Comparison of modelling accuracy with and without exploiting automated optical monitoring information in predicting the treated wastewater quality

Tomperi Jani, Leiviskä Kauko

Control Engineering, University of Oulu. Oulu, Finland.

Corresponding author: e-mail jani.tomperi@oulu.fi. University of Oulu, Control Engineering, P.O. Box 4300, FI-90014 University of Oulu, Finland.

Comparison of modelling accuracy with and without exploiting automated optical monitoring information in predicting the treated wastewater quality

Traditionally the modelling in an activated sludge process has been based on solely the process measurements but as the interest to optically monitor wastewater samples to characterize the floc morphology has increased, in the recent years the results of the image analyses have been more frequently utilized to predict the characteristics of wastewater. This study shows that the traditional process measurements or the automated optical monitoring variables by themselves are not capable of developing the best predictive models for the treated wastewater quality in a full-scale wastewater treatment plant, but utilizing these variables together the optimal models, that show the level and changes of the treated wastewater quality, are achieved. By this early warning, process operation can be optimized to avoid environmental damages and economic losses. The study also shows that specific optical monitoring variables are important in modelling a certain quality parameter regardless of the other input variables available.

Keywords: activated sludge process; cross-validation; image analysis; variable selection; wastewater treatment

Introduction

The main purpose of the wastewater treatment is to efficiently remove oxygen demanding substances, excessive nutrients and toxicants from treated wastewater that can be reused or discharged to waterways nearby. Wastewaters are most commonly treated in complex biological activated sludge processes (ASPs) where the operation of the treatment process is affected by several physical, chemical and microbiological factors. The key element for the efficient operating of an ASP is a good bacterial balance in biomass, which is very sensitive to internal and external disturbances like major changes in the quality and quantity of influent. The disturbances in the bacterial balance are most often shown as dysfunctional flocculation and settling properties and

the recovery from the disturbances is slow. This causes long-lasting challenges to process control and possible serious environmental effects as low quality effluent is discharged to waterways.

As the limitations to effluent discharges defined by the authorities are stringent and operating costs are constantly rising, more attention must be focused on the optimal operating of the wastewater treatment process. The efficiency of the wastewater treatment process can be assessed by measuring some physical, chemical and biological quality parameters from the effluent. Suspended solids (SS), eutrophication in waterways causing nitrogen (N) and phosphorus (P), and chemical oxygen demand (COD) and biochemical oxygen demand (BOD) that indicate the amount of dissolved oxygen required to oxidize the organic substances in wastewater are used traditionally to assess the quality of treated wastewater. [1]

The information received by the conventional measurements of the wastewater treatment process is not adequate to give an early warning of changes in the treated wastewater quality, which is essential in optimizing the process control and avoiding environmental damages. In a full-scale wastewater treatment plant (WWTP), on-line optical monitoring of floc morphological characteristics gives fast objective information about the quality of wastewater and the state of the treatment process, reveals the reasons for settling problems, and combined to predictive modelling shows the quality of the treated wastewater in advance hours before problems occur and would be noticed by traditional process measurements [2, 3, 4].

Due to the increasing interest towards the optical monitoring of floc morphology in activated sludge processes several measuring devices and methods have been developed into that purpose [5] and the utilization of the optical monitoring results in modelling the characteristics of wastewater treatment has grown. However, the

modelling of a complex nonlinear wastewater treatment process is very challenging. Since developing the activated sludge model 1 (ASM1) [6], several modelling techniques have been utilized to develop models related to active sludge processes [7]. For example, ASM1 modifications and artificial neural network (ANN) models have been used for modelling full-scale WWTPs [8, 9, 10] and the performances of WWTPs [11, 12]. Several studies have also utilized Partial Least Squares (PLS) technique to correlate the quantitative image analysis (QIA) data and parameters of a wastewater treatment process, sludge volume index (SVI) and total suspended solids (TSS) [13, 14, 15, 16].

In this study, predictive models for the treated wastewater quality parameters (BOD, COD, SS, N, and P) of a full-scale WWTP are developed first based on solely the variables of in-situ real-time optical monitoring of the wastewater, and secondly based on solely the process measurements. The results of these developed models are compared with the results of the earlier study where both the optical monitoring variables and the process measurements were together utilized in developing the quality parameter models. The purpose is to show how utilizing both datasets for predicting the quality of treated wastewater in a full-scale WWTP improves the performance of the model. It is also shown that certain optical monitoring variables are always important in developing a model for a quality parameter. The optimal subsets of input variables for the models are sought using five variable selection methods that are shortly presented.

Material and methods

The wastewater treatment plant

The data were collected from the largest WWTP in Finland, which processes daily in average 270,000 m^3 of wastewater from over 800,000 inhabitants and industrial sources

nearby. The treatment process is a three-phased activated sludge process that utilizes simultaneously a precipitation method for phosphorus removal. The 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. At normal flow, the delay between the aeration tank, where the optical monitoring was performed, and the output of the WWTP is about 13 hours. [17]

On-line optical monitoring and image analysis

A novel automatic optical monitoring device was developed to replace the traditional laborious, slow and subjective method to study wastewater samples under a microscope [18]. The small-scale monitoring device has been proved functional for reliable in-situ on-line monitoring of the floc morphology at a full-scale WWTP during the test period of several months [2, 4]. 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 of the nine activated sludge lines in the aeration tank, diluted and pumped through a cuvette, which was imaged with a high-resolution charge-coupled device (CCD) camera. A 1:100 dilution ratio was used in the on-line measurements because the laboratory test indicated that the use of different dilutions in the on-line imaging system does not affect the flocs and the image analysis results can be considered reliable even though the dilution at on-line is not as accurate as done in laboratory conditions [2]. The sensor of CCD camera is 5.5*3.7 mm (1392*1040 pixels) with a pixel size of 3.6*3.6 µm.

The automatic optical monitoring device measures several morphological features of the flocs and filaments. In addition to size parameters such as mean equivalent diameter, floc area and filament length, the calculated shape parameters includes, among others, the mass fractal dimension, form factor and roundness. The

parameters were calculated as an average of the values for individual objects over a single image. One analysed video contained about 1000 images and one image contained 150 flocs on average [2]. Thus, the obtained results of wastewater samples can be considered statistically reliable.

In the presented results, the amount of filaments is a ratio of filament length and floc area, the total filament length is the sum of the length of all filaments present in the image, and the number of small objects is calculated based on the size distribution where each object is assigned to a size category based on its equivalent diameter. The size distribution was calculated as the sum of the distributions of individual images. The case specific floc area threshold value for the calculated objects was 100 μm^2 because the boundaries of smaller objects may not have been sharp enough due to the resolution of the camera. The limit value for small objects was set in an equivalent diameter of under 25 µm. The mathematical formulas and more details of the calculated size and shape parameters are presented in [18].

Data pre-treatment

During the data collection period of over a year, the optical monitoring was carried out at least once a day but some process measurements were recorded only two to three times a week. These datasets were combined by date but the missing values were not interpolated because approximation always dilutes the results of analysis. Only the data from dates including all measurements were utilized in the study. Thus, the total amount of data samples was only 94 during a period of over one year. Before variable selection and modelling the data were scaled between [-2, 2] using a nonlinear scaling method based on generalized moments, norms and skewness presented in [19].

Variable selection

In modelling, using input variables that include noise, are correlated to each other or have no significant relationship with the output variable only increase the computational complexity and reduce the prediction result of a model. The amount of input variables should be kept decent because using too many input variables increases the risk to develop an over-fitted model which has an excellent training results but is not usable for prediction with new data.

In this work, five variable selection methods were used to select the optimal subsets of input variables to develop models for quality parameters (BOD, COD, SS, N, P) of treated wastewater utilizing solely the optical monitoring variables or the process measurements. Variable selection methods were correlation based selection, forward selection, stepwise selection, genetic algorithm (GA), and a successive projections algorithm (SPA) combined with a genetic algorithm, which are presented in detail in [4].

In correlation based selection, variables are selected by the absolute value of their correlation coefficient. Correlation coefficients are calculated and inspected to find variables that have a mutual correlation over |0.85|. From every found variable pair, the variable with a lower correlation coefficient is removed from the dataset and rest of the variables are arranged in downward order by their absolute correlation coefficient with the output variable.

A forward selection method adds one (the best) variable at a time to the model. Adding is continued until the performance of the model does not improve and the best combination of the variables is selected. No variables are removed from the variable subset once they are selected and thus the variables whose performance is strong together with other variables but poor alone are not selected due to the single selection

principle. [20, 21] A stepwise regression is a modified forward selection method, which adds the best variable to a variable subset or deletes the worst variable from a variable subset at each round. Adding and deleting is based on variable's statistical significance in regression. [22]

A successive projections algorithm is a forward selection method in multivariate calibration. SPA uses simple operations in the vector space to minimize collinearity between selected variables. The orthogonal projections of remaining variables to already selected ones are calculated and the variable which has the highest Euclidean length projection is selected. SPA selects variables whose information content is minimally redundant. [23]

Genetic algorithms (GAs) are optimization methods based on biological evolution. The new populations of chromosomes are generated using genetic operators, reproduction and mutation, to improve the population for solving an optimization problem. For feature selection, a subset is represented as a binary string (chromosomes) of the length of the total number of variables. The value of each position n in the string represents the presence or absence of a particular variable (1 for selected and 0 for not selected). Each variable is evaluated to determine its fitness, or its ability to survive and move into the next generation. New variables are created iterating crossover and mutation processes. The results of GA variable selection are highly dependent on the tuning parameter values, which are optimized manually one by one. [24, 25]

For a very large dataset one variable selection method, for example SPA, can be used for the variable elimination before the final variable selection by another method, for instance GA, to improve the reliability of selection [26].

Modelling

The quality of a developed model depends highly on the quality and length of the

dataset. Data should include a sufficient number of samples and it should also be fully representative of the full spectrum of all possible conditions. Especially in environmental related processes the source dataset should encompass at least one full year of measured data to ensure that all seasonal effects are included in data. In model development, efficient training and validation require long and representative enough subsets of data for both.

In this study, due to the small size of the dataset available a static split into the training and validation subsets of data was not advisable and therefore a five-fold cross-validation was used for validation of multivariable linear regression (MLR) models that predicted an output variable as a linear combination of selected input variables. The relative performances of the models were compared using Root Mean Square Error (RMSE) and coefficient of determination (R^2). In k-fold cross-validation, the whole data set is used for training and validating the model. 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. [27, 28]

Results and discussion

In the following, the modelling results for the five treated wastewater quality variables (BOD, COD, SS, N, and P concentrations) achieved using first only the optical monitoring variables and secondly only the process measurements are compared with the results of the earlier study [4] where the input variables included both the optical monitoring variables and the process measurements. A short comparison of the performances of the variable selection methods that were utilized for finding the

optimal subsets of input variables for the models is also presented.

In Table 1 the input variables selected from the optical monitoring variables and in Table 2 the input variables selected from the process measurement are presented. Variables are listed in the order of importance (the order of selection) and all the selected variables were used as input variables in the developed models.

The number of selected optical monitoring input variables in every subset is reasonably small (Table 1), from two to six, and thus the risk of developing an over-fitted model is reduced. Several methods found the identical subsets and certain variables are found important to develop a specific model. For example, four of five methods selected identical subset in developing BOD models and all methods selected aspect ratio and amount of filaments is inputs. Fractal dimension (5) and form factor (7) are found important variables to develop models for COD and SS, and median area of objects and form factor to develop models for nitrogen. These results confirm the importance of the certain optical monitoring variables in modelling a quality parameter of treated wastewater. In [4], fractal dimension (5) was found to be important input variable in the suspended solids model, aspect ratio (9) and filament length (1) in the BOD model, fractal dimension (5) and form factor (7) in the COD model, median area of objects (12) in the nitrogen model, and fractal dimension (5), amount of filaments (3) and filament length in the phosphorus model, as mainly also in this study.

The number of selected input variables from the process measurement in every subset (Table 2) is from four to seven, which is acceptable. Again, several methods found the identical subsets and certain variables are found important to develop a specific model. It is also notable that most of the selected subsets included the temperature of wastewater and anoxic proportion of volume. This is reasonable because according to the process personnel the treatment process control in the municipal WTTP

is strongly dependent on the season of the year, i.e. the temperature. Incoming load is partly flow and season dependent, and the quality of sludge and the sludge concentration depend on the influent load and the sludge age. The sludge age is one of the main factors that determine which bacterial groups are dominant and how these bacteria grow and form flocs. The sludge age is controlled mainly based on the wastewater temperature to ensure nitrification throughout the year, and is therefore dependent of the season of the year. The nitrate concentration after the active sludge process is affected by the anoxic volume, which depends on the temperature and the season of the year. Among others, these synchronous events cause quasi-correlations and are also shown in the results of the variable selection.

In the earlier study [4], in addition to the temperature and anoxic proportion of volume, influent total nitrogen (12), influent sulphate (23), and mechanically treated wastewater nitrate nitrogen (17) and iron (26) were found important input variables in the suspended solids model, iron (26) was important in modelling the BOD, sludge concentration (30) and PO4-P (11) in COD model and total nitrogen (13), pH (21) and total phosphorus (9) of mechanical treated wastewater were important in nitrogen model, as mainly also in this study.

Based on the selected subsets of input variables, MLR models for every quality parameter were developed. To evaluate the performances of the developed models the R^2 and RMSE of each model are listed in Table 3 and Table 4. As seen, the suspended solids models have the highest coefficient of determinations as also in the earlier study [4]. The phosphorus models also have satisfactory fitness but the models of other quality parameters did not yield as good. In this study, models developed using input variables selected by the genetic algorithm performed generally slightly better than other models but naturally similar subsets by other variable selection methods resulted

as good. The regression coefficients of the best models of every quality variable presented in Table 3 are listed in Table 5, where x_0 is bias and $x_{1\rightarrow n}$ is the selected input variable.

When the above mentioned modelling results are compared with the results presented in the earlier study (in Table 6) [4] it is showed that utilizing both the optical monitoring variables and the process measurements of the treatment plant yields better modelling performance for every quality variable than using only the optical monitoring variables or only the process measurements as input variables. Although the optical monitoring of the wastewater treatment process gives valuable additional information about the wastewater treatment process, all necessary information about the wastewater is not received by optical monitoring alone. Again, the process measurements alone are not sufficient to develop the best predictive models for the quality parameters. Thus, it is advisable to develop the models for the quality parameters utilizing the selected process measurements and the optical monitoring variables together. The accuracy of the best model of every quality variable is notably better in Table 6 than utilizing only either the optical variables (Table 3) or the process measurements (Table 4) expect for suspended solids models which performance improved only slightly. The reason for this is that the subset used in the best model presented in Table 6 included only one optical monitoring variable (fractal dimension) and the rest of the selected variables were nearly identical to the subset in Table 2. In this municipal WWTP, the suspended solids level is, among others, heavily related to the temperature of the incoming wastewater, which is included in every selected subset of input variables. Strong interdependence with temperature and high mutual correlations between variables may prevent the selection of other optical monitoring variables because the selection is made by mathematical grounds only. This may affect the fitness of the model.

Although the earlier studies reported in the literature have found good correlation between predicted and observed values, and used techniques provided important information for better understanding the behaviour of the activated sludge processes, the predicted parameters were measured from an aeration tank where the optical monitoring was also carried out or the studies were concentrated on the quality of the effluent in a laboratory scale process [29, 30, 31]. The test periods were often also short and in addition wide-range of process measurements were not taken into account in model development. Thus no evidence of the functionality of the methods in a full-size process or true predictive information on the quality of the effluent discharged to waterways were not achieved in many of the past studies reported, and therefore the comparison of the results based on a real-time monitoring in a full-scale treatment plant presented in this paper and in [3, 4] is not feasible.

Inspecting the results presented in this paper, it has to be pointed out that the optical monitoring was performed in one process line and the analysed samples of treated wastewater contained the wastewater from all the nine parallel treatment lines of the WWTP. It is also important to bear in mind that the variable selection methods do not take into account any deterministic models or additional chemical or biological knowledge about the activated sludge process but selections are performed based on mathematical ground only and despite the results of other selection methods (except the combination of SPA and GA selection). Without presumptions the methods are more generalizable but it has to be noted that the results based solely on a mathematical analysis may not accurately correspond the actual situation in the wastewater treatment process and that a high absolute correlation of variables not always means strong real-world causality. There also may be many hidden factors and indirect relations affecting

the real process but are not shown in the mathematical analysis due to the analysis method or the quality or length of the dataset.

However, the optical monitoring combined to predictive modelling has potential to be utilized in the process control, keeping it in stable conditions and avoiding environmental risks, as it shows the level and changes of a quality parameter.

Conclusions

A novel automatic optical monitoring device was used to image the wastewater samples in-situ in the full-scale WWTP during a period of over one year. Optical monitoring results were recorded together with the conventional process measurements. The optimal subsets of input variables for model development were searched using five variable selection methods based on mathematical grounds only and a five-fold cross-validation was used for evaluating the performance of the MLR models. The modelling results based on only the optical monitoring variables and only the process measurements were compared with the results of the earlier study, which utilized both the process measurements and the optical monitoring variables.

The comparison of the results showed that the best prediction accuracy is achieved by utilizing together both the traditional process measurements and the results of the optical monitoring and image analysis. Although a new valuable information and better understanding about the changes in the wastewater is received by the novel optical monitoring device, it is not enough to develop the best predictive model. Again, using only the process measurements the best possible fitness of models was not achieved. Using process measurements that are useful and reliable to measure from as an early stage of the process as possible with together the optical monitoring variables will improve the model accuracy and the developed model genuinely gives proactive information of the quality of the treated wastewater. The study also confirmed the

importance of the certain optical monitoring variables in modelling quality parameters of treated wastewater. The optical monitoring combined to the predictive modelling has potential to be utilized in process operation, keeping it in stable conditions and avoiding environmental damages, as it shows the level and changes of the treated wastewater quality.

Acknowledgements

The data used in this research was collected during the Measurement, Monitoring and Environmental Efficiency Assessment (MMEA), the research programme of CLEEN Ltd. – Cluster for Energy and Environment. The financial support of Riitta ja Jorma J. Takasen säätiö sr for finalizing this study is greatly acknowledged. D.Sc. (Tech.) Aki Sorsa (Control Engineering, University of Oulu) is acknowledged for help in variable selection issues and D.Sc. (Tech.) Elisa Koivuranta (Fibre and Particle Engineering, University of Oulu) is acknowledged for producing the original optical monitoring data.

References

[1] Tchobanoglous G, Burton FL, and Stensel HD. Wastewater Engineering: Treatment and Reuse. 4th ed. Boston: McGraw-Hill Education; 2003. 1819 p.

[2] Koivuranta E, Stoor T, Hattuniemi J, Niinimäki J. On-line optical monitoring of activated sludge floc morphology. Journal of Water Process Engineering. 2015;5:28–34.

[3] Tomperi J, Koivuranta E, Kuokkanen A, Juuso E, Leiviskä K. Real-time optical monitoring of wastewater treatment process. Environmental Technology. 2016;37(3):344-351. doi: 10.1080/09593330.2015.1069898

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

[5] Mesquita DP, Amaral AL, Ferreira EC. Activated sludge characterization through microscopy: A review on quantitative image analysis and chemometric techniques. Analytica Chimica Acta. 2013;802:14–28.

[6] Henze M, Grady CPL Jr, Gujer W, Marais GVR, Matsuo T. Activated Sludge Model No. 1. IAWQ Scientific and Technical Report No. 1, London, UK. 1987.

[7] Gernaey KV, van Loosdrecht MCM, Henze M, Lind M, Jørgensen SB. Activated sludge wastewater treatment plant modelling and simulation: state of the art. Environmental Modelling & Software. 2004;19:763–783.

[8] Keskitalo J, la Cour Jansen J, Leiviskä K. Calibration and validation of a modified ASM1 using long-term simulation of a full-scale pulp mill wastewater treatment plant. Environmental Technology. 2010;31(5):555-566.

[9] Keskitalo J, Leiviskä K. Application of evolutionary optimisers in data-based calibration of activated sludge models. Expert Systems with Applications. 2011;39(7):6609-6617. doi: 10.1016/j.eswa.2011.12.041

[10] Leiviskä K, Keskitalo J. Artificial neural network ensembles in hybrid modelling of activated sludge plant. IEEE IS 2014 – 7th IEEE Conference on Intelligent Systems, Warzaw 24-26.9.2014.

[11] Belanche L, Valde´s JJ, Comas J, Roda IR, Poch M. Prediction of the bulking phenomenon in wastewater treatment plants. Artificial Intelligence in Engineering. 2000;14:307–317.

[12] Mjalli FS, Al-Asheh S, Alfadala HE. Use of artificial neural network black-box modeling for the prediction of wastewater treatment plants performance. Journal of Environmental Management. 2007;83:329–338.

[13] Mesquita DP, Dias O, Dias AMA, Amaral AL, Ferreira EC. Correlation between sludge settling ability and image analysis information using partial least squares. Analytica Chimica Acta. 2009;642:94–101.

[14] Mesquita DP, Dias O, Amaral AL, Ferreira EC. Monitoring of activated sludge settling ability through image analysis: validation on full-scale wastewater treatment plants. Bioprocess Biosystem Engineering. 2009;32:361–367. doi: 10.1007/s00449-008-0255-z

[15] Amaral AL, Rodrigues S, Mota M, Ferreira EC. Morphological characterization of biomass in wastewater treatment using partial least squares. Proceedings of the Second IASTED International Conference Visualization, Imaging, and Image processing. September 9-12, 2002, Málaga, Spain.

[16] Amaral AL, Ferreira EC. Activated sludge monitoring of a wastewater treatment plant using image analysis and partial least squares regression. Analytica Chimica Acta. 2005;544:246–253.

[17] HSY Viikinmäki wastewater treatment plant webpage. Cited November 2016. Available from: https://www.hsy.fi/en/experts/water-services/wastewater-treatment-plants/viikinmaki/Pages/default.aspx.

[18] Koivuranta E, Keskitalo J, Haapala A, Stoor T, Sarén M, Niinimäki J. Optical monitoring of activated sludge flocs in bulking and non-bulking conditions. Environmental Technology. 2013;34(5-8):679–686.

[19] Juuso E. Integration of intelligent systems in development of smart adaptive systems: linguistic equation approach. - Acta Universitatis Ouluensis. Series C, Technica 476. Oulu. Dissertation. 258; 2013.

[20] Hall MA. Correlation-based feature selection for machine learning. The University of Waikato, New Zealand. Doctoral Thesis; 1999.

[21] Guyon I, Elisseeff A. An introduction to variable and feature selection. The Journal of Machine Learning Research. 2003;3:1157-1182.

[22] MathWorks. Statistics and Machine Learning Toolbox documentation. Cited November 2016. Available from: se.mathworks.com.

[23] Araujo MCU, Saldanha TCB, Galvao RKH, Yoneyama T, Chame HC, Visani V. The successive projections algorithm for variable selection in spectroscopic

multicomponent analysis. Chemometrics and Intelligent Laboratory Systems. 2001;57:65–73.

[24] Siedlecki W, Sklansky J. A Note on Genetic Algorithms for Large-Scale Feature Selection. Pattern Recognition Letters. 1989;10:335-347.

[25] Davis L. Handbook of genetic algorithms. Van Nostrand Reinhold; 1991. 385 p.

[26] Sorsa A, Leiviskä K, Santa-aho S, Vippola M, Lepistö T. An Efficient Procedure for Identifying the Prediction Model Between Residual Stress and Barkhausen Noise. Journal of Nondestructive Evaluation. 2013;32(4):341-349.

[27] Arlot S, Celisse A. A survey of cross-validation procedures for model selection. Statistics Surveys. 2010;4:40–79.

[28] Witten IH, Eibem F, Hall MA. Data Mining, Practical Machine Learning Tools and Techniques, 3nd edition, Elsevier, Burlington; 2011.

[29] Mesquita DP, Amaral AL, Ferreira EC. Estimation of effluent quality parameters from an activated sludge system using quantitative image analysis. Chemical Engineering Journal. 2016;285:349–357.

[30] Smets IY, Banadda EN, Deurinck J, Renders N, Jenne R, Van Impe JF. Dynamic modelling of filamentous bulking in lab-scale activated sludge process. Journal of Process Control. 2006;16:313-319.

[31] Banadda EN, Smets IY, Jenne R, Van Impe JF. Predicting the onset of filamentous bulking in biological wastewater treatment systems by exploiting image analysis information. Bioprocess Biosystem Engineering. 2005;27:339–348. doi: 10.1007/s00449-005-0412-6

Table 1. Selected subsets of input variables from the optical monitoring variables.

	BOD	COD	SS	N	P
Correlation analysis	9, 3, 17, 12	5, 3, 12, 7, 11	5, 3, 12, 11, 7	12, 11, 7, 3	5, 3, 12, 11
Stepwise selection	9, 3	5, 7	5, 7, 3	12, 7	8, 1, 5, 16
Forward selection	9, 3	5, 7	5, 2, 1, 7, 8, 17, 11	12, 7	5, 11, 1, 17
Genetic algorithm	3, 9	7, 16, 17	1, 2, 7, 8, 11, 17	1, 5, 7, 16, 17	1, 16, 17
SPA + GA	9, 3	7, 17	7, 11, 3, 5	7, 12	5, 3, 11, 17

Variables	
1	Filament length
2	Floc area
3	Amount of filaments
5	Fractal dimension
7	Form factor
8	Roundness
9	Aspect ratio
10	Equivalent diameter
11	Mean area of objects
12	Median area of objects
16	Number of objects
17	Number of small objects

Table 2. Selected subsets of input variables from the process measurements.

	BOD	COD	SS	N	P
Correlation analysis	29, 28, 17, 1	28, 4, 29, 17, 11, 31	28, 29, 17, 27, 16, 14	13, 19, 21, 28, 30, 17	28, 29, 17, 16, 26, 27
Stepwise selection	29, 26, 28, 16	28, 4, 11, 29, 30	28, 29, 17, 26, 12, 23	28, 21, 9, 29, 30, 15, 14	28, 29, 12, 17, 9
Forward selection	28, 29, 26, 16	28, 4, 29, 11, 30	28, 29, 12, 17, 26, 23	19, 28, 17, 5, 25	28, 29, 12, 9, 17, 11, 23
Genetic algorithm	16, 26, 28, 29	4, 11, 28, 29, 30	12, 17, 23, 26, 28, 29	9, 12, 15, 18, 21, 25, 28	9, 12, 17, 28, 29
SPA + GA	28, 16, 26, 29	28, 30, 11, 4, 29	15, 23, 28, 17, 26, 29	15, 28, 21, 30, 9, 29	15, 28, 17, 9, 29

Variables	
1	(I) BOD
4	(M) COD
5	(I) SS
9	(M) Total phosphorus
11	(M) PO4-P
12	(I) Total nitrogen
13	(M) Total nitrogen
14	(I) Ammonium nitrogen
15	(M) Ammonium nitrogen
16	(I) Nitrate nitrogen
17	(M) Nitrate nitrogen
18	(I) Alkalinity
19	(M) Alkalinity
21	(M) pH
23	(I) Sulphate
25	(I) Iron
26	(M) Iron
27	Flow
28	Anoxic proportion of vol.
29	Temperature
30	Sludge concentration
31	Sludge age

(I) influent, (M) mechanically treated wastewater

Table 3. The modelling results using only the optical monitoring variables.

	BOD		COD		SS		N		P	
	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE
Correlation analysis	0.41	0.73	0.39	0.74	0.62	0.63	0.31	0.84	0.56	0.62
Stepwise selection	0.41	0.73	0.37	0.75	0.62	0.63	0.31	0.84	0.61	0.58
Forward selection	0.41	0.73	0.37	0.75	0.67	0.58	0.31	0.84	0.61	0.58
Genetic algorithm	0.41	0.73	0.43	0.72	0.67	0.58	0.42	0.77	0.60	0.59
SPA + GA	0.41	0.73	0.38	0.75	0.62	0.63	0.31	0.83	0.62	0.58

Table 4. The modelling results using only the process measurements.

	BOD		COD		SS		N		P	
	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE
Correlation analysis	0.34	0.77	0.46	0.70	0.72	0.54	0.45	0.75	0.51	0.65
Stepwise selection	0.47	0.69	0.49	0.68	0.77	0.49	0.58	0.65	0.63	0.57
Forward selection	0.47	0.69	0.49	0.68	0.77	0.49	0.48	0.72	0.65	0.55
Genetic algorithm	0.47	0.69	0.49	0.68	0.77	0.49	0.50	0.71	0.63	0.56
SPA+GA	0.47	0.69	0.49	0.68	0.75	0.51	0.56	0.65	0.61	0.58

Table 5. The regression coefficients of the best developed models using optical monitoring variables.

BOD	-0.62 x_0	0.42 x_9	0.25 x_3				
COD	-1.06 x_0	3.83 x_5	-3.28 x_6	-0.31 x_7	0.45 x_{13}		
SS	-1.05 x_0	0.80 x_1	-1.55 x_2	-0.21 x_7	-0.50 x_8	0.48 x_{17}	0.82 x_{11}
N	0.02 x_0	0.50 x_1	-0.71 x_6	-0.33 x_7	-2.90 x_{15}	3.15 x_{17}	
P	0.04 x_0	-1.50 x_5	0.35 x_3	-0.39 x_4	0.45 x_{11}	1.55 x_{17}	