DOI: 10.1051/e3sconf/20171700089

Application of the selected classification models to the analysis of the settling capacity of the activated sludge – case study

Bartosz Szeląg[1,*], and *Piotr* Siwicki[2]

[1]Kielce University of Technology, Faculty of Environmental, Geomatic and Energy Engineering, Tysiąclecia Państwa Polskiego Ave. 7, Poland
[2]Warsaw University of Life Sciences - SGGW, Faculty of Civil and Environmental Engineering, Nowoursynowska St.166, Poland

Abstract. The study presents the development of classification models for sedimentation of activated sludge using the artificial neural networks (ANN), logistic-regression (RL), and linear discrimination model (LDM). The input consisted of indicators of wastewater quantity and quality (biochemical oxygen demand, chemical oxygen demand, total suspended solids, total nitrogen and total phosphorus at the inflow to the wastewater treatment plant) and operational characteristic of the bioreactor (pH, temperature of activated sludge, mixed liquor suspended solids, concentration of oxygen in the nitrification chamber, amount of PIX dosing).The prediction quality of the developed models was measured with: sensitivity, specificity, and computed errors. The calculations of the sedimentation were performed for sludge volume index (SVI). The results indicate that successful predictions were obtained using ANN, RL and LDM methods, which is supported by the fit of computations to measurement results. The study shows that for the wastewater treatment plant of concern, sedimentation properties can be obtained using only the loads of organic compounds, mixed liquor suspended solids, temperature, pH of activated sludge, concentration of oxygen in the nitrification chamber and amount of PIX dozing. Other analysed variables appear to be statistically insignificant for the sludge volume index.

1 Introduction

Engineering practice shows that the activated sludge method is one of the most frequently used and most effective technologies of domestic wastewater treatment. Sludge volume index (SVI) is one of the parameters employed to assess the performance of the wastewater treatment plant [1-3]. Sludge volume index makes it possible to predict metastable states, the occurrence of which is related to the excessive growth of filamentous bacteria. During wastewater clarifications in secondary settlement tanks, the presence of those bacteria can lead to an increase in the content of total suspended solids and of organic carbon in the outflow from the settler. Therefore, to be able to predict metastable states in

[*] Corresponding author: bszelag@tu.kielce.pl

good advance, and to make it possible for the treatment plant staff to select optimum settings of the bioreactor parameters to ensure effective wastewater treatment, it is necessary to develop mathematical models that are capable of determining continuous value or the linguistic value. In the former case, the discrete values of the measured dependent variable are predicted [4–6]. In the latter case, the range of variation of the measurement results provides a basis for a division into classes. The ranges of variation of the parameter under analysis, set by the researcher, correspond to those classes [7–8].

Numerous regressive methods are used to model the discrete value of the sludge volume index. The methods include Artificial Neural Networks and their modifications, Support Vector Machines, Random Forests, Boosted Trees, and k-Nearest Neighbour. For the analysis of the settling capacities, the classification models based on the methods listed above are also used. In those models, the SVI value is described by a binary (zero-one) variable, or linguistic values are assigned to SVI value ranges. In the classification models, the number of independent variables describing a given phenomenon can be smaller than in the regression models, because it is not the discrete value that is modelled but exclusively the class membership. That fact is of considerable importance in practical applications, especially those concerning wastewater treatment facilities. That means it is possible to reduce the number of variables needed to describe a given phenomenon. With the classification approach, sludge settling capacity can be estimated relying on the extreme SVI values. Additionally, SVI values can be corrected by means of altering the facility settings [9]. In the regression and classification black box methods, which include the ones mentioned above, the model structure being a basis for a given dependent variable prediction is generated at the training stage. However, to develop the model, it is necessary to implement complex numerical algorithms. Occasionally, model development would require designing specialist computer programs. Although such models show satisfactory predictive abilities in estimating SVI, their practical applications to technological areas will involve the use of additional software and make it necessary to integrate the computational and measurement modules, which can be a difficult task. The literature review shows that the logistic regression model was used to analyse the settling capacity. However, the correlation produced by [7] and also [8] did not allow the possibility of correcting the bioreactor technological parameters in order to improve the settling capacity of the activated sludge. In economic, and also medical and social sciences, to analyse binary data the so-called linear discriminant model is used. The results of computations demonstrate that the model has better predictive abilities than logit regression models or black box methods of ANN-type, and others.

The objective of this paper is to compare predictive abilities of the selected classification methods for the control of the settling capacity of the activated sludge. The paper analyses the possibility of modelling sludge settleability with the use of the black box models and dichotomy models, such as the logistic regression model and linear discriminant model. Additionally, the aim of the analyses is to develop a simple regression model for the continuous control and monitoring of the settling capacity of the activated sludge. That could be based on the quality indexes of wastewater influent and operational parameters of aeration tanks. Thus, it would not be necessary to implement complex numerical algorithms, or install additional measurement systems at the wastewater treatment plant.

2 The object of investigation

The analyses were conducted for the urban wastewater treatment plan (WWTP) having the design capacity of 72.000 m^3/d, located in the terrain of the commune of Sitkówka - Nowiny. The plant collects wastewater from the city of Kielce, the commune of Sitkówka -

Nowiny, and partially also from the commune of Masłów. The wastewater delivered to WWTP is mechanically pretreated using step screens and aerated grit chambers with grease separators. Next, wastewater is pumped to primary clarifiers, from which it is delivered to the biological unit, i.e. bioreactor with dephosphatation, denitrification and nitrification tanks. Then, together with activated sludge, wastewater is transferred to four secondary clarifiers, and after clarification it flows to the receiving water, i.e. the river Bobrza.

Continuous monitoring conducted by the company Wodociągi Kieleckie Sp. z o.o. at the treatment plant provides measurements of influent wastewater quantity, and also of the operational parameters of the bioreactor (Table 1).

3 Methodology

In this paper, the linear discriminant model, the logit model and the artificial neural network were employed to analyse sludge settling capacity. The results of measurements of influent wastewater quality and the reactor operational parameters were utilised to this end. The former included the following: biochemical oxygen demand, the content of ammonia, nitrate and nitrite nitrogen, total suspended solids, and total phosphorus. The bioreactor parameters were as follows**:** mixed liquor suspended solids (MLSS), activated sludge temperature and pH, amount of dosage of ferric coagulants (PIX), oxygen concentration in the nitrification tank. To identify the dependence between the variables of concern and activated sludge index, a correlation matrix was constructed based on the measurements of wastewater quantity and quality and the bioreactor parameters. In the analyses, the criterion for the assessment of the impact of the bioreactor operational parameters and influent wastewater quality was the value of the sludge volume index (SVI), equal to SVI_{lim} = 150 cm^3/g [7]. When the measured value of SVI was greater than SVI_{lim}, it was assigned the value of 1, otherwise it was given the value of 0.

Linear discriminant model (LDM) and logistic regression are used to analyse binary data and they offer the possibility of predicting the occurrence or non-occurrence of an event. In addition, those models are frequently applied to economical, and also medical and social sciences. The logit model is also often employed in river engineering, geomorphology, environmental protection, microbiology, wastewater treatment, and in analyses of the operation of components of sewage systems [8,10,11]. In addition to the applications above, the linear discriminant model was also used to assess the condition of stormwater drainage pipe and to analyse landslide occurrence [12,13]. Although linear discriminant model (LDM) has much better predictive abilities than the logit model, it has not been used for the analysis of the settling capacity of the activated sludge. The dichotomy models mentioned above can be described by the following dependences:

- the logit model:

$$p(\mathbf{X}) = \frac{exp(\mathbf{X})}{1+exp(\mathbf{X})} \tag{1}$$

- the linear discriminant model:

$$Z(\mathbf{X})_k = \mathbf{X}_k \tag{2}$$

in which: k – number of classes to be separated, and thus the number of discrimination functions; in the problem under consideration k = 2; in the first case, function Z_1 describes membership in a group, in which SVI < 150 cm^3/g, whereas function Z_2 expresses combinations of linear parameters x_i, for which SVI > 150cm^3/g, p – probability of exceeding the limiting value SVI_{lim} that is the basis for sludge volume index prediction, which correspond to the value p=0.50; X – vector made from linear combination of independent variables, expressed by the formula:

$$\mathbf{X} = \sum_{i=1}^{j} \beta_i \cdot x_i + \beta_0 \tag{3}$$

where: x_i – variables describing sludge volume index, which include the following: MLSS, sludge temperature and pH, amount of dosage of ferric coagulants (PIX), the content of ammonia nitrogen and that of organic substances in influent wastewater, and others; β_0, β_i – coefficients determined with the maximum likelihood method.

The advantage offered by the logit model is the possibility of interpreting the value of $\exp(\beta_i)$. For that purpose, the notion of chance value is used. It is defined as the ratio of the probability of an event occurrence to the probability of its non-occurrence (1 – p). The effect of the increment of independent variables by the value of Δx_i (i = 1, 2, …, j) on the chance of the event occurrence can be computed on the basis of the so-called chance ratio expressed as $OR=\exp(\Sigma\beta_i \cdot \Delta x_i)$.

Neural networks can be used to simulate linear and non-linear processes, optimisation, classification and control [5, 6, 14]. For the sake of modelling the processes taking place in the environment, the network termed the multilayer perceptron (MLP) is used most frequently. In MLP networks, input signals are multiplied by the values of weights and then transferred to the neurons of the hidden layer, where summation occurs in individual neurons. The sums received are transformed using a nonlinear activation function and conveyed to the output neurons. The optimal structure of the neural network for the prediction of SVI binary value was sought by computing the measures of fit between simulation results and measurements (SENS, SPEC, R_z^2). It was assumed that number of neurons in the hidden layer was (3–10) and the activation function was linear, expotential, sine, sigmoidal and hyperbolic tangent. For the output layer, the linear activation function was adopted.

The accuracy of predictions obtained with the statistical models was established based on the values of sensitivity (SENS), specificity (SPEC) and computed error (R_z^2), described in detailed in the study by Harrell [15], and other papers.

4 Results

Based on the results of measurements of wastewater quantity and quality and the bioreactor technological parameters, the range of parameter variation was determined (Table 1) in order to establish the applicability of the developed classification models.

The data provided in Table 1 indicate that daily wastewater inflows to the WWTP and the wastewater quality indexes varied substantially. That considerably affected the operational parameters of aeration tanks; the value of the mixed liquor suspended solids (MLSS) ranged 1.91–6.59 kg/m^3 and oxygen concentration in the nitrification tank changed 0.55–5.78 mg/dm^3, whereas the mean value of dissolved oxygen (DO) was 2.56 mg/dm^3. The data in Table 1 demonstrate that in the period of concern, the settling capacity of the activated sludge deteriorated, which is indicated by both the mean value of SVI = 186 cm^3/g and the maximum one that was 320 cm^3/g.

Based on the measurement results (Table 1), the values of the coefficient of correlation (r) between the considered variables describing influent quality and the bioreactor technological parameters were determined (Table 2).

Table 1. Range of variation of values of parameters describing wastewater quantity (Q) and quality (BOD_5, TSS, NH_4^+-N, TN) and aeration tank operational parameters (T_{sl}, pH, MLSS, RAS, PIX, DO).

Variable	Minimum	Mean	Maximum
Q, m^3/d	32564	40698	86592
T_{sl}, ^0C	10.0	15.9	23.0
pH	7.2	7.7	7.8
MLSS, kg/m^3	1.98	4.26	6.59
PIX, m^3/d	0.0	0.8	1.93
DO, mg/dm^3	0.55	2.56	5.78
SVI, cm^3/g	95	186	320
BOD_5, mg/dm^3	127	309	557
TSS, mg/dm^3	126	329	572
NH_4^+-N, mg/dm^3	24.4	49.4	65.9
TN, mg/dm^3	39.9	77.7	124.1

Table 2. Values of the coefficient of correlation (r) between the variables considered and the sludge volume index.

Variables	BOD_5	COD	TSS	NH_4^+-N	TN	pH	T_{sl}	MLSS	DO	PIX	Q
r	0.21	0.15	0.17	0.25	0.19	-0.21	-0.46	-0.61	-0.17	-0.12	-0.10

On the basis of the data in Table 2, it can be concluded that weak correlations hold between the value of sludge volume index and BOD_5 (r=0.21), COD (r=0.15), TSS (r=0.17), NH_4^+-N (r=0.25), and TN (r=0.19) in the influent. In addition, weak correlation was observed between SVI value and daily inflow into the wastewater treatment plant (r=-0.10). As regards the bioreactor parameters, such as DO, pH, and PIX, a weak dependence between SVI and the parameters of concern was found. Additionally, Table 2 indicates high and medium correlation between sludge volume index and activated sludge temperature (r =-0.46) and MLSS (r=-0.61). The results obtained (Table 1) are confirmed by investigations conducted by Lou and Zhao [4], Bayo et al.[7]. Due to the fact that the value of the sludge volume index alone does not give direct information on the aeration tank loading, in further analyses pollutant loads were introduced (L_{BOD5}). Based on the measurement data (Table 1) and the results of analyses (Table 2), binary variables were developed to describe sludge volume index. By applying the step regression method, variables x_i and coefficients β_i were determined in the model described by Eqs. (1) and (2). In the case under consideration, the equations can be expressed by the following dependences:

- the logistic regression:

$$p(\mathbf{X}) = \frac{\exp(15.79-0.39\cdot T_{sl}-1.33\cdot MLSS-1.83\cdot PIX-1.18\cdot DO+0.00008\cdot L_{BOD5})}{1+\exp(15.79-0.39\cdot T_{sl}-1.33\cdot MLSS-1.83\cdot PIX-1.18\cdot DO+0.00008\cdot L_{BOD})} \tag{4}$$

- the linear discriminant model:

$$Z = \begin{cases} 18,66\cdot MLSS - 11,36\cdot T_{sl} - 367,50\cdot pH - 7,48\cdot PIX + 0,001 L_{BOD5} - 0,14\cdot DO - 1528,19 \\ 20,24\cdot MLSS + 11,92\cdot T_{sl} + 370,19\cdot pH + 9,18\cdot PIX - 0,001 L_{BOD5} + 1,44\cdot DO - 1567,04 \end{cases} \tag{5}$$

As regards neural networks, the best classification abilities are demonstrated by the model that has 7 neurons in the hidden layer and the activation function takes the form of the hyperbolic tangent. The comparison of the parameters describing the fit of computational results to measurements, namely SPEC, SENS, and R_z^2 are presented in Table 3. In the logistic regression and LDA models, the remaining variables, i.e. pH, recirculation ratio, total suspended solids, total and ammonia nitrogen turned out to be statistically insignificant at the adopted confidence level p = 0.05. In the classification model based on artificial neural network method, the independent variables of SVI value, given in Eqs. (4) and (5), were adopted. The data listed in Table 3 show that the classification models based on the logit model and artificial neural networks have identical predictive abilities.

Table 3. Parameters of fit of computational results to measurements (SPEC, SENS, R_z^2) in models for sludge volume index classification.

Parameter	Methods		
	logit	LDM	ANN
SPEC	0.827	0.804	0.875
SENS	0.873	0.841	0.826
R_z^2	0.854	0.826	0.854

The lowest fit of computational results to measurements was obtained for the model based on the linear discriminant analysis (LDM). For instance, with the logit model, out of 55 events analysed, during which values of SVI < 150 cm^3/g, 46 events were classified correctly. Out of 55 events, during which sludge settling problems occurred (SVI > 150 cm^3/g), the model classified 49 events correctly. With the model based on the LDA method, out of 55 events considered, during which SVI < 150 cm^3/g, 44 cases were classified correctly, so were 47 events in 55 when SVI > 150 cm^3/g. Due to the fact that the ANN-based statistical model is an implicit dependence SVI = f(x_i) and has identical predictive abilities as the logistic regression model described by Eq.(4), binomial logit model was utilised to identify the regression dependence in further analyses.

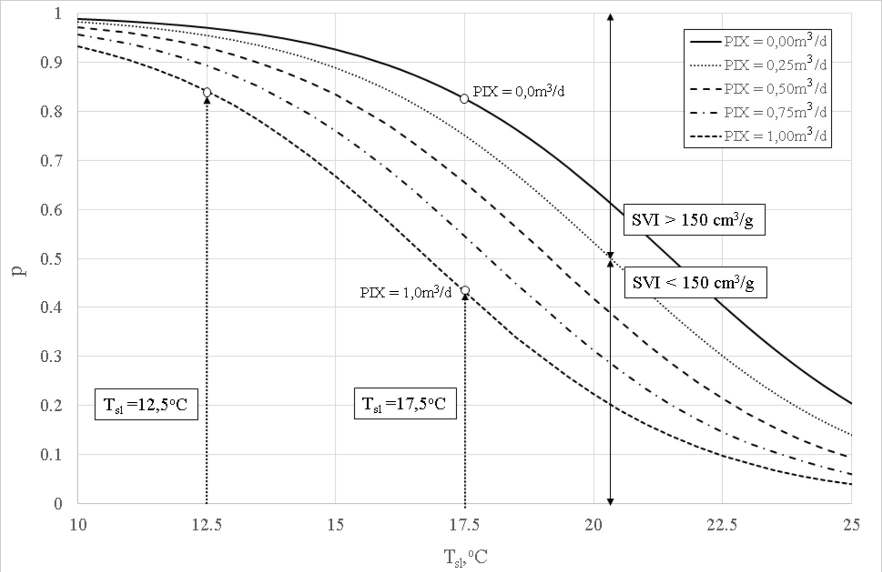

Fig. 1. The effect of activated sludge temperature (T_{sl}) and the amount of dosage of PIX on the probability of the SVI value exceeding SVI_{lim} = 150 cm^3/g.

The logistic regression model, described by formula (2), is an example of a simple regression dependence that makes it possible to find the value of sludge volume index. On the basis of this dependence, curves can be plotted which illustrate the impact of individual variables of the model on the activated sludge settling (Fig. 1).

The curves show the effect of sludge temperature and the amount of PIX dosed on the probability that SVI value will exceed SVI_{lim}=150 cm^3/g for mean values of MLSS, pH, BOD_5 load and oxygen concentration. Additionally, based on the results, chance ratios (OR) were computed. They allow the possibility of assessing the influence of increase in individual independent variables by the value of Δx_i (Table 4) on the risk of the event occurrence (exceeding of SVI_{lim}).

For instance, on the basis of the curves (Fig.1), it can be stated that temperature increase from T_{sl} = 12.5°C to T_{sl} = 17.5°C results in the reduced probability of SVI_{lim} value being exceeded from p = 0.85 to p = 0.42. In the latter case, that means the computed value of the sludge volume index will be lower than 150 cm^3/g. Additionally, the analysis of curves shows that when sludge temperature is T_{sl} = 17.5°C and PIX is not dosed, the probability of SVI_{lim} value being exceeded is equal to p = 0.83.

Table 4. Results of computations of chance ratios (OR) for individual independent variables of SVI.

ΔMLSS, kg/m^3	OR	ΔT_{sl}, ^{0}C	OR	ΔPIX, m^3/d	OR	ΔL_{BOD5}, kg/d	OR	ΔDO, mg/dm^3	OR
0.25	0.72	0.5	0.82	0.25	0.63	2000	1.17	0.1	0.89
0.50	0.51	1.0	0.68	0.50	0.4	3000	1.27	0.2	0.79
0.75	0.37	1.5	0.56	0.75	0.25	4000	1.38	0.3	0.7
1.00	0.26	2.0	0.46	1.00	0.16	5000	1.49	0.4	0.62

When the amount of PIX delivered into the aeration tank is 1m^3/d, the value p computed with formula (2) is 0.40. That means the value of sludge volume index determined for these conditions will be lower than 150 cm^3/g. The values of chance ratios (OR) in Table 4 show that the variables accounted for in the model substantially affect the probability that SVI_{lim} value will be exceeded. The data in Table 4 indicate that an increase in BOD_5 load in the influent by the value of 2000 kg/d leads to increased probability of SVI_{lim} being exceeded by 37%. However, e.g. increase in MLSS or T_{sl} by 0.50 kg/m^3 and 1°C, respectively, results in reduced probability of SVI_{lim} being exceeded by 49% and 32%.

5 Conclusions

On the basis of analyses conducted for the study, it is possible to state that the linear discriminant model, the logit model and artificial neural networks can be applied to the assessment of the settling capacity of the activated sludge. The best classification abilities with respect to sludge volume index were demonstrated by the logit model and artificial neural networks. That was confirmed by the determined values of parameters of simulation fit to the measurement results. Because the ANN model is a complex and implicit dependence that requires the implementation of complex numerical algorithms, its practical application under the treatment facility service conditions is limited. Therefore, for the control and monitoring of the sludge settling capacity, the logistic regression model proposed in the study is readily applicable. It has the form of a simple dependence and it does not require the implementation of complex algorithms as most software packages provide special modules for the computations of the model parameters.

Additionally, the analyses performed for the study demonstrate that the value of sludge volume index (SVI_{lim} = 150 cm^3/g) could be classified with satisfactory accuracy on the basis of the measurement results of the MLSS and activated sludge temperature, oxygen concentration in the nitrification tank and BOD_5 load in the influent wastewater. Taking into account that settling process of activated sludge is extremely complex, it is necessary to conduct further analyses. They should primarily rely on the logit model that would ultimately aim at taking into account the interaction of variables describing wastewater quantity and quality, and also the bioreactor parameters and the sludge biocenosis.

References

1. J. Łomotowski, A. Szpindor, *Nowoczesne systemy oczyszczania ścieków* (2002)
2. A.M.P. Martins, K.R. Pagilla, J.J. Heijnen, M.C.M. Van Loosdrecht, Bulking filamentous sludge - a critical review, Wat. Res. **38**, 793 (2004)
3. Z. Dymaczewski, J.A. Oleszkiewicz, M. M. Sozański, *Poradnik eksploatatora oczyszczalni ścieków*. Wydanie II (1997)
4. I. Lou, Y. Zhao , Sludge Bulking Prediction Using Principle Component Regression and Artificial Neural Network, Math Probl Eng **2012**, 1 (2012)
5. B. Szeląg, J. Gawdzik, Application of Selected Methods of Artificial Intelligence to Activated Sludge Settleability Predictions, Pol J Environ Stud **25**, 1709 (2016)
6. F. Li, J. Qiao J., Y. Han, C. Yang, A self - organizing cascade neural network with random weights for nonlinear system modeling, Appl Soft Comput **42**, 184 (2016)
7. J. Bayo, J.M. Angosto, J. M, J. Serrano-Aniorte, Evaluation of physicochemical parameters influencing bulking episodes in a municipal wastewater treatment plant. *Water Pollution VIII: Modelling, Monitoring and Management,* 531 (2006)
8. E. Bezak-Mazur, R. Stoińska, B. Szeląg, Ocena wpływu parametrów operacyjnych i występowania bakterii nitkowatych na objętościowy indeks osadu czynnego – studium przypadku, Rocznik Ochrona Środowiska **18**, 487 (2016)
9. L. Belanche, J. Valdes, J. Comas, I. Roda, M. Poch, Prediction of the bulking phenomenon in wastewater treatment plants, Artificial Intelligence in Engineering **14**, 307 (2000)
10. J. Łomotowski, M. Dańczuk, Application of the microwave energy to the hygienization of sewage sludge, EPE J **36**, 77 (2010)
11. T. Heyer, J. Stamm, Levee reliability analysis using logistic regression models – abilities, limitations and practical considerations, Georisk Assessment and Management of Risk for Engineered Systems and Geohazards **7**, 77 (2013)
12. D.H. Tran, A.W.M. Ng, B.J.C. Perera, S. Burn, P. Davis, Application of probabilistic neural networks in modelling structural deterioration of stormwater pipes, Urban Water J **3**, 175 (2006)
13. A.M. Ramos-Cañón, L.F. Prada-Sarmiento, M.G. Trujillo-Vela, J.P. Macías, A.C. Santos-R, Linear discriminant analysis to describe the relationship between rainfall and landslides in Bogotá, Colombia, Landslides **13**, 671 (2016)
14. E. Gatnar, *Podejście wielomodelowe w zagadnieniach dyskryminacji i regresji* (2012)
15. F. Harrell, *Regression Modeling Strategies with Application to Linear Models, Logistic Regression, and Survival Analysis* (2001)