# A Variogram-Based Tool for Variable Selection in a Wastewater Treatment Effluent Prediction

Markku Ohenoja Jani Tomperi

Control Engineering, University of Oulu, Finland, forename.surname@oulu.fi

## Abstract

In this study, a variogram method was utilized as a variable selection tool for finding the optimal subsets of variables for developing predictive models for the quality of wastewater treatment effluent. The quality of effluent was here assessed by biological and chemical oxygen demand and suspended solids in biologically treated wastewater. The dataset included, in addition to traditional process measurements, results of a novel optical monitoring device which was used for imaging an activated sludge process in-situ during a period of over one year. The study showed that the variogram based method has potential in fast and computationally easy variable selection. The developed models can be used for proactive monitoring and estimating the quality of effluent in several stages hours before in comparison to laboratory analysis taken from treated wastewater.

*Keywords: activated sludge process, BOD, COD, cross-validation, modeling, optical monitoring, suspended solids, variogram*

## 1 Introduction

While the amount of produced wastewater is increasing, the regulations for the quality of discharges by authorities are constantly tightening, and the operating costs are necessary to be minimized. Wastewaters are commonly treated in biological activated sludge processes (ASP), which are sensitive to external and internal disturbances, such as changing temperature, and varying quality and quantity of wastewater. Disturbances affect the bacterial balance of biomass and the optimum operating conditions, which are in a key role for a high pollution removal rate, low suspended solids in the effluent and a good settling properties of the sludge. Disturbances in the bacterial balance may have serious environmental and economic effects as they often produce dysfunctional flocculation and settling. The most common problem in ASP is filamentous bulking, which is caused when the secondary settler is unable to efficiently remove the suspended biomass from the wastewater. Recovery from the occurred disturbances is slow and the effects on process operation and purification result are long-lasting. (Tchobanoglous *et al*, 2003; Amaral, Ferreira, 2005; Mesquita *et al*, 2009)

On this account, an accurate operating of the wastewater purification process is required. The performance of a wastewater treatment process can be assessed analyzing the quality parameters of treated wastewater, such as biological and chemical oxygen demand (BOD, COD), suspended solids (SS), and sludge volume index (SVI). However, these parameters only show the poor quality of effluent when it already occurs and the corrective operations are inevitably late. Thus, there is a demand for new real-time monitoring tools and methods to be used in process control in parallel with the traditional offline analysis of wastewater samples and expert knowledge. The novel on-line optical monitoring method gives fast, objective information about the state of the wastewater treatment process, reveals some of the reasons for settling problems, and combined to a predictive model, shows the quality of effluent in advance (Koivuranta *et al*, 2015; Tomperi *et al*, 2017). In this study, a variogram method is utilized for finding the optimal subset of variables to develop predictive models for BOD, COD, and SS in biologically treated wastewater. The dataset from a period over one year included the results of the in-situ optical monitoring of an ASP, and the offline process measurements.

## 2 Material and methods

### 2.1 Wastewater Treatment Plant

The data used in this study was collected from the largest wastewater treatment plant (WWTP) in Finland, located in Helsinki. Viikinmäki WWTP processes daily 270,000 $m^3$ of wastewater from over 800,000 inhabitants around the Helsinki region. Part of the total flow (15%) come from industrial sources. This WWTP is a three-phased activated sludge process that utilizes the simultaneous precipitation method for phosphorus removal. Wastewater is processed in nine activated sludge process lines. In addition to mechanical, biological, and chemical treatment, a biological filter has been added to improve nitrogen removal. The unit operations of the process are intake, screening, grit, and grease removal, preliminary settling, aeration, degassing, secondary settling, biological de-nitrification filtration, and discharge (Figure 1). Screening removes the large solids from the water. Grit and grease removal separates rapidly settling, very coarse solids, as well as, greasy and oily substances that are lighter than water. In

the preliminary settling phase, easily settling material is separated from the water. The biological treatment is conducted by means of a de-nitrification-nitrification process in an aeration tank which is used to grow activated sludge. At the head of the aeration tank, there is a separate mixing area, where new wastewater entering the tank is reseeded with returned activated sludge from the secondary settling tank, and recycled sludge from the end of the aeration tank. Activated sludge, biomass which contains organic matter and nutrients, is separated from the treated wastewater by settling in the secondary settling tank and returned to the aeration tank. Part of the activated sludge is removed daily to maintain a suitable sludge age and sludge concentration in the aeration tank. After the secondary settling phase, wastewater is led to filtration based on bacterial action to enhance de-nitrification of the wastewater. (HSY, 2016)

## 2.2 Optical Monitoring and Image Analysis

To replace the slow, irregular, and subjective manual microscopic analysis of wastewater samples, a small-scale optical monitoring device and an image analysis method were developed (Koivuranta *et al*, 2013) and proved functional for monitoring the floc morphology reliably in-situ in full-scale municipal ASP (Koivuranta *et al*, 2015). The device consists of an imaging unit, a sample handling unit, and a control PC with an electronics unit. Wastewater samples were taken from one activated sludge line in the aeration tank, diluted, and pumped through a cuvette, which was imaged with a high-resolution charge-coupled device (CCD) camera. At normal flow, the delay between optical monitoring measurement and the output of the WWTP was about 13 hours. The optical monitoring device measured several morphological features of the flocs and filaments: in addition to the size parameters such as mean equivalent diameter, floc area, and filament length, the calculated shape parameters included, for example, fractal dimension, form factor, and roundness. The parameters were calculated as an average of the values for individual objects over a single image. The detailed description of the device and mathematical formulas of the calculated parameters are presented in (Koivuranta *et al*, 2013).

## 2.3 Variable Selection Using Variogram

Modern plants produce large amounts of data which often include irrelevant variables for a specific purpose, for instance modeling. Only significant input variables should be selected for model development. The greater number of variables does not necessary mean better prediction results because correlated, noisy and uninformative input variables increase the computational complexity, make the training of the model more difficult and worsen the prediction result. Over-fitting may occur if the model contains too many variables which are fitted not only to the data but also to the random noise. Additionally, the sampling rates of different input variables may differ significantly, therefore describing the process dynamics in different precisions.

In this work, a variogram-based method is utilized as a variable selection method in order to find the optimal subsets for modeling the suspended solids content, BOD, and COD in biologically treated wastewater. The idea of utilizing variogram for variable selection comes from the fact that a variogram of particular measurement holds the information about the relative error levels of the sampling and analysis of that measurement. Variogram is a fundamental tool within Theory of Sampling (Gy, 2004) and has already been considered in drift estimation (Paakkunainen *et al*, 2007), temporal uncertainty propagation (Jalbert *et al*, 2011), fault diagnosis (Kouadri *et al*, 2012), statistical process control (Minnit, Pitard, 2008), and as a process stability measure (Bisgaard, Kulachi, 2005).

Variogram is calculated from a set of systemically collected data. In this work, it is assumed that the data is systemically sampled and that the flow rate, or sample weight, is constant. Hence, the heterogeneity of the data can be interpreted as:

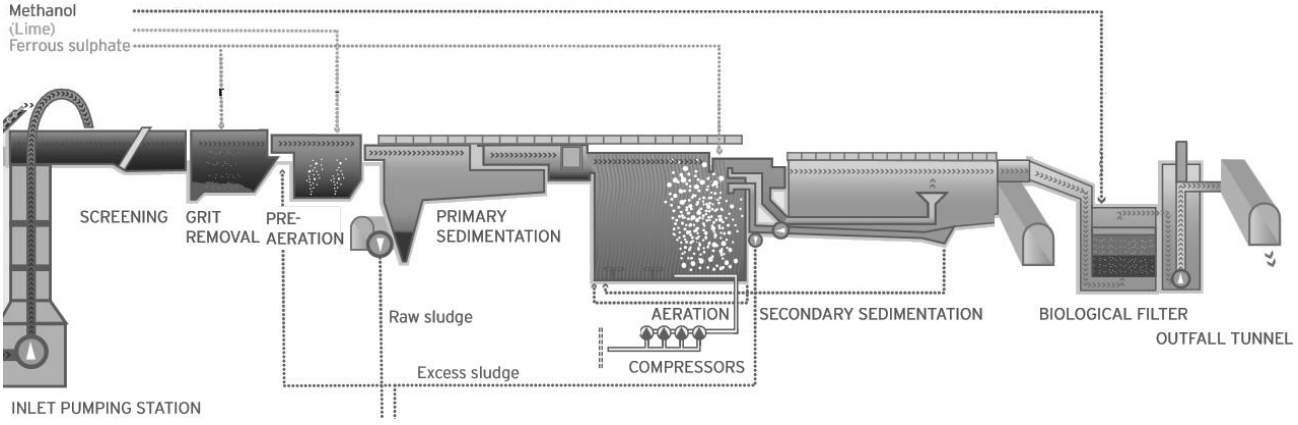


**Figure 1.** The wastewater treatment process at Viikinmäki. Modified from (HSY, 2016).

$$h_i = \frac{x_i - \bar{x}}{\bar{x}} \quad (1)$$

where $h_i$ and $x_i$ are the heterogeneity and the measurement result for sample $i$, respectively and $\bar{x}$ is the average of measurements $x_i$. The semi-variogram is calculated as:

$$v(j) = \frac{1}{2(N - j)} \sum_{i=1}^{N/2} (h_{i+j} - h_j)^2 \quad (2)$$

where $v(j)$ is the relative standard error between samples collected with lag $j$ and $N$ is the number of samples in the data set. The intercept $v(0)$ is estimated based on a linear extrapolation of the first $N/10$ (floored) points of the variogram. The index describing the relative information content of the measurement and thus the criterion for variable selection is calculated as relation between the estimated sampling error $v(1)$ and process variability $s_P$:

$$I = \frac{v(1)}{s_P} \quad (3)$$

A low value of the index $I$ indicates that a single sample of that measurement can describe the present process variability with good accuracy. On the other hand, high value for $I$ indicate that the relative information content of the measurement is low either due to higher sampling error or lower variability in the process.

### 2.4 Modeling

A k-fold cross-validation is a typical resampling method for predicting the fit of a model for a validation set, when dataset is small, and the split to separate training and validation subsets is not possible without a significant loss of data. Efficient training and validation require long and representative subset of data for both. In environmental related processes, the source dataset for model training should also encompass at least one full year of measured data because the temperature and rainfall, for instance, change depending on the season of the year and affect the process. In k-fold cross-validation, the original dataset is randomly partitioned into $k$ subsets of equal size. One subset is used as a validation data for testing the model and the remaining $k$–1 subsamples are used as training data. The cross-validation process is repeated $k$ times and each of the subsets is used only once as the validation data. A single estimation is then produced by combining these $k$ results of the folds. Optimal $k$ is often reported being between five and ten folds because statistical performance does not increase notably for larger values of $k$, and averaging over less than ten splits is computationally feasible. In this study, five-fold cross-validation was used. Multivariable linear regression (MLR) was used to predict an output variable as a linear combination of selected input variables as:

$$Y = b_0 + b_1 X_1 + b_2 X_2 + ... + b_n X_n + e \quad (4)$$

where $b_o$ is a constant value, $b_1 \ldots b_n$ are the $n$ regression coefficients, $X_1 \ldots X_n$ independent variables and $e$ is the error. The performance of the model was evaluated by using Root Mean Square Error (RMSE) and coefficient of determination ($R^2$), which can be used to compare the relative performance of the models. (Rao *et al*, 2008; Arlot, Celisse 2010)

## 3 Results and Discussion

The dataset used in this work consisted of optical monitoring results and wastewater treatment process measurements from a period of over one year. On-line optical monitoring measurements were carried out at least once a day, but the laboratory measurements, on the other hand, were done only two to three times a week. During the process maintenance stoppages or occasional problems with the device, the optical measurements could not be performed. The missing laboratory and on-line data was not interpolated in this study. Thus, the total number of data points was 94 observations for 50 variables. Measurement data was scaled to range [-2, 2] before variable selection as in (Tomperi *et al*, 2017). Only variables that are useful and reliable to measure were selected. The variables from as early stage of the process as possible were preferred in order to establish models which could give proactive information of the quality of biologically treated wastewater.

**Table 1.** Variable selection using variogram.

| Variable | Value of criterion |
|---|---|
| Fractal dimension[1] | 0.17 |
| Aspect ratio[2] | 0.18 |
| Temperature[3] | 0.22 |
| Median area of objects[4] | 0.25 |
| Filament length[5] | 0.27 |
| Roundness[6] | 0.30 |
| Sludge age[7] | 0.31 |
| Amount of filaments | 0.32 |
| Suspended solids | 0.33 |
| Number of small objects | 0.34 |

In this study, variogram was utilized as a variable selection tool for searching the optimal subset of variables for model development. The variogram-derived indices and ten first selected variables are presented in Table 1. As seen, the most of the variables are on-line optical monitoring variables and only three of ten variables are process measurements. In comparison, five other variable selection methods tested

in (Tomperi *et al*, 2017), resulted as a suspended solids model with only one on-line optical monitoring variable (fractal dimension) and six process measurements (influent total nitrogen and sulphate, mechanically treated wastewater iron and nitrate nitrogen, temperature and anoxic proportion). Plausible reason is the lower variance of the optical monitoring variables, which shows in variogram as lower error-estimate and lower value of criterion. The earlier data analysis also showed that the quality parameters of biologically treated wastewater (BOD, COD, SS) have high mutual correlation and follow the changes of the temperature: the quality of treated wastewater was good in summer time when wastewater was warmer. The optical monitoring parameters also have several mutual correlations. For example, at summer time the amount and length of filaments was low, flocs were larger, the roundness of flocs was higher and the number of objects was lower (Tomperi *et al*, 2017). Hence, several variables in Table 1 have high mutual correlations which affect the results of model validation.

Seven input variables (n=7, 1-7 in Table 1) were selected for developing linear models for suspended solids, BOD and COD in biologically treated wastewater. The fitness of the models was predicted using 5-fold cross-validation. The results of modeling, the $R^2$ and RMSE values, and the regression coefficients of each developed model, are presented in Table 2. The results presented here can be considered satisfactory although in (Tomperi *et al*, 2017) the $R^2$ values of the SS model were between 0.79 (received using the genetic algorithm subset variable selection) and 0.71 (received using the correlation based variable selection), and the RMSE values were between 0.47 and 0.55. In the same study, the $R^2$ values of the BOD model were between 0.55 and 0.45, and $R^2$ values of the COD model were between 0.56 and 0.45.

The performance of variogram-based model for suspended solids in biologically treated wastewater is presented in Figure 2, together with the correlation-based model from the earlier study (Tomperi *et al*, 2017). All models developed using input variables selected by the variogram method have the most difference to the measured value of BOD, COD and SS at the same point: between 10-20 data point and around 40 and 75 data point. The visual interpretation also indicates that the modeling results with the variogram based variable selection contains less fast fluctuations.

The results of this study show that the variogram based tool has potential in selecting input variables for developing predictive models of treated wastewater quality even though the performance of the models was not as high as in the earlier study. Expert knowledge is required to improve the performance of the models. However, it should also be noted that the computational effort of variogram-based variable selection was minimal (less than 0.5 sec.) and implementation of the method was considerably easier than for example with genetic algorithm and successive projections algorithm, whose computational time was tens of minutes. Although the variogram-based variable selection has limited performance in the tested dataset, the method is seen interesting as it could also be developed into a recursive variable selection method due to its computational performance.

**Table 2.** The modeling results and the regression coefficients of input variables

| Variable | $R^2$ | RMSE | $b_0$ | $b_1$ | $b_2$ | $b_3$ | $b_4$ | $b_5$ | $b_6$ | $b_7$ |
|---|---|---|---|---|---|---|---|---|---|---|
| BOD | 0.45 | 0.71 | -0.64 | 0.47 | 0.67 | 0.09 | -0.08 | 0.60 | 0.12 | -0.11 |
| COD | 0.37 | 0.75 | -0.35 | -0.43 | 0.39 | 0.22 | 0.02 | 0.37 | 0.49 | -0.13 |
| SS | 0.60 | 0.64 | -0.61 | -0.47 | 0.33 | -0.25 | 0.15 | 0.52 | 0.67 | -0.20 |

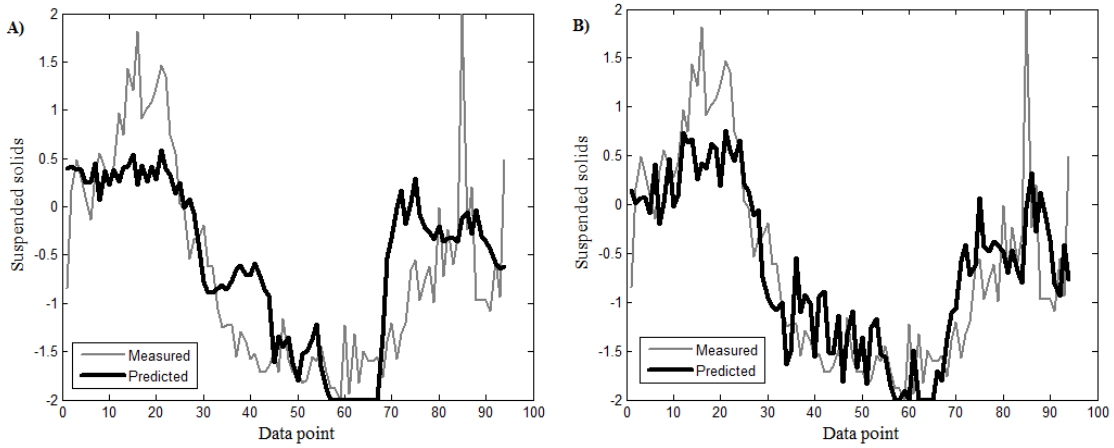


**Figure 2.** Measured and predicted suspended solids in biologically treated wastewater as scaled values, A) variogram based selected variables, B) correlation based selected variables (Tomperi *et al*, 2017).

# 4 Conclusions

In this study, a variogram method was utilized as a variable selection tool. Selected variables were used as input variables in predictive models of BOD, COD and suspended solids, which are important and critical quality parameters of the wastewater treatment process efficiency. Dataset included process measurements and the results of a novel optical monitoring method from a period of one year. Five-fold cross-validation was used to evaluate the performance of the developed models.

The presented results of variable selection show that the variogram based tool has potential in selecting input variables for developing predictive models of treated wastewater quality even though the fitness of the developed models was not as high as in the earlier study. The variogram method is, however, easier to implement and faster to use than some traditional variable selection methods. Nevertheless, the results can be considered satisfactory and the developed models can be used for proactive monitoring and estimating the quality of treated wastewater in several stages hours before in comparison to laboratory analysis taken from the treated water.

## Acknowledgements

This research was carried out as part of the Measurement, Monitoring and Environmental Efficiency Assessment (MMEA), the research program of CLEEN Ltd. – Cluster for Energy and Environment.

Ms. Elisa Koivuranta (Fibre and Particle Engineering, University of Oulu) and Ms. Anna Kuokkanen (Helsinki Region Environmental Services Authority) are acknowledged for producing the original data used in this study.

## References

A.L. Amaral and E.C. Ferreira. Activated sludge monitoring of a wastewater treatment plant using image analysis and partial least squares regression. *Analytica Chimica Acta*, 544:246–253, 2005. doi: 10.1016/j.aca.2004.12.061.

S. Arlot and A. Celisse. A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4:40–79, 2010. doi: 10.1214/09-SS054.

S. Bisgaard and M. Kulahci. Checking process stability with the variogram. *Quality Engineering*, 17:323–327, 2005. doi: 10.1081/QEN-200056505.

P. Gy. Sampling of discrete materials: III. Quantitative approach—sampling of one-dimensional objects. *Chemometrics and Intelligent Laboratory Systems*, 74:39–47, 2004. doi: 10.1016/j.chemolab.2004.05.011.

HSY Viikinmäki wastewater treatment plant webpage, May 2016 https://www.hsy.fi/en/experts/water-services/wastewater-treatment-plants/viikinmaki/Pages/default.aspx.

J. Jalbert, T. Mathevet. and A. Favre. Temporal uncertainty estimation of discharges from rating curves using a variographic analysis. *Journal of Hydrology*, 397:83–92, 2011. doi: 10.1016/j.jhydrol.2010.11.031.

E. Koivuranta, J. Keskitalo, A. Haapala, T. Stoor, M. Sarén and J. Niinimäki. Optical monitoring of activated sludge flocs in bulking and non-bulking conditions. *Environmental Technology*, 34:679–686, 2013. doi: 10.1080/09593330.2012.710410.

E. Koivuranta, T. Stoor, J. Hattuniemi and J. Niinimäki. On-line optical monitoring of activated sludge floc morphology. *Journal of Water Process Engineering*, 5:28–34, 2015. doi: 10.1016/j.jwpe.2014.12.009.

A. Kouadri, M. Aitouche and M. Zelmat. Variogram-based fault diagnosis in an interconnected tank system. *ISA Transactions*, 51:471–476, 2012. doi: 10.1016/j.isatra.2012.01.003.

D.P. Mesquita, O. Dias, A.L. Amaral and E.C. Ferreira. Monitoring of activated sludge settling ability through image analysis: validation on full-scale wastewater treatment plants. *Bioprocess Biosyst Eng*, 32:361–367, 2009. doi: 10.1007/s00449-008-0255-z.

R. Minnitt. and F. Pitard. Application of variography to the control of species in material process streams: %Fe in an iron ore product. *Journal of SAIMM*, 108:109–122, 2008.

M. Paakkunainen, S. Reinikainen and P. Minkkinen. Estimation of the variance of sampling of process analytical and environmental emissions measurements. *Chemometrics and Intelligent Laboratory Systems*, 88:26–34, 2007. doi: 10.1016/j.chemolab.2006.11.001.

R.B. Rao, G. Fung and R. Rosales. On the dangers of cross-validation: An experimental evaluation. *Proceedings of the 2008 SIAM International Conference on Data Mining*, Atlanta, GA, pp. 588–596, 2008. doi: 10.1137/1.9781611972788.54.

G. Tchobanoglous, F.L. Burton and H.D. Stense. *Wastewater Engineering: Treatment and Reuse*, 4th ed., Boston: McGraw-Hill, 2003.

J. Tomperi, E. Koivuranta, A. Kuokkanen and K. Leiviskä. Modelling effluent quality based on a real-time optical monitoring of the wastewater treatment process. *Environmental Technology*, 38:1–13, 2017. doi: 10.1080/09593330.2016.1181674.