jiang. X., Zhang. L., and Chen. X., 2014. "Short-term forecasting of high-speed rail demand: a hybrid approach combining ensemble empirical mode decomposition and gray support vector machine with real-world applications in China", *Transportation Research Part C: Emerging Technologies*, vol. 44, pp. 110-127.
Kecman, P., Goverde, R. M. P., 2015. "Predictive modelling of running and dwell times in railway traffic", Public Transport, vol. 7, pp.295-319.
Krstajic, D., Buturovic, L. J., Leahy, D. E., Thomas, S., 2014. "Cross-validation pitfalls when selecting and assessing regression and classification models", *Journal of Cheminformatics*, vol. 6, pp. 10.
LeCun, Y., Bengio, Y., Hinton, G.E., 2015. "Deep learning", *Nature*, vol. 521, pp. 436-444.
Lee, S., Fambro, D.B., 1999. "Application of subset autoregressive integrated moving average model for short-term freeway traffic volume forecasting", *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1678, pp. 179-188.
Leng, B, Zheng, J.B, Xiong, Z., Lv. W.F, Wan. Y.L., 2013. "Probability tree-based passenger flow prediction and its application to the Beijing subway system", *Frontiers of Computer Science*, vol. 7, pp. 195-203.
Li, Y., Wang, X., Sun, S., Ma, X., Lu. G., 2017. "Forecasting short-term subway passenger flow under special events scenarios using multiscale radial basis function networks", *Transportation Research Part C: Emerging Technologies*, vol. 77, pp. 306-328.
Liu, L., Chen, R., 2017. "A novel passenger flow prediction model using deep learning methods", *Transportation Research Part C: Emerging Technologies*, vol. 84, pp.74-91.
Lv, Y.S., Duan, Y.J., Kang, W.W., Li, Z.X, Wang, F.Y., 2015. "Traffic flow prediction with big data: a deep learning approach", *Intelligent transportation systems*, vol.16, pp. 865-873.
Mackenzie, J., Roddick, J.F., Zito, R., 2018. "An evaluation of HTM and LSTM for short-term arterial traffic flow prediction", *Intelligent transportation systems*, pp.1-11.
Manoel, C.N., Jeong, Y.S., Jeong, M.K., Lee, D.H., 2009. "Online-SVR for short-term traffic flow prediction under typical and atypical traffic conditions", *Expert Systems with Applications*, vol. 36, pp. 6164-6173.

Okutani, I., and Stephanedes, Y.J., 1984. “Dynamic prediction of traffic volume through Kalman filtering theory”, *Transportation Research Part B: Methodological*, vol.18, pp. 1-11.

Polson, N.G., Sokolov, V.O., 2017. “Deep learning for short-term traffic flow prediction”, *Transportation Research Part C: Emerging Technologies*, vol. 79, pp. 1-17.

Park, D., Rilett, L.R., 1999. “Forecasting freeway link travel times with a multilayer feedforward neural network”, *Computer-Aided Civil and Infrastructure Engineering*, vol. 14, pp. 357-367.

Sun, Y, Leng, B., Guan, W., 2015. “A novel wavelet-SVM short-time passenger flow prediction in Beijing subway system”, *Neurocomputing*, vol. 166, pp. 109-121.

Tsai, T., Lee, C., Wei, C., 2009. “Neural network based temporal feature models for short-term railway passenger demand forecasting”, *Expert Systems with Applications*, vol. 36, pp. 3728-3736.

Vlahogianni, E.I., Karlaftis, M.G., Golias, J.C., 2014. “Short-term traffic forecasting: Where we are and where we’re going”, *Transportation Research Part C: Emerging Technologies*, vol. 43, pp. 3-19.

Wei, Y., Chen, M.C., 2012. “Forecasting the short-term metro passenger flow with empirical mode decomposition and neural networks”, *Transportation Research Part C: Emerging Technologies*, vol. 21, pp. 148-162.

Williams, B.M., Durvasula, P.K., Brown D.E., 1998. “Urban freeway traffic flow prediction: Application of seasonal autoregressive integrated moving average and exponential smoothing models”, *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1644, pp. 132-141.

Williams, B.M., Hoel, L.A., 2003. “Modeling and forecasting vehicular traffic flow as a seasonal ARIMA process: Theoretical basis and empirical results”, *Journal of Transportation Engineering*, vol. 129, pp. 664-672.

Zhang, L., Liu,Q.C., Yang,W.C., Wei, N., Dong, D.C., 2013. “An improved K-nearest neighbor model for short-term traffic flow prediction”, *Procedia-Social and Behavioral Sciences*, vol. 96, pp. 653-662.

Zhang, Y., Liu, Y., 2009. “Traffic forecasting using least squares support vector machines”, *Transportmetrica*, vol, 5, pp.193-213.

Zhang, Y., Xie, Y., 2007. “Forecasting of short-term freeway volume with v-support vector machines”, *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2024, pp. 92-99.

Zheng, W.Z., Lee, D.H., Shi, Q.X., 2006. “Short-term freeway traffic flow prediction: Bayesian combined neural network approach”, *Journal of Transportation Engineering*, vol. 132, pp. 114-121.

Zhong. C., Batty, M., Manley, E., Wang, Z., Chen, F., 2016. “Variability in regularity: Mining temporal mobility patterns in London, Singapore and Beijing using smart-card data”, *Plos One*, vol. 11, e0149222.

Railway Infrastructure Capacity Utilization Description through Data Integration in Blocking Time Theory

Qinglun Zhong [a], Shaoquan Ni [b,c,d], Shengdong Li [b,e], Chang'an Xu [b,1]
[a] Institut für Eisenbahnwesen und Verkehrssicherung, Technische Universität Braunschweig Pockelsstr. 3, 38106 Braunschweig, Germany
[b] School of Transportation and Logistics, Southwest Jiaotong University 610031 Chengdu, China
[1] E-mail: bk20085673@my.swjtu.edu.cn, Phone: +86 (0) 28 87600706
[c] National Railway Train Diagram Research and Training Centre, Southwest Jiaotong University 610031 Chengdu, China
[d] National and Local Joint Engineering Laboratory of Comprehensive Intelligent Transportation, Southwest Jiaotong University 610031 Chengdu, China
[e] Department of Management, Technical University of Denmark 2800 Kgs. Lyngby, Denmark

Abstract
We propose a method to describe capacity utilization for railway infrastructure that applies blocking time theory to managing train runs. Different from traditional capacity evaluation, infrastructure capacity utilization description shows detailed information on infrastructure utilization hidden in timetabling data instead of sheer number of trains that can be operated, or capacity consumed. Using a function system defined upon necessary operational inputs for timetabling in blocking time theory, we can obtain the feasibility condition for operating consecutive trains. Thus, the method to identify critical block section can be deduced from the feasibility condition. Structural indication determines the capacity utilization of consecutive train paths, which can be further integrated into a bi-directional graph to model infrastructure capacity utilization description followed by infrastructure time allocation. Consumed capacity of railway infrastructure by operating train runs can be formulated. Besides, a general procedure is proposed to analyse the sensitivity of consumed capacity to operational inputs. An experimental case study is conducted to demonstrate the application of this method in analysing the impact of speed and recovery time.

Keywords
Blocking time theory, Capacity analysis, Infrastructure capacity utilization description, Timetabling data

1 Introduction

Railway capacity analysis or calculation is almost an ancient problem in the field of railway operations. Yet it has not been eliminated from a rather critical role in infrastructure utilization management and rolling stock utilization. And the increasing emphasis on energy efficiency, CO_2 emission reduction, and environmental protection can be better met by new generation railway services, especially the so-called high-speed railway. With the fourth railway package issued, a more open and competitive railway market reinforced by relevant administrative and technical artifices can be expected. The increasing demands on railway services also require highly efficient managerial techniques of railway infrastructure and

rolling stocks for proper strategizing as a potential reply.

Along with the introduction of German railway reform, blocking time theory was employed in service planning by the German railway infrastructure manager DB Netz AG, when developing its computer-aided timetabling system in the late 1990s. Blocking time theory was developed by Happel (Happel, 1959) and is now widely used for timetabling in Europe. It allows the description of train runs on different railway networks with different signalling and control systems. It also enhances railway operations with competitive edges over train diagramming in many ways, one of which is that blocking time theory visualizes the infrastructure occupation by a specific train. And this is especially significant in terms of conflict detection and resolution in a competitive business environment under the duality of infrastructure-operation. Blocking time theory will continue to dominate European railway operations in a foreseeable future, which further drives managerial innovations.

The overall structure of this paper is organized as follows. Section 2 reviews related literatures. Section 3 introduces the operational inputs required to perform the analysis. Section 4 presents the method to critical block section identification and parametrically indicating the structure of feasible train paths, providing the theoretical foundation for the description of infrastructure capacity utilization. Section 5 integrates the results achieved in previous steps into infrastructure capacity utilization description and gives the method to calculate fragmented infrastructure capacity, which constitutes the allocation of infrastructure time along with structural indicators. An experimental case study is reported in section 6, and the impact of parameters, such as speed and recovery time, on consumed is discussed. And section 7 concludes this work briefly.

2 Literature Review

Numerous methods have been developed to analyse and calculate railway capacity. Many scholars have classified these methods from different perspectives. Pachl (2018) classified the capacity methodologies as two major classes: analytic and simulation. Another well-received overview concerning railway capacity issues was provided by Abril et al. (2008), which further divided the relevant methods into three levels: analytical, optimization, and simulation methods. Sameni et al. (2011) categorized capacity evaluation methods to be timetable-based and non-timetable-based. Among all classifications, the classification presented by Abril et al. is most widely noted (Abril et al., 2008), based on which this research summarizes existing methodologies on capacity researches.

2.1 Analytical Methods

Analytical methods are designed to model the railway environment by means of algebraic expressions or mathematical formulae (Abril et al., 2008). They usually obtain theoretical capacities and determine practical capacities either as a percentage of the theoretical capacity or by including regularity margins (Yaghini et al., 2014). The UIC method proposed by the International Union of Railways (UIC) is an important one within this category, which is based on visually compressing timetable (UIC, 2004). This method measures the consumed capacity of sections for a given infrastructure based on pre-determined timetable (Landex et al., 2006; Jensen et al., 2017), though it is also argued that the method can also be applied when the infrastructure is not divided into sections (Landex, 2008). Many researches have been produced in terms of analysing the method (Landex, 2009), including propositions to improve it following different ideologies, such as (Lindner

& Pachl, 2010; Lindner, 2011), which eventually resulted in an improved update (UIC, 2013). Other important analytical methods include the subtraction factor method (Yan, 1997; Zhao, 2001), the minimum interval method (Zhang, 2015; Jamili, 2018), and parametric method (Lai and Barkan, 2009; Lai and Barkan, 2011).

In general, analytical methods are useful for the calculation of railway capacity at a planning level, as well as for the identification of bottlenecks in the infrastructure. However, different methods may provide very different results when studying the same line since they are very sensitive to the parameters used and variations in the composition of trains (Riejos et al., 2016).

2.2 Optimization Methods

Optimization methods are designed to provide more strategic methods for solving the railway capacity problem other than purely analytical formulae (Abril et al., 2008). The ideological basis of optimization methods is usually timetable saturation through mathematical programming. Optimization techniques, such as tabu search (Higgins, 1998), branch-and-bound (Higgins et al., 1996), Lagrangian relaxation (Caprara et al., 2002) and heuristic algorithms (Carey and Lockwood, 1995) are designed to solve railway capability analysis problems.

The railway capacity optimization methods can be roughly divided into deterministic optimization methods and stochastic optimization methods. For deterministic optimization method, an initial timetable is required. Recent contributions that belong to deterministic optimization method include Yaghini et al. (2014), Harrod (2009), Petering et al. (2015), and Burdett (2015). But for stochastic optimization method, it does not need an initial timetable, instead it requires the probability distribution of relevant time variables and dwell times (de Kort et al., 2003). Recent papers of this type include Burdett and Kozan (2005). Kroon et al. (2008), and Medeossi et al. (2011).

Optimization methods may be useful for problems of uncomplicated nature, but it could be very difficult to solve a model with very complex capacity and traffic constraints.

2.3 Simulation Methods

Simulation methods are usually provided a model as close to reality as possible, to validate a given timetable (Abril et al., 2008). These methods attempt to replicate the actual operation of trains within a line or a railway network (Riejos et al., 2016). Excellent surveys of the railway capacity simulation methods have been done by Pouryousef et al. (2015). There are two basic simulation models: microscopic and macroscopic model. And some works are based on the integration of both models (Kettner et al., 2003). While most simulation models fall into these two categories, mesoscopic models can be created by simplifying microscopic model or enriching details in macroscopic model with proper skills (Gille et al., 2008; Marinov and Viegas, 2011; Jensen et al., 2017).

Realization of simulation models requires specific tools. Current mainstream railway capability simulation software includes SIMONE, RailSys, and OpenTrack. More information about railway simulation tools can be found in Barber et al. (2007).

Simulation is most effective method to analyze capacity for infrastructure of limited size (Lai et al., 2014), and they become computationally intensive when applied in network level. In addition, these models are sensitive to data because of their dependency of complex operational data as inputs, such as geometrical configuration, velocity of trains, and

movement rules.

In conclusion, existing methods for capacity analysis basically focus on infrastructure capacity determination or evaluation in terms of time consumed or numbers of trains that can be operated based on ready timetables or other operational parameters. However, other useful information in timetabling data remains unrevealed. We feel that capacity researches can be approached from another angle where the deterministic relationship between timetabling data and infrastructure capacity utilization can be clarified and utilized. For instance, formulations for timetable optimization programs are generally based on timetabling, which lacks the insight from this connectivity that can potentially simplify computations. Therefore, this paper proposes a description of capacity utilization for railway infrastructure that applies blocking time theory.

3 Operational Inputs

This paper considers one direction of double-track railway infrastructure whose operation is based on blocking time theory.

An arbitrary train i operating on the infrastructure is always defined in section $[O_i, D_i]$, where O_i and D_i denote the first and last block section on the operation route of train i. The operation route of train i does not necessarily overlap the infrastructure which we analyze.

In order to clarify the relationship between consecutive train paths, define two train paths as a train pair if they meet following conditions:

(i) they directly follow each other over certain section of railway infrastructure;

(ii) the lower blocking time of the leading train can be scheduled at the same time as the upper blocking time of the following train in certain block section, without causing conflicts to any other train.

Condition (i) demands that train i is followed by $i+1$ during a certain section on the infrastructure. Interpretation of condition (ii) involves feasibility issues and can be referred to section 4.2 and 4.3. Let train i and $i+1$ form a train pair, denoted as $(i, i+1)$, on their common operation route $[o_{i,i+1}, d_{i,i+1}]$. Noticeably, a train pair is sequence-relevant. There is always an arbitrary block section $j \in [o_{i,i+1}, d_{i,i+1}]$ when we talk about train pair $(i, i+1)$ unless specified otherwise.

And the information required for analyzing the utilization of infrastructure from timetabling data can be called operational inputs, as in the following explanations.

(i) Time for signal setup A_i^j denotes the time needed to set up the signal to operate in block section j for train i.

(ii) Time for signal confirmation B_i^j denotes the time needed for the train driver to confirm the signal to approach in block section j for train i.

(iii) Approach time C_i^j denotes the time needed for train i to end block section j.

(iv) Running in a block section r_i^j denotes the time needed for train i to cover the whole length of block section j. It is usually the sum of pure calculated running time and recovery margin which makes up certain percentage of the total running time.

(v) Time for clearance D_i^j denotes the time needed for train i to clear block section j.

(vi) Time for release E_i^j denotes the time needed for railway operation system to release

the signal of block section j after the traverse by train i.

(vii) Scheduled stop d_i^j denotes the duration of a scheduled stop of train i at station in block section j.

(viii) Operation sequence $(1,\cdots,i,\cdots,m)$ denotes the sequence of train departing from certain block sections of the infrastructure.

(ix) Overtaking arrangement $(i,i+1)\xrightarrow{j}(i+1,i)$ denotes a change of operation sequence from $(i,i+1)$ to $(i+1,i)$ at station in block section j. It is noteworthy that $(i,i+1)$ and $(i+1,i)$ should be treated as two train pairs on different sections.

All mathematical notations used in this paper are listed in the Appendix.

4 Capacity Analysis of Consecutive Train Runs

Before analysing infrastructure capacity utilization that is determined by its timetable, the method to study its occupation that is determined by the structure of consecutive train runs is introduced in this section.

4.1 Function System

The time spent from the departure of a train at a certain node to another node on its route of operation, can be calculated and used to describe the temporal proceeding of that train. It is called process time, different from departure and arrival time in a ready timetable, with which timetable structure can be restated.

(1) Single train path

Define the process time of train i when entering block section $j\in[O_i,D_i]$ from block section O_i as the entry process time of train i in block section j, denoted as $p_i^{O_i,j}$, and it can be given by

$$p_i^{O_i,j}=\sum_{k=O_i}^{j-1}(r_i^k+d_i^k), \tag{1}$$

where k is the universal serial number. Process times are but intermediate to model train runs in blocking time theory, so that capacity analysis can be performed. Since planning timetables in blocking time theory relies on blocking time, the upper blocking time of train i in block section $j\in[O_i,D_i]$ from block section O_i can be given by

$$b_{i,up}^{O_i,j}=p_i^{O_i,j}-A_i^j-B_i^j, \tag{2}$$

where $j=O_i$, or $d_i^{j-1}\neq 0$.

And

$$b_{i,up}^{O_i,j}=p_i^{O_i,j}-A_i^j-B_i^j-C_i^{j+h}, \tag{3}$$

where $j\neq O_i$, or $d_i^{j-1}=0$. And the lower blocking time of train i in block section $j\in[O_i,D_i]$ from block section O_i can be given by

$$b_{i,low}^{O_i,j}=p_i^{O_i,j}+D_i^j+E_i^j+r_i^j+d_i^j. \tag{4}$$

(2) Train pair

Let the blocking time difference of train pair $(i,i+1)$ on block section $j\in[o_{i,i+1},d_{i,i+1}]$

be $t_{i,i+1}^{o_{i,i+1},j}$, given by

$$t_{i,i+1}^{o,j} = b_{i+1,up}^{o,j} - b_{i,low}^{o,j}, \tag{5}$$

where the subscript $i,i+1$ of notation $o_{i,i+1}$ is intentionally left out given there is no confusion, just in case equations get too long and unreadable. Replacing the right-hand sided blocking times of equation (5) with equation (2-4) yields

$$t_{i,i+1}^{o,j} = p_{i+1}^{o,j} - p_i^{o,j} - (A_{i+1}^{j+1} + B_{i+1}^{j+1} + C_{i+1}^{j+1}) - (D_i^{j+1} + E_i^{j+1}) - (r_i^j + d_i^j). \tag{6}$$

4.2 Feasibility

A complex train path structure can always be decomposed into several train pairs, each of which be conflict-free, when analyzing the infrastructure occupation of train paths. Suppose that train pair $(i,i+1)$ exists in section $[o_{i,i+1}, d_{i,i+1}]$ and its blocking time difference at block section j can be given by

$$t_{i,i+1}^{o,j} = p_{i+1}^{o,j} - p_i^{o,j} - (A_{i+1}^{j+1} + B_{i+1}^{j+1} + C_{i+1}^{j+1}) - (D_i^{j+1} + E_i^{j+1}) - (r_i^j + d_i^j). \tag{7}$$

The departure time when train i entries section $j \in [o_{i,i+1}, d_{i,i+1}]$ can be denoted as $y_{i,j}$. Departure time $y_{i,1}$ denotes the departure time of train i from its origin block section. It is obvious that

$$y_{i,j} = y_{i,j-1} + r_i^j + d_i^j. \tag{8}$$

And departure times can be calculated using process times using following equations

$$y_{i,j} = y_{i,o} + p_i^{o,j}, \tag{9}$$

$$p_i^{o,o} = 0. \tag{10}$$

Using equation (11), the first two items on the right-hand side of equation (9) can be written as

$$p_{i+1}^{o,j} = y_{i+1,j} - y_{i+1,o}, \tag{11}$$

$$p_i^{o,j} = y_{i,j} - y_{i,o}. \tag{12}$$

Substituting equation (13) and (14) into equation (10) yields

$$t_{i,i+1}^{o,j} = (y_{i+1,j} - y_{i+1,o}) - (y_{i,j} - y_{i,o}) - (A_{i+1}^j + B_{i+1}^j + C_{i+1}^j) - (D_i^j + E_i^j) - (r_i^j + d_i^j). \tag{13}$$

Denote the mark of the lower blocking time of train i in block section j on the time axis as $m_{i,j}^{low}$, and it can be expressed by

$$m_{i,j}^{low} = y_{i,j} + (D_i^j + E_i^j). \tag{14}$$

Denote the mark of the upper blocking time of train i in block section j on the time axis as $m_{i,j}^{up}$, and it can be expressed by

$$m_{i+1,j}^{up} = y_{i+1,j} - (A_{i+1}^j + B_{i+1}^j + C_{i+1}^j). \tag{15}$$

Substituting equation (14) and (15) into equation (13) yields

$$m_{i+1,j}^{up} - m_{i,j}^{low} = t_{i,i+1}^{o,j} + y_{i+1,o} - y_{i,o} - r_i^j - d_i^j. \tag{16}$$

Train pair $(i,i+1)$ is feasible if and only if the $m_{i+1,j}^{up} - m_{i,j}^{low}$ is nonnegative for $\forall j \in [o_{i,i+1}, d_{i,i+1}]$, or the right-hand side of equation (16) being nonnegative.

Thus, the nonnegativity of $m_{i+1,j}^{up} - m_{i,j}^{low}$ can be called the feasibility condition of train

pair $(i, i+1)$ in block section $j \in [o_{i,i+1}, d_{i,i+1}]$

4.3 Critical Block Section

There is at least one block section on the common operation route of a train pair, where their blocking time squares elapse earlier than those in other block sections when pushing their train paths closer together. It supports the train path structure of a train pair and determines the occupation of infrastructure by them. This section presents the method to its identification using operational data.

Non-Overtaking Operation

Let train i and $i+1$ depart into section $o_{i,i+1}$ at the same time, meaning $y_{i+1,o} = y_{i,o}$. Train i and $i+1$ are obviously conflicted in section $[o_{i,i+1}, d_{i,i+1}]$. Thus, a value must be added to the right-hand side of equation (16), which is denoted as $I_{i,i+1}^{g_{i,i+1}}$, and equation (16) transforms into

$$m_{i+1,j}^{up} - m_{i,j}^{low} = t_{i,i+1}^{o,j} - r_i^j - d_i^j + I_{i,i+1}^j. \tag{17}$$

Substituting $m_{i+1,j}^{up} = m_{i,j}^{low}$ into equation (17) yields

$$I_{i,i+1}^j = d_i^j + r_i^j - t_{i,i+1}^{o,j}. \tag{18}$$

Equation (18) gives the minimum value needed to make train pair $(i, i+1)$ feasible in section $j \in [o_{i,i+1}, d_{i,i+1}]$, and anything more than that might be considered as buffer time.

Adding a positive value to the right-hand side of equation (17) signifies letting the train path $i+1$ translate away from train path i by that value. Notice that during the transition of train pair that follows a fixed operation sequence on the time axis, their blocking time differences on block sections within a fixed section increase or decrease proportionately during the process. This shows the structural stability of a train pair given their parameters constant, meaning $\{I_{i,i+1}^j\}$ is certain. In order to make train path i and $i+1$ conflict-free, a value large enough should be added. There could be more than one case that can make them so, and we define the section that is traversed the latest among the sections with the same largest $I_{i,i+1}^j$ to be the critical block section of train pair $(i, i+1)$, which can be mathematically expressed as

$$g_{i,i+1} = max\{ j \mid max[I_{i,i+1}^j] \} \quad j \in [o_{i,i+1}, d_{i,i+1}]. \tag{19}$$

Notice that it is unnecessary to distinguish between homogeneous and heterogeneous train operations for a train pair, since the method presented can treat them in general.

Complex Overtaking Operation

There are often complex overtakes which involve more than two train paths, when scheduling timetables for railway network of limited scope. There should be quite some instances that are of this kind when considering railway network covering a considerably large area. The method to obtain a feasible schedule is to examine all trains according to the operation sequence.

Consider the scenario that train i acts as the leading train in the train pair formed with train $\mu \in \{\mu\}$ in section $[o_{i,\mu}, d_{i,\mu}]$. Thus, the critical block section of train pair (i, μ) can

be expressed as $g_{i,\mu} \in [o_{i,\mu}, d_{i,\mu}]$. And their largest value to be added can be given by $I^{g}_{i,\mu}$.

Assume that trains within set $\{\mu\}$ are scheduled conflict-free and involving overtakes. All train pair involving train i and $\mu \in \{\mu\}$ with train i being the leading train, if and only if a critical block section g_i exists for the complex structure and satisfies the following

$$g_i = max\{k \mid max[I^{g_{i,\mu}}_{i,\mu}]\}. \tag{20}$$

4.4 Structural Indication

The structural stability of a train pair can be exploited to describe the capacity utilization of a train pair. For this purpose, define the structural indicator of train pair $(i, i+1)$ on block section $j \in [o_{i,i+1}, d_{i,i+1}]$ to be the difference between the added value of train pair $(i, i+1)$ on its critical block section $g_{i,i+1} \in [o_{i,i+1}, d_{i,i+1}]$ and block section $j \in [o_{i,i+1}, d_{i,i+1}]$, and its mathematical expression can be written as

$$s^{j}_{i,i+1} = I^{g}_{i,i+1} - t^{j}_{i,i+1}. \tag{21}$$

Structural indicator $s^{j}_{i,i+1}$ can be used to denote the minimum infrastructure time interval to operation two consecutive train paths that form a train pair on their common section $[o_{i,i+1}, d_{i,i+1}]$.

5 Infrastructure Capacity Analysis

We consider describing infrastructure capacity utilization based on analytical results from previous steps. And a general method to analyse the impact of timetabling data is summarized based on the formulation of consumed capacity.

5.1 Infrastructure Capacity Utilization Description

As suggested by equation (17), adding $I^{g}_{i,i+1}$ to its left-hand side is same as to move train path $(i, i+1)$ away so that they can be feasible, thus producing a compressed train pair. Repeat the process so that all the train pairs are feasible. And the utilization of infrastructure capacity by train pairs can all be indicated using methods presented in section 4.

The blocking time graph originally calculated by blocking time theory can be improved by integrating structural indicators into an infrastructure capacity utilization description, abbreviated as ICUD. As can be seen in Fig. 1, denote the two edges of a time square that are parallel to the time axis of the timetable as time edges, and the two edges of time square that are parallel to the distance axis as distance edges. The distance edge of any time square does not concern capacity analysis and therefore is deemed 0.

Denote the blocking time square representing the occupation of infrastructure by train path i in block section j as $U(i, j)$. The weight of the time edge of $U(i, j)$ can be given by

$$L(i, j)^{U} = \sum_{j} (r^{j}_{i} + d^{j}_{i}) + A^{j}_{i} + B^{j}_{i} + C^{j}_{i} + D^{j}_{i} + E^{j}_{i}. \tag{22}$$

Define the time square representing the occupation of infrastructure by train pair

$(i, i+1)$ in block section j as structural time square, which is also suggested by its structural indicator $s_{i,i+1}^{j}$. The weight of the time edge of $V(i, i+1, j)$ can be given by

$$L(i, i+1, j)^{V} = s_{i,i+1}^{j}. \tag{23}$$

With the weight given by equation (22) and (23), relevant information on ICUD is sufficiently provided. And an obvious and useful property of graph ICUD is its strong connectedness. It is easily noticeable that ICUD is uniquely defined by operational inputs (or timetabling data). Notice that the method introduced in this paper should be performed on a relatively integral infrastructure, which is illustrated in section 6. See (Lindner, 2011) for more details.

5.2 Infrastructure Time Allocation

As in Fig. 1, there are several time squares that are neither blocking time square nor structural time square. And notice that train $i-1$ and $i+1$ do not constitute a train pair in block section $j+1$, neither can train $i+1$ and $i+2$ in block section $j-1$ or j.

They are either the product of imperfect timetabling in terms of capacity utilization, or the result of acceptable marketing strategies. And those infrastructure time squares can sometimes be used to operate other trains, and sometimes not. They can be intuitively regarded as infrastructure time fragments. This happens when the operation routes of two or more trains partially overlap or overtakes occur.

Define the time square formed by train path i and $i+1$ in block section j, where train i and $i+1$ do not form a train pair in block section j, as fragment time square, and denote as $W(i, i+1, j)$. This is the reason why compression cannot be conducted partially on certain section of infrastructure, namely the nature of infrastructure time utilization in

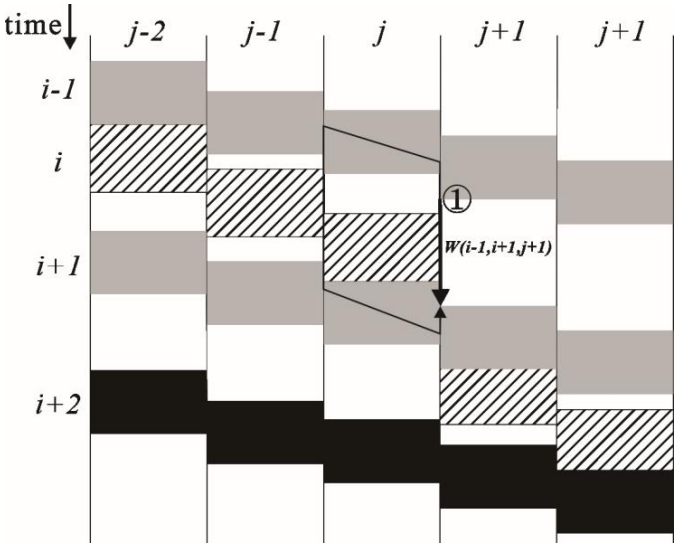


Figure 1: Infrastructure capacity utilization description

railway operations is not all identical. Mixing fragmented infrastructure time with structural indicator only produces meaningless results.

The lower blocking time of train $i-1$ in block section $j+1$ corresponds to time $y_{i-1,j+1}+r_{i-1}^{j+1}+d_{i-1}^{j+1}+D_{i-1}^{j+1}+E_{i-1}^{j+1}$, and the upper blocking time of train $i+1$ in block section $j+1$ corresponds to time $y_{i+1,j+1}-A_{i+1}^{j+1}-B_{i+1}^{j+1}-C_{i+1}^{j+1}$. The weight of the time edge on fragment time square $W(i-1,i+1,j+1)$ can be expressed as

$$L(i-1,i+1,j)^{W}=(y_{i+1,j+1}-A_{i+1}^{j+1}-B_{i+1}^{j+1}-C_{i+1}^{j+1})-(y_{i-1,j+1}+r_{i-1}^{j+1}+d_{i-1}^{j+1}+D_{i-1}^{j+1}+E_{i-1}^{j+1}). \quad (24)$$

A path made of relevant elements, which are calculable using the function system and the given information in ICUD, can be found in ICUD that linking the upper- and lower-time edges of the fragment time square. Path ① calculates the weight of the time edge of $W(i-1,i+1,j+1)$.

Incorporating information on fragment time squares produces an improved ICUD that can better visualize the allocation of infrastructure time.

5.3 Consumed Capacity

Consumed capacity, or capacity consumption, is used to express the total consumption of infrastructure capacity due to certain purpose of calculation. The consumed capacity of an infrastructure during a time period can be expressed as time needed to correspondingly go through the first occupation of the infrastructure till the last occupation of the infrastructure concerned on ICUD.

Like the calculation of infrastructure time fragment, a path can be found linking the upper time edge of time square denoting the first occupation of the infrastructure and the lower time edge of the time square denoting the last occupation of the infrastructure. And all elements can be calculated based on applying function system on given information. As in Fig. 2, a bolder polyline linking the upper edge of *Ub* and the lower edge of *Ue* presents the consumed capacity determined by timetabling data, where *Ub* and *Ue* denote the first

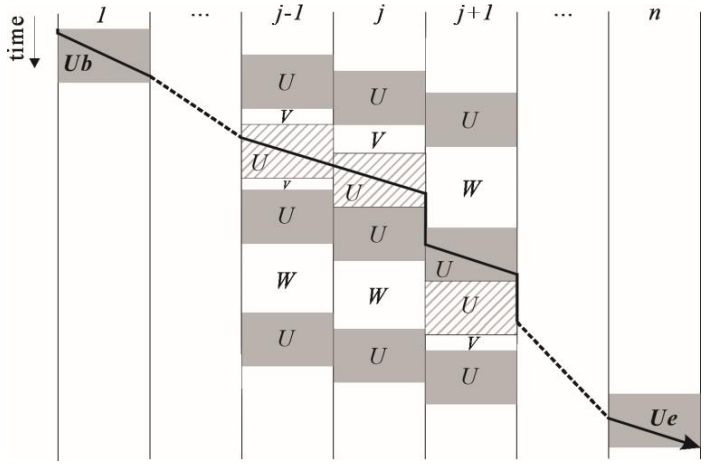


Figure 2: Calculation of consumed capacity

and last occupation of the infrastructure respectively.

As a matter of fact, more than one path of the like can be found. Among them, one path uniquely made up of only blocking time squares and critical block sections of all trains operating on the infrastructure, which we denote as the critical path and denote as P_c. Denote the distance of P_c as $L(P_c)$, and it can also be calculated in a vector-based way.

5.4 Sensitivity Analysis

In order to address the constantly required changes in operational inputs in real operations, impact of operational inputs on consumed capacity should be considered. For that purpose, the connection between capacity utilization and operational inputs must be shown.

Suppose that the operational inputs of train i are changed, which mainly includes (iv) and (vii) as in section 3. Other terms are rarely subject to changes in the short run, which can be dealt with in the same vein. A general procedure is proposed as follows:

a) Renew the sets of feasibility additives, typically set $\{I^j_{i-1,i}\}$ and $\{I^j_{i,i+1}\}$;

b) Renew the critical block sections of relevant train pairs, typically train pair $(i-1,i)$ and $(i,i+1)$;

c) Renew the blocking time squares of train i, structural time squares of relevant train pairs, typically $(i-1,i)$ and $(i,i+1)$, in ICUD;

d) Renew infrastructure time allocation in ICUD;

e) Renew P_c', and calculate $L(P_c')$, where P_c' denotes the renewed critical path.

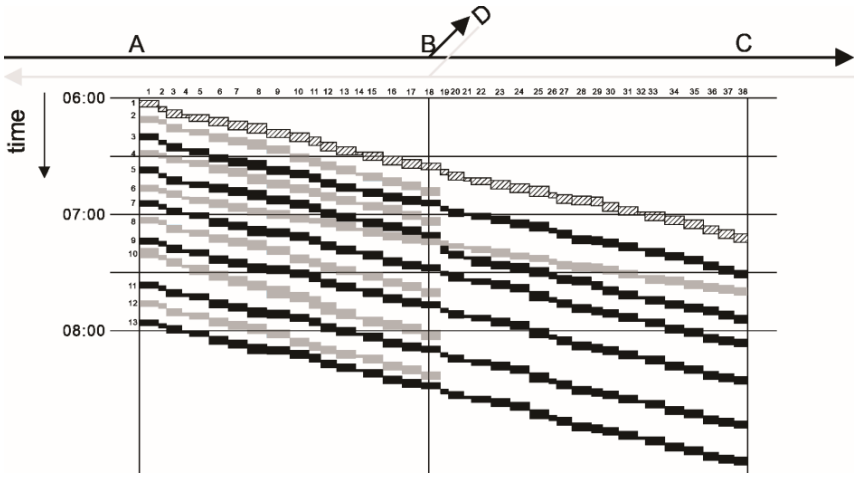
In real application of this method, step d) can be skipped when only $L(P_c')$ is required, since the renewed P_c' share certain section of the original path P_c.

6 Case Study

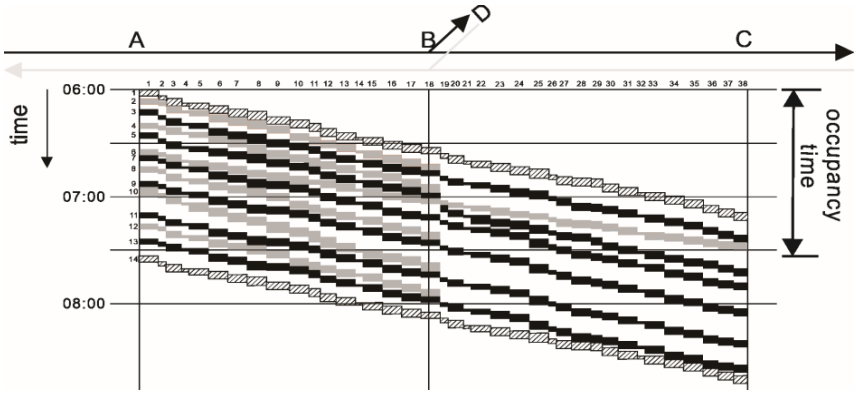
In order to demonstrate the application of proposed method in analyzing railway capacity, including the calculation of consumed capacity and its relationship with relevant parameters, an experimental case in analyzing one direction of a double-track railway infrastructure's capacity is considered in this section.

6.1 Calculation of Consumed Capacity

We consider analysing the capacity utilization of railway infrastructure from A-B-C from 06:00 to 08:00, as shown in Fig. 3 a). As an example, the timetabling follows the basic structure of blocking time theory. Station A, B, and C are terminals, and in between there are intermediate stations that operate passenger transport. As presented in the figure, there is an extra double-track railway line linking station D that is also a terminal, which in real operation causes fragmented use of railway infrastructure. On infrastructure A-B-C, there operate 13 trains of 4 types from 06:00 to 08:00 according to the definition of UIC code



a) Original timetable for infrastructure A-B-C



b) Graphic output of ICUD (compressed timetable)

Figure 3: Capacity analysis of infrastructure A-B-C

406 (UIC, 2013). Regional services 1, 3, 5, 7 ,9 11, 13 operate in section A-B-C. Regional services 2, 4, 8, 12 operate in section A-B-D. Intercity service 5 operates in section A-B-C on the infrastructure. And freight train 10 operates in section A-B on the infrastructure.

The compressed timetable on infrastructure A-B-C is presented in Fig. 3 b). Train 14 acts as the repeated train path of train 1. The first occupation of infrastructure A-B-C is by train 1 in block section 1, and the end of occupation is denoted by the upper blocking time of train 14 in block section 1. The occupancy time in section A-B-C is 93.4 min, which accounts for 77.8% of the chosen period.

Graphic representation of ICUD is a saturated timetable, which is the same as the compressed timetable generated by UIC compression. The difference of ICUD to the compressed timetable is that time edge' weight of all the time squares formed by train 1 to 13 and section 1 to 38 are calculated (which is impossible to show in the picture), presenting the capacity utilization pattern determined by timetabling data.

Using ICUD to formulate the consumed capacity according to the definition of UIC compression, the path found to calculate the consumed capacity when operating timetable shown in Fig. 3 a) can be either one that links the beginning of blocking time square (1,1) and the beginning of blocking time square (14,1). As reported in table 1, the column contribution expresses that respective train's blocking time square path contributes to the overall consumed capacity positively or negatively. The calculation result based on ICUD is $L(P_c) = 93.4\ min$, the same as that from UIC compression. As a matter of fact, the consumed capacity can be viewed as the sum of time components in vectors that denote complete infrastructure occupation. Thus, the calculation process of UIC compression can be regarded as a simplified calculation process using ICUD.

Table 1: Blocking time squares defining critical blocking time path P_c

Train	**Blocking Time Squares**		**Contribution**
	From	To	
1	U(1,1)	U(1,8)	+
2	U(2,8)	U(2,18)	+
3	U(3,18)	U(3,7)	-
4	U(4,7)	U(4,8)	+
5	U(5,8)	U(5,18)	+
6	U(6,18)	U(6,5)	-
7	U(7,5)	U(7,8)	+
8	U(8,8)	U(8,18)	+
9	U(9,18)	U(9,1)	-
10	U(10,1)	U(10,18)	+
11	U(11,18)	U(11,8)	-
12	U(12,8)	U(12,10)	+
13	U(13,10)	U(13,29)	-
14	U(14,29)	U(14,1)	-

Data source of UIC compression is ready timetable, or timetable information such as departure and arrival times at each block sections, while ICUD is based on processing operational inputs using function system. UIC compression generates compressed timetable to determine the infrastructure occupancy so that the utilization rate of the whole infrastructure can be analysed. In the meantime, ICUD comprehensively presents the utilization of railway infrastructure through the distribution of structural time squares and fragmented time squares, which can be used for various purposes.

6.2 Speed and Consumed Capacity

Consider increasing the running time of train 4 within all block sections by 3%, which influences train pair $(3,4)$ and $(4,5)$. Apply the procedure for sensitivity analysis as follows:

a) Renew set $\{I_{3,4}^j\}$ and $\{I_{4,5}^j\}$, where $j \in \{1,\ldots,18\}$;

b) Renew the critical block sections of train pair $(3,4)$ and $(4,5)$, and they are

respectively $g_{3,4} = 6$ and $g_{4,5} = 18$;

c) Renew the blocking time squares of train 4, structural time squares of relevant train pair $(3,4)$ and $(4,5)$, in ICUD;

d) Renew infrastructure time allocation in ICUD;

e) After increasing running time, train 4 contributes more to the total consumed capacity. And the distance of renewed critical path is around 94.3 min, or 78.5% in terms occupancy rate.

The previous analytical process shows that ICUD can present the impact of timetabling parameters on infrastructure utilization as well as on consumed capacity. The advantage of ICUD lies in the unnecessity to repeat the whole analytical process to generate a complete new ICUD. Instead, it is done in a rather limited scope which only involves trains whose timetabling data is changed.

6.3 Recovery Time and Consumed Capacity

Recovery time is added to train running time within a block section. Using equation (18), the feasibility constant is influenced by adding recovery margin. Thus, recovery margin influences the distribution of critical block sections. Since it changes the process times of trains, which is immediately related to the contribution in consumed capacity from that train. Therefore, the real influence of recovery time must be determined through the analytical procedure described in section 5.4. In order to show the impact of recovery margin, we present a comparison of consumed capacity with and without recovery time.

Suppose that evenly-spread regular recovery time addition in every train path is 5%. Now we consider the scenario without added recovery time. The critical path P_c'' without adding recovery time is as shown in table 2.

Table 2: Blocking time squares defining critical blocking time path P_c''

Train	**Blocking Time Squares**		**Contribution**
	From	To	
1	U(1,1)	U(1,8)	+
2	U(2,8)	U(2,18)	+
3	U(3,18)	U(3,7)	-
4	U(4,7)	U(4,18)	+
5	U(5,18)	U(5,18)	+
6	U(6,18)	U(6,5)	-
7	U(7,5)	U(7,8)	+
8	U(8,8)	U(8,18)	+
9	U(9,18)	U(9,1)	-
10	U(10,1)	U(10,18)	+
11	U(11,18)	U(11,8)	-
12	U(12,8)	U(12,18)	+
13	U(13,18)	U(13,29)	-
14	U(14,29)	U(14,1)	-

Based on equation (18), running time in block section influences the calculation of

feasibility constants. In this case, the adding running times only changes the critical block section of train pair (12,13). The consumed capacity without recovery time is 88.8 min, which accounts for the 74.0% within the total 2 hours. In comparison with the consumed capacity with recovery time, a 5% recovery time addition with evenly spread pattern to the timetabling data partake 4.6 min of the total consumed capacity in real timetable, which takes up 4.92% of the total consumed capacity, slightly less than 5%. Therefore, it is easy to conclude that there is no linear correlation between the claimed percentage of recovery time addition and its real influence due to the existence of blocking time elements other than running time in a block section. And it is foreseeable that the percentage representing the real influence will be smaller as more trains are included in the analysis.

7 Conclusion

In this paper, we propose an analytical tool to present capacity utilization of railway infrastructure whose operation is based on blocking time theory. The basic assumption concerning the structure of timetable is that operational inputs, mainly comprised of timetabling data, remain constant during scheduling, thereupon the sensitivity of consumed capacity to operational inputs can be considered. The critical block section of train pair is determined through comparing its feasibility additives in all block sections and can then be used for describing the capacity utilization of a train pair. A simple overtake can be viewed to be composed of several train pair during analysis, while a complex overtake can be analysed by examining the structure of each train pair composing the complex overtake. Infrastructure capacity utilization can be formulated as a graph of distributed blocking time squares and structural time squares, which can be improved by an infrastructure time allocation process that determines fragmented infrastructure usage. Based on ICUD, the overall consumed capacity can be computed, along with the general procedure to analyse the impact of parameter variations on the utilization of infrastructure capacity.

An experimental case study was reported to support the method, in which the differences of this method to UIC compression were demonstrated. Based on the results, ICUD can be used for calculating consumed capacity. And it proved to be a better tool that presents infrastructure capacity utilization when it comes to utilization analysis of railway infrastructure whose operation is dependent on a conflict-free timetable. And parametric connection between operational parameters and consumed capacity was also tested. In the paper, speed and recovery time correspond to running time a in block section. It was demonstrated in both cases that the proposed formulation of ICUD is capable of presenting influence of important parameters owing to the connection between operational inputs (mainly timetabling data) and the utilization of infrastructure.

Acknowledgements

This research was supported by the National Key R&D Program of China (2017YFB1200702), National Natural Science Foundation of China (Project No. 61703351), Sichuan Science and Technology Program (Project No. 2018RZ0078), Science and Technology Plan of China Railway Corporation (Project No.: 2016X006-D), Chengdu Soft Science Research Project (Project No.: 2017-RK00-00028-ZF, 2017-RK00-00378-ZF), Fundamental Research Funds for the Central Universities (2682017CX022, 2682017CX018), Service Science and Innovation Key Laboratory of Sichuan Province (KL1701), and Doctoral Innovation Fund Program of Southwest Jiaotong University (D-

CX201829).

Appendix

Mathematical notations used in this paper are listed as follows:

Notation	Description
i	An arbitrary train operating on the infrastructure
O_i	The first block section on the operation route of train i
D_i	The last block section on the operation route of train i
$[O_i, D_i]$	The operation route of train i
$(i, i+1)$	Train pair comprised of train i and $i+1$
$o_{i,i+1}$	The first block section on the common operation route of train pair $(i, i+1)$
$d_{i,i+1}$	The last block section on the common operation route of train pair $(i, i+1)$
$[o_{i,i+1}, d_{i,i+1}]$	The common operation route of train pair $(i, i+1)$
j	An arbitrary block section on the operation route of train i or train pair $(i, i+1)$
A_i^j	The time needed to set up the signal to operate in block section j for train i
B_i^j	Time needed for the train driver to confirm the signal to approach block section j for train i
C_i^j	The time needed for train i to end block section j
r_i^j	The time needed for train i to cover the whole block section j
D_i^j	The time needed for train i to clear block section j
E_i^j	The time needed to release the signal of block section j after the traverse by train i
d_i^j	The duration of a scheduled stop of train i at station in block section j
$(1, \cdots, i, \cdots, m)$	The sequence of train departing from certain block section of the infrastructure
$(i, i+1) \xrightarrow{j} (i+1, i)$	A change of operation sequence from $(i, i+1)$ to $(i+1, i)$ at station in block section j
$p_i^{O_i, j}$	The entry process time of train i from block section O_i to j
$b_{i,up}^{O_i, j}$	The upper blocking time of train i in block section j from the origin section O_i
$b_{i,low}^{O_i, j}$	The lower blocking time of train i in block section j from the origin section O_i
$t_{i,i+1}^{o_{i,i+1}, j}$	The blocking time difference of train pair $(i, i+1)$ in block section j

Notation	Description
$y_{i,j}$	The departure time of train i from block section j
$m_{i,j}^{low}$	The mark of the lower blocking time of train i in block section j on the time axis
$I_{i,i+1}^{j}$	The feasibility constant of train pair $(i,i+1)$ in block section j
$g_{i,i+1}$	The critical block section of train pair $(i,i+1)$
μ	An arbitrary train that forms a train pair with train i in a complex overtake
$s_{i,i+1}^{j}$	The structural indicator of train pair $(i,i+1)$ in block section j
$U(i,j)$	The blocking time square formed by train i in block section j
$L(i,j)^{U}$	The weight of blocking time square's time edge
$V(i,i+1,j)$	The structural time square formed by train pair $(i,i+1)$ in block section j
$L(i,i+1,j)^{V}$	The weight of structural time square's time edge
$W(i,i+1,j)$	The fragment time square form by train i and $i+1$ in block section j
$L(i,i+1,j)^{W}$	The weight of fragment time square's time edge
P_c	The critical path
$L(P_c)$	The distance of critical path P_c

References

Abril, M., Barber, F., Ingolotti, L., Salido, M. A., Tormos, P., Lova, A., 2008. "An assessment of railway capacity", *Transportation Research Part E: Logistics and Transportation Review*, vol. 44 (5), pp. 774-806. https://doi.org/10.1016/j.tre.2007.04.001

Barber, F., Abril, M., Salido, M.A., Ingolotti, L., Tormos, P., Lova, A., 2007. Survey of automated Systems for Railway Management, Technical Report DSIC-II/01/07, Department of Computer Systems and Computation, Technical University of Valencia.

Burdett, R.L., 2015. "Multi-objective models and techniques for analysing the absolute capacity of railway networks", *European Journal of Operational Research*, vol. 245 (2), pp. 489-505. https://doi.org/10.1016/j.ejor.2015.03.020

Burdett, R.L., Kozan, E., 2006. "Techniques for absolute capacity determination in railways", *Transportation Research Part B: Methodological*, vol. 40 (8), pp. 616-632. https://doi.org/10.1016/j.trb.2005.09.004

Caprara, A., Fischetti, M., Toth, P., 2002. "Modeling and solving the train timetabling problem", *Operations Research*, vol. 50 (5), pp. 851-861. https://doi.org/10.1287/opre.50.5.851.362

Carey, M., Lockwood, D., 1995. "A model, algorithms and strategy for train pathing", *Journal of the Operational Research Society*, vol. 46, pp. 988-1005. https://doi.org/10.1057/jors.1995.136

De Kort, A. F., Heidergott, B., Ayhan, H., 2003. “A probabilistic (max, +) approach for determining railway infrastructure capacity”, *European Journal of Operational Research*, vol. 148(3), pp. 644-661. https://doi.org/10.1016/S0377-2217(02)00467-8

Gille, A., Klemenz, M., Siefer, T., 2008. “Applying multiscaling analysis to detect capacity resources in railway networks”, *WIT Transactions on The Built Environment*, vol. 103, pp. 595-603.

Hansen, I.A. (ed.), 2008. *Railway Timetable & Traffic: Analysis, Modelling, and Simulation*, Eurailpress, Hamburg.

Happel O., 1959. “Sperrzeiten als Grundlage für die Fahrplankonstruktion (Blocking times as basic elements for the railway timetabling)”, *Eisenbahntechnische Rundschau*, vol. 8(2), pp. 79-90.

Harrod, S., 2009. “Capacity factors of a mixed speed railway network”, *Transportation Research Part E Logistics and Transportation Review*, vol. 45(5), pp. 830-841. https://doi.org/10.1016/j.tre.2009.03.004

Higgins, A., 1998. “Scheduling of Railway Track Maintenance Activities and Crews”, *Journal of the Operational Research Society*, vol. 49(10), pp. 1026-1033. https://doi.org/10.1057/palgrave.jors.2600612

Higgins, A., Kozan, E., Ferreira, L., 1996. “Optimal scheduling of trains on a single line track”, *Transportation Research Part B: Methodological*, vol. 30(2), pp.147-161. https://doi.org/10.1016/0191-2615(95)00022-4

UIC (International Union of Railways), 2004. CODE 406: Capacity, Paris, France.

UIC (International Union of Railways), 2013. CODE 406: Capacity, 2nd Edition, Paris, France.

Jamili, A., 2018. “Computation of practical capacity in single-track railway lines based on computing the minimum buffer times”, *Journal of Rail Transport Planning & Management*, vol. 8(2), pp. 91-102. https://doi.org/10.1016/j.jrtpm.2018.03.002

Jensen L.W., Landex A., Nielsen O. A., Kroon L. G., Schmidt M., 2017. “Strategic assessment of capacity consumption in railway networks: framework and model”, *Transportation Research Part C: Emerging Technologies*, no. 74, pp. 126-149. https://doi.org/10.1016/j.trc.2016.10.013

Kettner, M., Sewcyk, B., Eickmann, C., 2003. Integrating microscopic and macroscopic models for railway network evaluation. In: *European Transport Conference (Straburg2003)*, Strasburg, France.

Kroon, L., Maróti, G., Helmrich, M.R., Vromans, M., Dekker, R., 2008. “Stochastic improvement of cyclic railway timetables”, *Transportation Research Part B: Methodological*, vol. 42(6), pp. 553-570. https://doi.org/10.1016/j.trb.2007.11.002

Lai, Y.C., Barkan, C. P.L., 2009. “Enhanced parametric railway capacity evaluation tool”, *Transportation Research Record Journal of the Transportation Research Board*, vol. 2117, pp. 33-40. doi:10.3141/2117-05

Lai, Y.C., Barkan, C. P.L., 2011. “A comprehensive decision support framework for strategic railway capacity planning”, *Journal of Transportation Engineering*, vol. 137(10), pp. 738-749. https://doi.org/10.1061/(ASCE)TE.1943-5436.0000248

Lai, Y.C., Huang, Y.A., Chu, H.Y., 2014. “Estimation of rail capacity using regression and neural network”, *Neural Computing and Applications*, vol. 25 (7-8), pp. 2067-2077. https://doi.org/10.1007/s00521-014-1694-x

Landex, A., 2008. Methods to estimate railway capacity and passenger delays. Technical University of Denmark (DTU), Kgs. Lyngby, Denmark.

Landex, A., 2009. “Evaluation of railway networks with single track operation using the UIC 406 capacity method”, *Networks & Spatial Economics*, vol. 9 (1), pp. 7-23. https://doi.org/10.1007/s11067-008-9090-7

Landex, A., Kaas, A. H., Schittenhelm, B., Schneider-Tilli, J., 2006. “Practical use of the UIC 406 capacity leaflet by including timetable tools in the investigations”, *WIT Transactions on the Built Environment*, vol. 1, pp. 643-652.

Lindner, T., 2011. “Applicability of the analytical UIC code 406 compression method for evaluating line and station capacity”, *Journal of Rail Transport Planning & Management*, vol. 1(1), pp. 49-57. https://doi.org/10.1016/j.jrtpm.2011.09.002

Lindner, T., Pachl, J., 2010. “Recommendations for Enhancing UIC Code 406 Method to Evaluate Railroad Infrastructure Capacity”, In: *Transportation Research Board 89th Annual Meeting (Washington2010)*, Washington, USA.

Marinov, M. Viegas, J., 2011. “A mesoscopic simulation modelling methodology for analyzing and evaluating freight train operations in a rail network”, *Simulation Modelling Practice and Theory*, vol. 19(1), pp. 516-539.

Medeossi, G., Longo, G., Fabris, S.D., 2011. “A method for using stochastic blocking times to improve timetable planning”, *Journal of Rail Transport Planning & Management*, vol. 1(1), pp. 1-13. https://doi.org/10.1016/j.jrtpm.2011.07.001

Pachl J., 2018. *Railway operation and control*, 4th Edition, VTD Rail Publishing, WA United States.

Petering, M.E., Heydar, M., Bergmann, D.R., 2015. “Mixed-integer programming for railway capacity analysis and cyclic, combined train timetabling and platforming”, *Transportation Science*, vol. 50(3), pp. 892-909. https://doi.org/10.1287/trsc.2015.0652

Pouryousef, H., Lautala, P. White, T., “Railroad capacity tools and methodologies in the U.S. and Europe” *Journal of Modern Transportation*, vol. 23(1), pp. 30–42. DOI 10.1007/s40534-015-0069-z

Riejos, F. A. O., Barrena, E., Ortiz, J. D. C., Laporte, G., 2015. “Analyzing the theoretical capacity of railway networks with a radial-backbone topology”, *Transportation Research Part A: Policy and Practice*, vol.84, pp. 83-92. https://doi.org/10.1016/j.tra.2015.03.018

Sameni M.K., Dingler M., Preston J.M., Barkan C. P.L., 2011. Profit-generating capacity for a freight railroad. In: *Transportation Research Board 90th Annual Meeting (Washington2011)*, Washington, USA.

Yan Y. S., 1997. “A study on the calculating method for removal coefficient of passenger trains on double track railway with automatic block system”, *Journal Southwest Jiaotong University*, vol. 32, pp. 11-15.

Yaghini M., Nikoo N., Ahadi H.R., 2014. “An integer programming model for analysing impacts of different train types on railway line capacity”, *Transport*, vol. 29 (1), pp: 28-35.

Zhang, J.M., 2015. “Analysis on line capacity usage for china high speed railway with optimization approach”, *Transportation Research Part A: Policy and Practice*, vol. 77, pp. 336-349. https://doi.org/10.1016/j.tra.2015.04.022

Zhao L.Z., 2001. “Calculation and analysis of carrying capacity of high-speed railway section”, *China Railway Science*, vol. 22 (6), pp. 54-58.