Modelling the effluent quality based on a real-time optical monitoring of the wastewater treatment process

Tomperi Jani[a], Koivuranta Elisa[b], Kuokkanen Anna[c], Leiviskä Kauko[a]

[a]*Control engineering, University of Oulu, Oulu, Finland.*

[b]*Fibre and Particle Engineering, University of Oulu, Oulu, Finland.*

[c]*HSY Helsinki Region Environmental Services Authority, Helsinki, Finland*

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

Modelling the effluent quality based on a real-time optical monitoring of the wastewater treatment process

A novel optical monitoring device was used for imaging an activated sludge process in-situ during a period of over one year. In this study, the dependencies between the results of image analysis and the process measurements were studied, and the optical monitoring results were utilised to predict the important quality parameters of the wastewater treatment process efficiency: suspended solids (SS), biological oxygen demand (BOD), chemical oxygen demand (COD), total nitrogen and total phosphorous in biologically treated wastewater. The optimal subsets of variables for each model were searched using five variable selection methods. It was shown that the on-line optical analysis results have clear dependencies on some process variables and the purification result. The model based on optical monitoring and process variables from the early stage of the treatment process can be used to predict the levels of important quality parameters, and to show the quality of the biologically treated wastewater hours in advance. This study confirms that the optical monitoring method is a valuable tool for monitoring a wastewater treatment process and receiving new information in real-time. Combined to predictive modelling it has potential to be used in process control, keeping the process in a stable operating condition and avoiding environmental risks.

Keywords: activated sludge process; BOD; COD; suspended solids; variable selection methods

Introduction

Activated sludge processes (ASP) are widely used for treating industrial and municipal wastewaters. These complex nonlinear biological processes are very sensitive to external and internal disturbances (the seasonally changing temperature, the varying quality and quantity of wastewater, heavy rains, etc.) affecting the optimum operating conditions and a good bacterial balance which is a key role for an efficient ASP. Efficient ASP has a high pollution removal rate, low suspended solids in the effluent and good settling properties of the sludge, among others. The disturbances in the bacterial balance will most often be shown as dysfunctional flocculation and settling properties which may have serious environmental and economic effects. The most common problems in ASP are filamentous bulking, pinpoint flocs formation, dispersed growth, and viscous (zoogleal) bulking. For example, in filamentous bulking the biomass is strongly colonized by long filaments that hold the flocs apart and hinder the sludge settling, which causes the amount of total suspended solids content in the effluent to rise and reduce the quality of treated wastewater. Recovering from the disturbances is slow and the effect on the process operation and purification results are long-lasting. [1-3]

A more accurate operating of the wastewater purification process is required to meet the constantly tightening regulations for effluent discharges and for keeping the operation costs at a minimum. Although the expert knowledge and traditional offline analysis of wastewater samples are important in process monitoring and control, there is a demand for new on-line monitoring tools and methods. Assessing the process performance by offline analysis of the key quality parameters, such as biological and chemical oxygen demand (BOD, COD), suspended solids (SS) and sludge volume index (SVI), only shows the poor quality of effluent when it already occurs and it is too late for corrective operations. Both BOD and COD show the amount of organic matter in wastewater as they are a measure of dissolved oxygen required to oxidize the dissolved organic matter in wastewater. In routine process control SVI is used to monitor the settling characteristics of activated sludge. However, these analyses either do not provide essential information about the reasons behind the problems in the process. On-line optical monitoring of floc morphological characteristics, on the other hand, gives fast objective information about the quality of wastewater and the state of the treatment process, reveals some of the reasons for settling problems and combined to a predictive model shows the quality of effluent in advance. [1-7]

In this study, the results of the in-situ optical monitoring of an ASP from a period over one year were studied together with the offline process measurements data. The longer and completely new data included also seasonal changes in operating conditions and gave more precise results than in the earlier study [6] with short dataset from the same ASP. The optical monitoring data were collected using a new high-resolution charge-coupled device (CCD) camera. Dependencies were sought between the image analysis results and the process measurements focusing on the quality parameters of the process efficiency: suspended solids content, biological oxygen demand, chemical oxygen demand, total nitrogen and total phosphorus in biologically treated wastewater. Variable selection methods were used for selecting the optimal subsets of variables for each developed predictive model for the quality parameters measured from biologically treated wastewater. The models can be used to evaluate the performance of the wastewater treatment plant (WWTP) and for a better control regarding stable effluent quality. Control actions can be comprehended on chemical dosing, sludge concentration and sludge age, for instance.

Material and Methods

Wastewater treatment plant

The optical monitoring device was sited at one of the nine activated sludge process lines of Viikinmäki WWTP. The largest wastewater treatment plant in Finland processes the wastewaters of over 800,000 inhabitants. The average daily flow through the process is about 270,000 m^3 of wastewater but melting snow and hard rains often multiply the flow. 85% of the total flow is from domestic and 15% of total flow is from industrial sources. The time period spent in each process stage depends on the flow, but on average the wastewater stays in the plant for about 24 hours. Viikinmäki WWTP 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. The unit operations of the process are intake, screening, grit and grease removal, preliminary settling, aeration, degassing, secondary settling, biological denitrification filtration and discharge (Figure 1). Screening removes the large solids from the wastewater. Grit and grease removal separates rapidly settling, coarse solids and greasy and oily substances that are lighter than water. In the preliminary settling phase, easily settling material is separated from the wastewater. The biological treatment is conducted by means of a denitrification/nitrification process in an aeration tank which is used to grow activated sludge. At the head of the aeration tank, in a separate mixing area, 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 the denitrification of the wastewater. The entire process removes 95% of the phosphorus, 90% of the nitrogen together with nearly all solids and oxygen-consuming substances from the wastewater. The BOD removal rate has been between 89% and 98% during the last decade. [8]

Optical monitoring and image analysis

Optical monitoring of wastewater samples is traditionally performed manually under a microscope, which is a slow, irregular and subjective method. On this account, several optical monitoring and image analysis methods have been developed in recent years for wastewater monitoring and characterisation [9]. However, many of the developed methods have been tested only in a laboratory or at a pilot-scale and are not suitable for full-scale and in-situ use. The data for this study was measured using the novel optical monitoring device and image analysis methods described in detail in [10]. The small-scale on-line optical monitoring device was specially designed for in-situ use and it has been tested in full-scale municipal ASP for a period of over eight months and proved functional for reliable in-situ monitoring of the floc morphology [4]. The device consists of an imaging unit, a sample handling unit and a control PC with an electronics unit (Figure 2). Wastewater samples were taken from the aeration tank, diluted and pumped through a cuvette, which was imaged with a high-resolution charge-coupled device (CCD) camera. To determine the appropriate dilution rate for the on-line use three different dilutions (1:50, 1:100 and 1:200) were tested at the laboratory [4]. No significant differences in floc morphology were found using different dilutions and 1:100 ratio was selected to be used in the on-line measurements. The laboratory test indicates 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.

The sensor of CCD camera is 5.5 mm * 3.7 mm (1392 * 1040 pixels) with a pixel size of 3.6 µm * 3.6 µm. The 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 (FD), form factor (FF) and roundness (RO). The parameters are calculated as an average of the values for individual objects over a single image. One analysed video contains about 1000 images and one image contains 150 flocs on average. As two videos were analysed, the obtained results of wastewater sample can be considered statistically reliable. The amount of filaments is presented as a ratio of filament length and floc area. The total filament length is the sum of the filaments length of all filamentous bacteria present in the image. 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 is calculated as the sum of the distributions of individual images. The case specific floc area threshold value for the calculated parameters is set in the user interface of the image analysis software and in this study it was 100 µm^2. The limit value for small objects was set in an equivalent diameter of under 25 µm.

The mathematical formulas of the calculated size and shape parameters are presented in [10]. The equivalent diameter is the diameter of a circle with an area equal to the object's area. The form factor is affected by the irregularity or roughness of the object's boundary. It is 1.0 for a perfect circle and below 1.0 for any other shape. Objects with more irregular boundaries have a longer perimeter per surface area and therefore have smaller form factors. Roundness (also 1.0 for a perfect circle) is defined as the ratio between the area of an object and the area of a circle with a diameter equal to the object's length. The floc structure can be also described using the fractal dimension. The fractal dimension is calculated by the box counting concept: the image is covered with squares of a certain size using the minimum number of squares to completely cover the object in the image. The square size is then reduced and the process is repeated. The fractal dimension is obtained from the slope of the curve from the logarithmic plot of the size of squares against the logarithm of number of the squares: D = – log (N) / log (s), where s is the box size and N represents the number of boxes of size s needed to cover the object. [11, 12]

As mentioned, the wastewater samples for on-line imaging were taken from one activated sludge line in the aeration tank, before secondary sedimentation tank. At normal flow, the delay between optical monitoring measurement and the output of the WWTP is about 13 hours. 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. In the datasheet where the optical monitoring and laboratory measurements were combined by date, the missing laboratory data was not interpolated because interpolation was found to dilute the results of data analysis. In addition, due to setup problems some of the data was corrupted and during the process maintenance stoppages measurements could not be performed. Thus, the total amount of data available was 94 data samples (measurement times).

Data pretreatment

Before data analysis and modelling, a dataset including several measurements has to be scaled or normalized, to facilitate analysis, to avoid incorrect conclusions and to reveal all the noteworthy changes or states. In this work, the dataset was scaled between [-2, 2] using the nonlinear scaling method based on generalized moments, norms and skewness [13]. Before scaling, a dataset should be inspected and feasible incorrect values deleted or replaced by interpolation if necessary. In this study, no interpolation or deletion of values was not needed to perform. A nonlinear mapping function has been developed to extract the meanings of variables from measurement signals. These functions are called membership definitions which map the real values of variables to the linguistic range of [-2, 2]. Thus, a normal scaling to range [-1, 1] is combined with the handling of warnings and alarms. A trapezoidal membership function which is based on the support and core areas [c_l, c_h] defined by fuzzy set theory is used to define the concept of the feasible range. The support area is defined by the minimum and maximum of the values of the variable x. The value range is divided into two parts by the central tendency value c and the core area is limited by the central tendency values of the lower and upper part. The tuning approach based on the generalised skewness is used for estimating the central tendency value and the core area. The central tendency value is chosen by the point where the skewness changes from positive to negative. Then the dataset is divided into a lower part and an upper part. The same analysis is done for these two datasets. The estimates of the corner points are the points where the direction of the skewness changes for the lower and upper data set, respectively. The iteration is performed with generalised norms. The mapping function contains one monotonously increasing function for the values between -2 to 0 and one monotonously increasing function between values 0 to +2. In order to keep the functions monotonous and increasing, the derivatives of the functions should always be positive. Membership functions consist of two second order polynomials: one for the negative values and one for the positive values. Because the scaling idea is based on the membership functions of fuzzy set systems these values are called linguistic values. The coefficients of the polynomials are defined by the corner points. The main concepts are presented in Figure 3, which shows the connection between the membership definition and the corresponding fuzzy membership functions.

Variable selection

Modern plants most likely have several sensors to measure on-line data and large number of manually collected samples around the process are also analysed offline. Large datasets often include irrelevant variables for a specific purpose, for instance modelling. Variable selection is one of the most important steps in the data analysis and model development. Only significant variables must be selected because a greater number of variables does not necessary mean better prediction results. Some input variables may be correlated with each other, noisy or have no significant relationship with the output variable and thus will not be informative. Selecting non-essential input variables increases computational complexity, makes the training of the model more difficult and prediction results worse. 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. An over-fitted model has an excellent performance in training but is not general usable with new upcoming data. In this work, in addition to selecting a subset of input variables manually after the correlation and visual analysis of the dataset, five variable selection methods were used to find the optimal subsets for modelling the suspended solids content, BOD, COD, total nitrogen and total phosphorus in biologically treated wastewater.

Variable selection methods can be roughly grouped into wrapper and filter methods. In a filter method, variables are selected or deleted according to the formed ranking which is based on the correlation coefficients. Filter methods are very efficient but the model is seldom optimal. In a wrapper method, a subset of variables is assessed according to their usefulness to a given predictor. Wrapper methods wrap around an appropriate learning machine which is employed as the evaluation criterion, such as prediction or classification error. Wrappers often give better results but are slower than filters. For example, correlation-based selection and a successive projections algorithm (SPA) are classified as filters and forward selection and a genetic algorithm are classified as wrappers. [15, 16]

For very large datasets two variable selection methods can be used one after another: one variable selection method is used for the variable elimination before the final variable selection by another method. In Sorsa et al. [17] a successive projections algorithm (SPA) together with a modified genetic algorithm method was used and it was found that SPA applied before genetic algorithm search greatly improves the reliability of the genetic search. The genetic algorithm is able to find the global optimum well and the computational load is greatly reduced by SPA. The usage of SPA greatly reduces the variability in the selection results.

Correlation-based selection

In correlation-based selection, variables are selected by the absolute value of their correlation coefficient. The selected subset should contain variables that have a high correlation coefficient with the output variable but low correlation between each other. For this study, variables that had mutual correlation |0.85| or larger were removed from the set. The variable with a lower correlation coefficient was removed and variables were arranged in downward order by their correlation coefficient with the output variable.

Forward selection

A forward selection is a simple variable selection method which goes through the data set and adds one variable at a time to a model. Initially, every variable from the variable set is evaluated and the best variable is added. At the following steps, variables that are not already included in the model are evaluated and variables are added one at a time to the model based on their performance. Adding is continued until the performance of the model does not improve more and the best combination of the variables is selected. The drawback of the forward selection method is that no variables are removed from subset once they are selected and it may easily get trapped to local optimum. Variables whose performance is strong together but poor alone are not selected due to the single selection principle. [15, 16]

Stepwise regression

Stepwise regression function in Matlab is a modified forward selection method, which adds the best variable to, or deletes the worst variable from a variable subset at each round. Adding and deleting is based on variable's statistical significance in regression. It starts with an initial model and continues until either no further model changes occur over one complete round or a preset number of variable selections and deletions occur. Depending on the variables included in the initial model and the order in which variables are added and removed, the method may build different models from the same set of variables. Stepwise models are locally optimal, but may not be globally optimal. [18]

Successive projections algorithm

A successive projections algorithm (SPA) is a forward selection method in multivariate calibration. SPA uses simple operations in a 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. The method starts with one variable and adds a new variable at each iteration until a specific number of variables is reached. SPA selects variables whose information content is minimally redundant. SPA steps are described in details in [19].

Genetic algorithms

Genetic algorithms (GAs) are evolutionary optimization methods for various problems. GAs are based on the biological evolution. The new populations of chromosomes are generated using genetic operators, reproduction (selection and crossover) and mutation, to improve the population for solving an optimization problem. A genetic algorithm usually starts from the random initial population, with a few tuning parameters, for example the population size, crossover and mutation probabilities. Each chromosome in the initial population represents a different solution to the problem. The initial population is evaluated and a new population is created generating offspring. Two members are selected using for example tournament or roulette wheel selection to be parents to crossover. In roulette wheel selection parents are selected according to their fitness: the better the chromosomes are the more they have chances to be selected. In crossover, parts of two different parent chromosomes are mixed to create offspring. A random number is compared with predefined crossover probability and if the random number is smaller than the probability, the parents are crossed with the selected crossover method. If the random number is larger, parents are moved to the new population. The idea is to combine the good values of chromosomes and create better chromosomes. Mutation makes random changes to the population and it occurs if the mutation probability is larger than the random number. Population is evaluated and the new population is finalized with elitism which moves a predefined number of the best chromosomes to the new population. This prevents the disappearance of good solutions but also decreases the diversity of the population by increasing the dominance of the best chromosomes. Steps are repeated until the population of predefined size is created. [20]

Siedlecki and Sklansky [21] introduced the use of genetic algorithms for feature selection. 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.

In this study, genetic algorithm selection uses multivariable linear regression models and leave-multiple-out cross-validation.

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 but it should also be fully representative of the full spectrum of all possible conditions. In environmental related processes the source dataset should 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 model development, efficient training and validation require long and representative enough subsets of data for both training and testing. With small data sets, the split to these subsets is not possible without a significant loss of data. A cross-validation is a typical resampling method and one way to predict the fit of a model for a validation set when dataset is small and an explicit validation set is not available. Three cross-validation methods are available: leave one out (LOO), leave multiple out (LMO) and k-fold. In all methods, the whole data set is used for training and validating the model by using part of the data for training and the rest of the data for validation and repeating this until the whole dataset is processed. Thus the largest possible test set can be used, which is a great advantage especially with a small dataset. 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 (averaging) these k results. Optimal k is often reported to be 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 more feasible. In this study, a five-fold cross-validation was used to evaluate the model accuracy. A multivariable linear regression (MLR) was used to predict an output variable as a linear combination of selected input variables. 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. [22, 23]

Results and Discussion

In this work, the dataset, which consisted of optical monitoring results and wastewater treatment process measurements from a period of over one year, was analysed. Variable selection methods were used in searching the optimal subsets of variables for developing a predictive model for suspended solids, BOD, COD, total nitrogen and total phosphorus in biologically treated wastewater. The variable selection methods, contrary to the manual selection, did not take into account any deterministic models or chemical or biological knowledge about the activated sludge process but selections were performed based on mathematical ground only. Without presumptions the methods are, thus, more generalizable and easily usable in different process cases. However, 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 correlation coefficients between variables not always mean strong real-world causality. There are also many hidden factors that affect in the real process but are not shown in this data analysis. The data used in this study included also seasonal changes (temperature changes, heavy rains, melting snow, variations in the quality and quantity of wastewater, etc.) which gives a more accurate analysis of the process operation than the earlier study of Tomperi et al. [6] with a short dataset because many factors affecting the quality of sludge and the purification process are dependent, for example, on the temperature.

The dependencies between the on-line image analysis results and the WWTP process measurements were studied using correlation analysis and visual examination of scaled variables. Correlation analysis is very challenging to perform and correctly interpret in absolute due to the complexity of the wastewater treatment process. According to the process personnel, in the Viikinmäki WWTP the incoming load is partly dependent on the flow and the season of the year. The quality of sludge and the sludge concentration (mixed liquor suspended solids) depend on the influent load and the sludge age. To ensure nitrification throughout the year, the sludge age is controlled mainly based on wastewater temperature and is, thus, dependent on the season of the year. The sludge age is one of the main factors that determine which bacterial groups are dominant and how these bacteria grow and form flocs. For example, if the sludge age is too high related to the temperature, the amount of filamentous bacteria may increase and cause filamentous bulking. The nitrate concentration after the activated sludge process is affected by the anoxic volume, which is dependent on the temperature and the season of the year. Among others, these synchronous events cause quasi-correlations.

Suspended solids level, BOD and COD contents in biologically treated wastewater and influent, important operating conditions that affect the process efficiency (flow, the temperature, sludge age, sludge concentration and anoxic proportion of volume) and selected image analysis variables are presented in Figure 4 as scaled values, which enables better observation of even small changes in a plotted trend. Minimum, maximum, median and standard deviation values calculated from the raw data of the variables in Figure 4 are listed in Table 1. The level of suspended solids, BOD, COD content and total nitrogen and phosphorus (which are not shown in Figure 4) in biologically treated wastewater (B) are mutually correlated and follow the changes of the temperature: for example, the level of suspended solids was very low in summer time (in Figure 4 data points from 32 to 57 represent the days from June to end of August) when wastewater was warmer and rose when the temperature decreased at the beginning of winter. This is typical in wastewater treatment plants with longer sludge age in winter. The process operating is more efficient when the temperature is warmer and external disturbances in flow are not present. In summer and fall when the quality of the biologically treated wastewater was at (very) good level the amount (ratio of filament length and floc area) and length of filaments were low, flocs were larger, the roundness of flocs was higher and the number of objects was lower. When the quality of the treated wastewater deteriorated the amount and length of filaments notably increased, the roundness of the flocs decreased and number of the objects increased. The results show that the on-line optical monitoring variables measured from the aeration tank indicate the quality of the biologically treated wastewater.

The correlation analysis showed that the selected image analysis variables have correlations with several process measurements. In Table 2 and Table 3, the correlation coefficients greater than |0.50| of selected variables are shown. Biologically treated wastewater (B) samples are taken after biological treatment (secondary settling) and effluent (E) samples are taken from the outfall of the process after the excessive biological filter. Several optical monitoring parameters were removed from the tables because they had no significant correlation with any interesting process measurement or due to high cross correlation with the other optical monitoring parameters. The optical monitoring parameters have several mutual correlations, for example when the amount and length of filaments is high the roundness and area of flocs are low and the number of small objects is high. Although correlations between process variables and optical monitoring variables were found, no significantly high correlation coefficients were revealed except for the temperature of wastewater. The optical monitoring variables correlated, in addition to BOD and SS, mainly on phosphorous and nitrogen content in biologically treated wastewater and effluent from denitrifying filters. According to this analysis, the quality parameters of biologically treated wastewater (BOD, COD, SS, total phosphorus and nitrogen) have high mutual correlation.

The temperature, flow, sludge age, sludge concentration (or mixed liquor suspended solids) and anoxic proportion of volume are considered important operating conditions that affect the process efficiency. Sludge age has no correlation coefficient above |0.50| with any variable and is not shown in Table 3. A reason for this may be that the sludge age is manually controlled based on the season of the year and condition of the biomass. As it can be seen, the temperature and wastewater flow have the large number of correlation above |0.50| with other process measurements, but flow has no significant correlation with any of the image analysis variables or the suspended solids, BOD and COD in biologically treated wastewater (B). However, the flow may have indirect effect on these parameters which is not revealed on mathematical analysis. The temperature and flow have a mutual correlation of -0.52. Based on the correlation analysis, when the flow is high several variables measured from the incoming wastewater are at low level. This is due to the diluting effect of rainfall and melting waters. The temperature correlates with several image analysis variables and it can be assumed that when the temperature is high, the filament length, amount of filaments and number of measured objects are low, and on the other hand, the roundness of the flocs is high.

The usefulness of the image analysis as a novel informative monitoring tool is proved as any other variable measured before the aeration tank does not have as high correlation with suspended solids in biologically treated wastewater as the image analysis variables. Suspended solids also correlate with several measurements in effluent and biologically treated wastewater, for example iron, COD, total phosphorus and total nitrogen. However, as inspecting the results, it has to be pointed out that the optical monitoring was performed from one of the nine parallel process lines whereas the suspended solid and other measurements of biologically treated wastewater included wastewater from all nine lines. Although the results in this work are similar to the results presented in [6], minor differences occurs mostly due to the length of the dataset and included seasonal changes. Using the new camera with higher resolution also affected the results positively as more accurate imaging of shape and size of flocs and filaments was achieved. It is advisable to use a camera that has as high resolution as possible to image the wastewater samples, yet a camera with a lower resolution as in the earlier study [6] also yields sufficient results. The longer dataset used in this study confirmed that the variables that had a high correlation coefficient in the earlier analysis have high coefficients also in the present study.

Optical monitoring results were utilised together with traditional process measurements to develop predictive models for suspended solids, biological oxygen demand, chemical oxygen demand, total nitrogen and total phosphorus in biologically treated wastewater. Only variables that are useful and reliable to measure were selected, from as an early stage of the process as possible, so that the developed models could genuinely give proactive information for the quality of biologically treated wastewater. In addition to manually selected variables based on the visual inspection of data, expert knowledge and trial and error, the optimal subsets of input variables were searched using five variable selection methods based on mathematical grounds only: correlation-based selection, forward selection, Matlab stepwise selection, a genetic algorithm and a successive projections algorithm combined with a genetic algorithm. Resulted subsets are listed in Table 4 (SS), Table 5 (BOD), Table 6 (COD), Table 7 (Nitrogen) and Table 8 (Phosphorus) in order of the significance of variables. Altogether, 12 variables were selected in different subsets to develop models for suspended solids (Table 4). For suspended solids model, forward selection, the stepwise regression method and the genetic algorithm gave similar subsets. The similarity is explained as they are all wrapper selection methods. The similarity proves the functionality of these selection methods. The correlation-based method and SPA+GA method subsets included a few different variables. Several variables (fractal dimension, anoxic proportion and nitrate-nitrogen, for instance) were present in all subsets and can be considered important variables in this case. Because three of five variable selection methods gave similar subsets, the model developed using these seven selected variables (fractal dimension, influent total nitrogen and sulphate, mechanically treated wastewater iron and nitrate nitrogen, the temperature and anoxic proportion) can be considered the optimal with this dataset. The found subsets of variables to develop an optimal predictive model for BOD in biologically treated wastewater are shown in Table 5. Altogether, 14 variables were selected in different subsets to develop models for BOD content and three methods (stepwise, forward and SPA+GA) gave identical subsets (aspect ratio, anoxic proportion, iron in mechanically treated wastewater and the length of filaments) of four variables. Thus these variables can be considered important and optimal in developing a BOD model. To develop models for COD content, 12 variables were selected in different subsets. Stepwise, forward, GA and GA+SPA methods found similar, yet not identical, subsets of variables. The anoxic proportion of volume, (M) PO4-P, sludge concentration and (M) BOD, for instance, are present in several subsets and have an important role to model the COD. Altogether, 18 variables were selected in different subsets to develop models for total nitrogen (Table 7) and 12 variables were selected in different subsets to develop models for total phosphorus (Table 8). No identical subsets were found but several variables were selected for many subsets. For example (M) total nitrogen and phosphorus, area of flocs, (M) pH and anoxic proportion of volume seem to be important variables to model the total nitrogen in biologically treated wastewater and fractal dimension, the anoxic proportion of volume, and (M) total nitrogen seem to be important variables to model the total phosphorus.

The fitness of the models for suspended solids, BOD, COD, total nitrogen and total phosphorus contents were estimated by five-fold cross-validation and the R^2 and RMSE values of the developed models are listed in Table 9. Models were developed using all the selected variables of each subset presented in Table 4, Table 5, Table 6, Table 7 and Table 8. Similar subsets naturally resulted in similar fitness of models. The best modelling results for BOD (R^2=0.55), COD (R^2=0.55), nitrogen (R^2=0.63) and phosphorus (R^2=0.69) were satisfactory, yet far worse than the best modelling result for suspended solids (R^2=0.79). The fitness of the worst model for suspended solids, which was yet acceptable, was better than the fitness of the best models for other quality parameters. Supposedly, some unknown factors that are not present in the dataset affect the other quality parameters and the modelling results are lower that for suspended solids. The optimal variable set for suspended solids is different and the modelling result slightly worse than in the earlier study [6]. This is possible due to the length of dataset including seasonal changes which add variation and some noise to the data. However, in model development for a process where seasonal changes have an important role, the results achieved by longer data can be considered more reliable. The differing result also shows that the developed ASP models are not generalizable although certain variables have an important role in every ASP, and that the developed models should be actively updated. The presented results can be considered acceptable taking into account that the optical monitoring was done only from one of nine parallel process lines whereas the offline analysed manually collected samples of biologically treated wastewater included wastewaters from all lines. Adding more process measurements to the dataset or using more complicated modelling methods the modelling results of all quality variables could be improved. Nevertheless, the best developed models can be used for proactive monitoring and estimating the level of suspended solids, BOD, COD, total nitrogen and total phosphorus content in several conditions during the year hours before in comparison to offline laboratory analysis taken from treated water. Thus, the in-situ imaging method has potential to be used as assistance in process control.

Conclusion

In this study, the results of in-situ performed optical monitoring of wastewater were compared with the data obtained from the offline process measurements. A novel optical monitoring device was used to image wastewater samples from one of the nine process line during a period of over one year. The goal was to study the dependencies between process measurements and optical analysis variables, and utilise the optical monitoring results to predict the important and critical quality parameters of the wastewater treatment process efficiency: suspended solids, biological oxygen demand, chemical oxygen demand, total nitrogen and total phosphorus in biologically treated wastewater. The optimal subsets of input variables for model development were searched using five variable selection methods based on mathematical grounds only.

It was found that the on-line optical analysis results have clear dependencies on some process variables and in the purification result. For example, the amount and length of filaments and roundness of the flocs correlated strongly with suspended solids in biologically treated wastewater which on the other hand have a high correlation with BOD and COD content. The temperature of wastewater has high correlation coefficients with several image analysis variables. The usefulness of the optical monitoring method was proved in correlation analysis where quality variables in biologically treated wastewater were found to have the higher correlation with the optical monitoring variable than with any process measurement from the early part of the process.

In addition, five variable selection methods were used to find the optimal subset of variables for developing a simple but useful model to predict the suspended solids, BOD, COD, total nitrogen and total phosphorus content in biologically treated wastewater. Five-fold cross-validation was used to evaluate the performance of the developed models. The model based on the optical monitoring and process variables from the early stage of the WWTP can be used to predict the level of the quality parameters but predicting the exact values is challenging. It has to be pointed out that the optical monitoring was performed from the one process line and the analysed samples of biologically treated wastewater contained the wastewater from all the nine parallel lines. It is also important to bear in mind that the results based solely on a mathematical analysis may not accurately correspond the reality in the process and that a high correlation coefficient between variables not always mean strong real-world causality. Again, indirect effects between parameters may occur in the process even though they are not revealed on a mathematical data analysis. The modelling results could be improved in the future by using a more complicated modelling method or including some additional information to the models.

The study confirmed that the optical monitoring method is a valuable tool for monitoring the wastewater treatment process and the changes in floc morphology. The received new information gives better understanding about the quality changes of the effluent and the reasons for problems in the process. The advantages of the on-line optical monitoring against traditional microscopic monitoring are remarkable: the method is objective, continuous, fast, includes several morphological characterisation variables and enables observing the changes at an early stage before they show as problems in the sludge and treated water quality. The optical monitoring combined to predictive modelling has potential to be utilized in controlling the process, keeping it in stable operating conditions and avoiding environmental risks.

Acknowledgements

The data used in this study was collected during the Measurement, Monitoring and Environmental Efficiency Assessment (MMEA), the research programme of CLEEN Ltd. – Cluster for Energy and Environment. Aki Sorsa, D.Sc. (Tech.), is greatly acknowledged for assisting in variable selection issues. Maa- ja vesitekniikan tuki ry supported the finalizing this study and is thereby greatly acknowledged.

References

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

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

[3] Mesquita D. P., Dias O., Amaral A. L., Ferreira E. C. Monitoring of activated sludge settling ability through image analysis: validation on full-scale wastewater treatment plants. Bioprocess Biosyst Eng 2009;32:361–367.

[4] 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.

[5] Banadda E. N., Smets I. Y., Jenne R., Van Impe J. F. Predicting the onset of filamentous bulking in biological wastewater treatment systems by exploiting image analysis information. Bioprocess Biosyst Eng. 2005;27:339–348.

[6] 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.

[7] Yu R-F., Chen H-W., Cheng W-P., Chu M-L. Simultaneously monitoring the particle size distribution, morphology and suspended solids concentration in wastewater applying digital image analysis (DIA). Environ. Monit. Assess. 2009;148:19–26.

[8] HSY Viikinmäki wastewater treatment plant webpage. Cited May 2015. Available from https://www.hsy.fi/en/experts/water-services/wastewater-treatment-plants/viikinmaki/Pages/default.aspx

[9] Mesquita D.P., Amaral A.L., Ferreira E.C. Activated sludge characterization through microscopy: A review on quantitative image analysis and chemometric techniques. Analytica Chimica Acta. 2013;802:14–28.

[10] 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.

[11] Russ J. Computer-assisted Microscopy, Plenum Press, New York.;1990.

[12] Bushell G., Yan Y., Woodfield D., Raper J., Amal R. On techniques for the measurement of the mass fractal dimension of aggregates, Adv. Colloid Interface Sci. 2012;95:1–50.

[13] 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. http://urn.fi/urn:isbn:9789526202891

[14] Juuso E.K. Integration of Intelligent Systems in Development of Smart Adaptive Systems. International Journal of Approximate Reasoning. 2004;35:3:307-337.

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

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

[17] Sorsa A., Leiviskä K., Santaaho 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.

[18] MathWorks Statistics and Machine Learning Toolbox documentation. Se.mathworks.com, sited May 2015

[19] Araujo M. C. U., Saldanha T. C. B., Galvao R. K. H., Yoneyama T., Chame H. C., Visani V. The successive projections algorithm for variable selection in spectroscopic multicomponent analysis. Chemometrics and Intelligent Laboratory Systems, 2001;57:65–73.

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

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

[22] Rao R.B., Fung G., Rosales R. On the dangers of cross-validation: An experimental evaluation. In SIAM Data Mining, Philadelphia, PA. 2008.

[23] Arlot S., Celisse A. A survey of cross-validation procedures for model selection. Statistics Surveys. 2010;4:40–79. [1] Tchobanoglous G., Burton F. L., Stense H. D. Wastewater Engineering: Treatment and Reuse. 4th edition. 1819 p.; 2003

Table 1. Raw data values of selected measurements.

		min	max	median	std
(B) Suspended solids	mg/l	4.8	32.3	8.9	6.2
(B) BOD	mg/l	3.8	21.0	6.7	3.4
(B) COD	mg/l	36.0	72.0	50.0	7.9
(I) Suspended solids	mg/l	132.0	926.7	278.8	111.0
(I) BOD	mg/l	163.3	481.1	255.8	51.0
(I) COD	mg/l	275.0	1309.0	584.0	155.3
Flow	m^3/d	24364	51235	32961	5549
Anoxic proportion of vol	%	16.0	50.0	34.1	9.2
Temperature	ºC	10.7	20.4	14.8	2.7
Sludge concentration	g/l	2.0	4.6	3.1	0.6
Sludge age	d	6.0	15.5	8.5	3.2
Length of filaments	µm	115.7	2503.5	1130.6	626.9
Amount of filaments		0.000264	0.0026	0.0012	0.000478
Fractal dimension		1.5715	1.6104	1.5843	0.0104
Roundness		0.5114	0.6431	0.5655	0.0289
Equivalent diameter	µm	60.6	75.8	67.3	3.9
Number of objects		71.7	399.4	212.2	83.1
Number of small objects		61.2	586.6	267.3	123.2

(B) biologically treated wastewater, (I) influent

Table 2. The correlation coefficients of selected image analysis variables and WWTP variables.

	Filament length	Amount of filaments	Fractal dimension	Roundness	Aspect ratio	Median area of objects	Number of small objects
Filament length		0.82	-0.92	-0.91	0.71	-0.50	0.93
Total floc area	0.90	0.50	-0.82	-0.92	0.66		0.87
Amount of filaments	0.82		-0.78	-0.64	0.60	-0.75	0.72
Roundness	-0.91	-0.64	0.94		-0.88		-0.86
Aspect ratio	0.71	0.60	-0.86	-0.88			0.63
Equivalent diameter		-0.68				0.96	
Number of objects	0.93	0.68	-0.90	-0.86	0.62	-0.57	0.99
(B) BOD	0.55	0.50	-0.58	-0.59	0.62		
(B) Suspended solids	0.68	0.66	-0.73	-0.66	0.63		0.65
(B) Total phosphorus	0.67	0.68	-0.70	-0.58	0.51	-0.55	0.67
(B) Iron	0.51	0.54	-0.58		0.51		0.50
(E) Total nitrogen	0.52		-0.58	-0.59	0.59		0.52
(E) Ammonium nitrogen				-0.51	0.55		
(E) Nitrate-nitrogen		0.57				-0.54	
(E) pH			0.56	0.55	-0.67		

(B) biologically treated wastewater, (E) effluent

Table 3. The correlation coefficients of selected image analysis variables and important operating condition variables.

	(B) BOD	(B) COD	(B) Suspended solids	Flow	Anoxic proportion of volume	Temperature	Sludge concentration
Filament length	0.55		0.68			-0.82	
Total floc area			0.50			-0.72	0.52
Amount of filaments	0.50		0.66		-0.55	-0.71	
Fractal dimension	-0.58	-0.53	-0.73			0.83	
Roundness	-0.59		-0.66			0.84	-0.52
Aspect ratio	0.62		0.63			-0.78	
Number of objects			0.63			-0.74	
Number of small objects			0.65			-0.75	
(B) BOD		0.70	0.75		-0.50	-0.51	
(B) COD	0.70		0.73		-0.52		
(B) Suspended solids	0.75	0.73			-0.67	-0.62	
(I) Total phosphorus				-0.69		0.56	
(B) Total phosphorus	0.66	0.64	0.86		-0.58	-0.57	
(I) Total nitrogen				-0.73		0.58	
(E) Total nitrogen			0.59		-0.51	-0.60	
(I) Ammonium nitrogen				-0.70		0.67	
(M) Ammonium nitrogen				-0.73		0.65	
(B) Ammonium nitrogen				0.51		-0.57	
(I) Nitrate nitrogen				0.56			
(M) Nitrate nitrogen			0.61				
(I) Alkalinity				-0.71		0.54	
(M) Alkalinity				-0.70		0.54	
(E) pH			-0.56				
(B) Iron	0.71	0.72	0.91		-0.61		

(I) influent, (M) mechanically treated wastewater, (B) biologically treated wastewater, (E) effluent

Table 4. Variable selection for suspended solids content in biologically treated wastewater.

Correlation analysis	Stepwise selection	Forward selection	Genetic algorithm	Successive projections algorithm + GA	Manual selection	Variables	
5	5	5	1	1	5	1	Filament length
46	46	46	30	12	46	3	Amount of filaments
3	35	35	35	46	35	5	Fractal dimension
47	30	30	41	41	30	9	Aspect ratio
35	41	41	44	35		12	Median area of objects
12	47	44	46	27		27	(M) Total phosphorus
	44	47	47	44		30	(I) Total nitrogen
				9		35	(M) Nitrate-nitrogen
						41	(I) Sulphate
						44	(M) Iron
						46	Anoxic proportion of vol
						47	Temperature

(I) influent, (M) mechanically treated wastewater

Table 5. Variable selection for BOD in biologically treated wastewater.

Correlation analysis	Stepwise selection	Forward selection	Genetic algorithm	Successive projections algorithm + GA	Manual selection	Variables	
9	9	9	5	1	5	1	Filament length
47	46	46	12	46	12	2	Total floc area
3	44	44	2	44	2	3	Amount of filaments
46	1	1	16	9	16	5	Fractal dimension
17			27		27	9	Aspect ratio
35			41			12	Median area of objects
			44			16	Number of objects
			46			17	Number of small objects
						27	(M) Total phosphorus
						35	(M) Nitrate-nitrogen
						41	(I) Sulphate
						44	(M) Iron
						46	Anoxic proportion of vol
						47	Temperature

(I) influent, (M) mechanically treated wastewater

Table 6. Variable selection for COD in biologically treated wastewater.

Correlation analysis	Stepwise selection	Forward selection	Genetic algorithm	Successive projections algorithm + GA	Manual selection	Variables	
5	7	46	17	46	5	3	Amount of filaments
46	46	21	19	29	7	5	Fractal dimension
3	19	17	29	19	19	7	Form factor
12	5	29	46	48	29	12	Median area of objects
21	48	48	48	16	46	16	Number of objects
47	29	19			48	17	Number of small objects
						19	(M) BOD
						21	(M) COD
						29	(M) PO4-P
						46	Anoxic proportion of vol
						47	Temperature
						48	Sludge concentration

(M) mechanically treated wastewater

Table 7. Variable selection for total nitrogen content in biologically treated wastewater.

Correlation analysis	Stepwise selection	Forward selection	Genetic algorithm	Successive projections algorithm + GA	Variables	
31	31	31	2	46	2	Total floc area
38	12	12	8	39	8	Roundness
12	27	27	21	48	9	Aspect ratio
39	43	46	22	27	10	Equivalent diameter
11	46	39	27	9	11	Mean area of objects
46	39	9	33	31	12	Median area of objects
		16	47		16	Number of objects
		10	48		21	(M) COD
					22	(I) Suspended solids
					27	(M) Total phosphorus
					31	(M) Total nitrogen
					33	(M) Ammonium nitrogen
					38	(I) pH
					39	(M) pH
					43	(I) Iron
					46	Anoxic proportion of vol
					47	Temperature
					48	Sludge concentration

(I) influent, (M) mechanically treated wastewater

Table 8. Variable selection for total phosphorus content in biologically treated wastewater.

Correlation analysis	Stepwise selection	Forward selection	Genetic algorithm	Successive projections algorithm + GA	Variables	
5	5	5	1	1	1	Filament length
3	46	46	2	12	2	Total floc area
46	27	27	16	46	3	Amount of filaments
47	9	1	27	27	5	Fractal dimension
12	3	2	41		9	Aspect ratio
35		16	46		12	Median area of objects
		41			16	Number of objects
					27	(M) Total phosphorus
					35	(M) Nitrate-nitrogen
					41	(I) Sulphate
					46	Anoxic proportion of vol
					47	Temperature

(I) influent, (M) mechanically treated wastewater

Table 9. The fitness of developed models for suspended solids, biological oxygen demand, chemical oxygen demand, total nitrogen and total phosphorus in biologically treated wastewater.

	SS		BOD		COD		Nitrogen		Phosphorus	
Variable selection method	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE
Correlation analysis	0.71	0.55	0.45	0.71	0.45	0.71	0.47	0.73	0.57	0.61
Stepwise selection	0.79	0.47	0.50	0.67	0.55	0.63	0.54	0.68	0.67	0.54
Forward selection	0.78	0.48	0.50	0.67	0.54	0.64	0.58	0.65	0.69	0.52
Genetic algorithms	0.79	0.47	0.55	0.64	0.55	0.64	0.63	0.61	0.69	0.52
SPA+GA	0.78	0.48	0.50	0.67	0.54	0.64	0.54	0.68	0.66	0.54
Manual selection	0.74	0.51	0.43	0.72	0.56	0.63				

Figure 1. The wastewater treatment process at Viikinmäki. Modified from [8]

Figure 2. The on-line optical monitoring device for imaging activated sludge process. Modified from [4]

Figure 3. (A) The feasible range, (B) scaled value, and (C) membership functions. Redesigned from [14].

Figure 4. Selected process variables of the WWTP and image analysis variables. (B) biologically treated wastewater, (I) influent.