DE GRUYTER OPEN

DOI: 10.1515/eces-2017-0009 ECOL CHEM ENG S. 2017;24(1):119-127

Bartosz SZELĄG[1], Alicja GAWDZIK[2] and Andrzej GAWDZIK[2*]

APPLICATION OF SELECTED METHODS OF BLACK BOX FOR MODELLING THE SETTLEABILITY PROCESS IN WASTEWATER TREATMENT PLANT

ZASTOSOWANIE WYBRANYCH METOD CZARNEJ SKRZYNKI DO MODELOWANIA OPADALNOŚCI W OCZYSZCZALNI ŚCIEKÓW

Abstract: The paper described how the results of measurements of inflow wastewater temperature in the chamber, a degree of external and internal recirculation in the biological-mechanical wastewater treatment plant (WWTP) in Cedzyna near Kielce, Poland, were used to make predictions of settleability of activated sludge. Three methods, namely: multivariate adaptive regression splines (MARS), random forests (RF) and modified random forests (RF + SOM) were employed to compute activated sludge settleability. The results of analysis indicate that modified random forests demonstrate the best predictive abilities.

Keywords: multivariate adaptive regression splines, random forests, modified random forests, sludge settleability

Introduction

One of the parameters that determine the efficiency of wastewater treatment sludge technology is the capacity of sedimentation. In the case of its deterioration increase the amount of slurry and coal in wastewater effluent from the secondary settling tank over the limit values in treated sewage. Therefore, in order to prevent the above-mentioned problems, mathematical models should be used for predicting the parameters describing the settleability process. Due to the complex biochemical processes in the settlings and its metastable character, the index of sediment being inherently variable intense, will not be in any case an optimal source of data for forecasting WWTP. Therefore, an alternative approach based on extensive variable expressed by settling seems appropriate. In view of the above observations, it is advisable to create mathematical models to predict the above discussed parameter processing. A review of the literature [1] shows that there were no attempts to develop mathematical models of a physical or statistical character.

[1] Faculty of Environmental, Geomatic and Energy Engineering, Kielce University of Technology, al. Tysiąclecia Państwa Polskiego 7, 25-314 Kielce, Poland
[2] Department of Process Engineering, University of Opole, ul. R. Dmowskiego 7/9, 45-365 Opole, Poland, phone +48 77 401 67 00, email: kip@uni.opole.pl
[*] Corresponding author: kip@uni.opole.pl

To assess the functioning of sewage treatment plants (STP) and the variability of technological parameters of activated sludge and wastewater are used computer programs (WEST, STOAT, BioWIN, SIMBA, etc.), But in order to calibrate mathematical prams in the above-mentioned models, it is necessary to gather information about the parameters that describe the processes occurring in the individual objects of a STP. In practice, there are many interactions between the parameters that describe the processes running on the STP which leads to numerous problems at the stage of calibration models. Therefore, the forecasts of the STP used models of black box, which at the stage of learning model is generated structure model without the need to know the physics of the analyzed phenomenon. One of the most commonly used for this purpose methods are artificial neural networks [2-5], but there are also used other methods such as vectors carrying, tree reinforced models, autoregressive MARS (multivariate adaptive regression splines) and the like. The former method is a modification of the classical model MLP (multi-layer perceptron) where the hidden layer represents a non-linear projection of a vector set of input data characteristics of the space in which they are linearly separable. Trees are strengthened modification of the classical model of regression trees, while in the method MARS relationship is linear regression spline function of a conditional character. Within the methods of the black box an interesting solution is also random forests [6, 7] increasingly used for modeling both the quantity and quality of upstream and downstream of the WWTP as well as forecasting processes taking place in different places of the plant.

When, in the analyzed phenomenon there may occur seasonal changes, it is advisable to its forecasts to create hybrid models combining a model of a classification with regression models mentioned above. In these models the first stage of calculating training set, using a suitable algorithm discriminatory type of method k - medium [8], c - medium [9], self-organizing neural networks [10], is divided into appropriate classes, within which they are made predictions based on independent models. In many cases [11, 12], this approach can improve the ability of predictive mathematical model, but in a situation of extremely diverse number of data in each class or insufficient input data there may be problems at the stage of learning and testing statistical models.

In view of the above comments in the publication presents the concept of the hybrid model as a combination of self-organizing artificial neural networks and methods of random forests. In this method, at the stage of classification, data from the training set are divided into classes and information about the appropriate allocation to them is an additional variable in the forest random model. A hybrid model presented in this paper was used to predict settleability. The results of calculations were compared with the results of the analyzes received by MARS and random forests methods.

The object of investigations

The facility in Cedzyna located in the province Swietokrzyskie, Poland, is a mechanical-biological WWTP with aerobic sludge stabilization having a capacity Q_{av} of 1215 m^3/d and *EP* of 9466. Raw sewage inflowing via sanitation channel go to the grit chamber and grease trap and then to the pumping station, where they are introduced a biological reactor, wherein the effluent of the aerobic part are recycled to the anaerobic section. Wastewater from biological reactor flows into the secondary settling tank, from where it is pumped for recirculation to the biological deposit. Purified wastewater is discharged into the Lubrzanka river.

Methodology

In the present study to determine the settleability (*SE*) in the biological reactor chamber the results of the following measurements were used: the temperature of waste water in the chamber (*T*), the inflow (*Q*), degree of external (REC_{ext}) and inernal (REC_{int}) recirculation conducted during the period from 17.06.2011 to 31.12.2015. To determine settleability some methods of black box were used, which include MARS, random forests and modified random forests method basing on self-organizing neural networks.

In this paper examined combinations of variables described above, where in consideration of each analyzed scenario that takes into account the degree of external and internal recirculation. Results of analyzes examined the possibility of predicting the value of settleability *SE*(*t*) on the basis of the temperature of the sludge *T*(*t*) and inflow *Q*(*t*) to the WWTP. Next, we analyzed the possibility of predicting the *SE*(*t*) value basing only on the instantaneous values of inflows to the WWTP - *Q*(*t*-1) and *Q*(*t*-2) and the temperatures *T*(*t*-1) and *T*(*t*-2). In the last stage of the analysis examined variants that include the ability to model the *SE*(*t*) values on the basis of the previous measurement results of the settleability.

Prior to the creation of mathematical models the input and output signals should be standardized using the equation [4, 13]:

$$\bar{x}_i = \frac{x_i - min.X}{maxX - min.X} \tag{1}$$

wherein: $\bar{x}_i$ - the normalized value of *i*-th element of *X* by min-max method, x_i - the value of *i*-th element of *X* registered at the time of measurement, max *X* - the maximum value of a single element in the parameters set *X*, min. *X* - the minimum value of a single element of the set *X*.

The methods of black box utilized for determining the settleability process in this work are shortly described in the next part of this paper.

MARS (Multivariate adaptive regression splines)

Method MARS is one of the many exploration methods for solving problems of regression and is an extension of the classical approach of explanatory variables in the regression model. Besides taking into account the overall impact of predictors (as in the classical regression model) method MARS ranges of variation in the input data are divided into compartments in which individual predictors may have different effects on sludge settleability. The limits of separation are based on the threshold values (*t*), which means that depending on whether the variable has a value below or above the parameter *t*, this can be included in the model with various weights, or other signs. The distinction between the various input variables for smaller and larger values than the threshold value t_i is made using the basis function [14]:

$$h(X) = \alpha_i \cdot (max(0, X - t)) \tag{2}$$

where *h*(*X*) - vector of the basis functions for the selected variables for which the relationship is satisfied:

$$x_i - t_i = \begin{cases} x_i - t_i; & for \quad x_i > t_i \\ 0; & for \quad x_i \le t_i \end{cases} \tag{3}$$

The regression relationship in the MARS method makes the spline function obtained from the product of the linear combination of the basis functions and respective weights, which can be expressed:

$$f(X) = \alpha_0 + \sum_{m=1}^{M} \alpha_m \cdot h_m(X) \tag{4}$$

where: $X = [x_1, x_2, ..., x_i]$ - vector input, α_m - the weight, h_m - basis functions.

Equation (4) shows that the mathematical model of method MARS can be represented as a weighted sum of the selected basis functions from all of the available functions $h(x_i)$, taking into account the values of the explanatory variables included in the input data. For the estimation of the model parameters, a special algorithm is used, by means of which observation space is searched to determine the threshold values (nodes) [14].

RF (Random forests)

Random forests algorithm was proposed in paper [15]. In the first stage sampling from the training set of n elements is carried out *k* times, permitting repetition. Subsequently, based on the received collections are created regression trees, where the construction process is modified so that each node of the tree makes the best division which is not based on all attributes but only on the drawn ones (explanatory variables). In this way, a *k* - regression trees forming the forest are obtained on the basis of which the forecast is determined, which consists in calculating the arithmetic average of the individual forecasts of individual trees as a result of the entire model.

The most important advantages of random forests is their empirically demonstrated effectiveness [15] and a small number of parameters control their activity. In this case, the user must actually primarily decide how many trees will comprise forest. This is a significant advantage compared to the neural network or a support vector regression method, where the number of parameters (and thus the number of possible variants of settings) is large. Random forests algorithm can not check out a significant number of explanatory information, where only a small number of them has a decisive influence on the analyzed phenomenon. In this publication, due to the small number of variables, there is no such risk.

SOM (Self-organizing neural networks)

The most common type of network referred to as self-organized have been presented in the paper [16]. They belong to the network learned without a teacher, which means that at the training stage, for setting the input is not provided on the output patterns. So, the task of the network is to create these patterns on the stage of learning and self-classify the input data. With the algorithm, which is used in most of the SOM neurons representing similar classes which are next to each other to form an ordered map, making it possible to determine the relationship between the classes received. Self-organizing neural networks are composed of two layers: the input and output (competitive) in the form of one or

two-dimensional array. In contrast to other types of networks discussed neural networks do not have the hidden layer, each neuron of competitive layer is connected to all the neurons of the input layer. Discussed neural network is one way in which individual neurons are connected to all components of the n - dimensional input vector X. Scales connections of neurons form the vector $w_i = [w_{i1}, w_{i2}, .., w_{in}]^T$. Vector of input signals before learning process is subjected to a standardized, which can be written:

$$\bar{x}_i = \frac{x_i}{\sqrt{\sum_{k-1}^{n} (x_k)^2}} \tag{5}$$

Once stimulated the network by the input vector x the competition stage will win such a neuron, for which the weights at least differ from the corresponding components of the vector x. Winner, the w_m neuron, satisfies the relationship [16]:

$$d(x,w_m) = \min_{1<i<n} d(x,w_i) \quad (i = 1,2,\ldots,n) \tag{6}$$

where: w_m - weight vector of the neuron winning, $d(x, w_i)$ - distance (usually Euclidean) between the vector presented input pattern (x) and the weight vector (w_i), i - the number of outputs.

At the stage of learning self-organizing neural network is defined by the number of neurons forming topological layers and the number of teaching periods.

Considering the fact that the methods of classification are used in conjunction with models of regression and improve the ability of predictive mathematical models, which is why they are being increasingly used in the discussion of practical documented through publications [11, 12, 17]. The review of the literature shows [10, 18], that these analyzes are made by hybrid models which are usually a combination of self-organizing artificial neural network with another type of neural network (probabilistic, multi-layer perceptron, recurrent). However, the above-described approaches has its disadvantages, especially in case they have a small amount of input data or when the yields as a result of the classification is characterized by significant differences. In these cases there may be problems at the stage of learning and testing a regression mathematical model. For this purpose, it seems logical to use the discriminatory model for the distribution of training set for the classification purpose of $X = [x_1, x_2, ..., x_n]$ and determine the vector:

$$u = [u_1, u_2, u_3, \ldots, u_n], \text{ where: } u_i = 0 \text{ or } 1 \ (i = 1, 2, \ldots, n) \tag{7}$$
$$\text{and } u_i = 1 \rightarrow u_{n \neq i} = 0$$

constituting an additional explanatory variable (Boolean) in a mathematical model of classification or regression. Noteworthy is the fact that the approach outlined above requires no additional separation of training data on the appropriate classification. The resultant vector u of using discriminatory model complements the input data contained in the form of a vector X. Finally, one can say that predictor explaining the value of y (independent variable) in this case (vectors carrying, artificial neural networks, tree reinforced etc.) is a vector $X' = [X, u]$. In the present study to determine the vector u being the basis for classification of input data applied self-organizing Kohonen neural network type of a topological layer $2 \cdot 2$, while the number of periods at the stage of learning was a 1000. The thus obtained vector (u) provide an additional variable of a Boolean model to forecast settleability using random forests method in which they have already taken into

account the sediment temperature, instantaneous flow and the value of *SE* in the final measurements. In order to carry out a proper learning of the analyzed in this study models (RF, RF + SOM, MARS) and the appropriate evaluation of their operation, a division of the data sets (learner - 50%, validating - 25% and the test - 25%) have been done.

Criteria for evaluation of the models

In order to assess the predictive ability of the above described models commonly used measures were applied which include:

- mean error (*MAE*)

$$MAE = \frac{1}{n} \cdot \sum_{i=1}^{n} \left| y_{i,obs} - y_{i,pred} \right| \tag{8}$$

- mean percentage error (*MAPE*)

$$MAPE = \frac{1}{n} \cdot \sum_{i=1}^{n} \left| \frac{y_{i,obs} - y_{i,pred}}{y_{i,obs}} \right| \cdot 100\% \tag{9}$$

Where the subscripts: *obs* - refers to measured values, *pred* - refers to calculated values, *n* - refers to number of elements in the set, in our case $n = 4652$.

Calculation results

The results of calculation carried out using statistical methods (classical RF and modified RF + SOM) and MARS method are presented in Table 1. For the WWTP in Cedzyna, being under consideration in this paper, a random number of forest trees varied from 50 to 100 and in the MARS method the amount of the base functions varied in the range of 5-10.

Analyses show that in the case of the forecast value $SE(t)$ based on the temporary inflow of $Q(t)$ and the temperature of the sludge $T(t)$, the best results of settleability simulation obtained using modified random forests method (RF + SOM), for which the error values are: *MAE* = 57.8 cm^3/dm^3 and *MAPE* = 13.9%. On the other hand, the worst performance of $SE(t)$ prediction obtained from the mathematical model developed using the method of MARS. In that case, the error values are equal: *MAE* = 84.8 cm^3/dm^3 and *MAPE* = 18.1%. Slightly worse results of analyzes then for the previous model (Q, T) were obtained when explanatory variables of settleability sludge was the values of $Q(t\text{-}1)$. In this case, the smallest error values were obtained using the modified random forests method (*MAE* = 53.1 cm^3/dm^3 and *MAPE* = 12.5%), and significantly higher in the statistical models developed based on the method MARS (*MAE* = 111.1 cm^3/dm^3 and *MAPE* = 25.4%) and random forests (*MAE* = 113.1 cm^3/dm^3 and *MAPE* = 26.1%).

Furthermore, the simulations show that the flow rate $Q(t\text{-}2)$ has no effect on improving the prediction accuracy of statistical models of settleability. This is confirmed by the evaluated model prediction errors in the methods: MARS (*MAE* = 113.1 cm^3/dm^3 and *MAPE* = 25.6%), RF (*MAE* = 110.5 cm^3/dm^3 and *MAPE* = 25.1%), and (RF + SOM) (*MAE* = 61.7 cm^3/dm^3 and *MAPE* = 14.3%). Much better results than before abutting against the $Q(t\text{-}1)$ and $Q(t\text{-}2)$ were obtained with the temperatures $T(t\text{-}1)$ and $T(t\text{-}2)$.

In such a case, there were used for predicting the settleability calculated values of the prediction error $SE(t)$ for the methods MARS, RF and RF + SOM are respectively equal to: MAE = 86.9 cm^3/dm^3 and $MAPE$ = 18.3%, MAE = 72.2 cm^3/dm^3 and $MAPE$ = 15.9% and MAE = 34.9 cm^3/dm^3 and $MAPE$ = 8.5%. Similar values of prediction errors of $SE(t)$ were obtained when explanatory variables in the model were the values of the temperature $T(t\text{-}1)$.

Table 1

Comparison of the accuracy of forecasts for the developed mathematical models by methods: MARS, random forests (RF) and the modified random forests (RF + SOM)

Variables	MARS				RF				RF+SOM			
	Education		Test		Education		Test		Education		Test	
	MAE	***MAPE***	***MAE***	***MAPE***	***MAE***	***MAPE***	***MAE***	***MAPE***	***MAE***	***MAPE***	***MAE***	***MAPE***
Q,T	104.9	30.8	84.8	18.1	76.5	22.4	69.1	15.5	48.7	12.9	57.8	13.9
$Q(t\text{-}1)$	132.6	39.9	111.1	25.4	1113	33.7	111.3	26.1	47.8	13.1	53.1	12.5
$Q(t\text{-}1)$, $Q(t\text{-}2)$	131.8	39.6	113.1	25.6	113.9	33.9	110.5	25.1	66.2	19.1	61.7	14.3
$T(t\text{-}1)$, $T(t\text{-}2)$	108.7	33.4	86.9	18.3	80.9	24.2	72.2	15.9	31	8.2	34.9	8.5
$T(t\text{-}1)$	108.8	33.5	86.8	18.3	78.7	23.6	71.5	15.9	32.6	9.4	35.8	8.9
$SE(t\text{-}1)$	40.9	12.4	43.1	10.3	40.8	12.3	42.0	10.1	33.5	8.6	35.9	8.9
$SE(t\text{-}1)$, $T(t\text{-}1)$	40.4	12.3	42.4	10.0	41.0	12.0	41.0	9.4	33.8	8.6	31.9	8.8

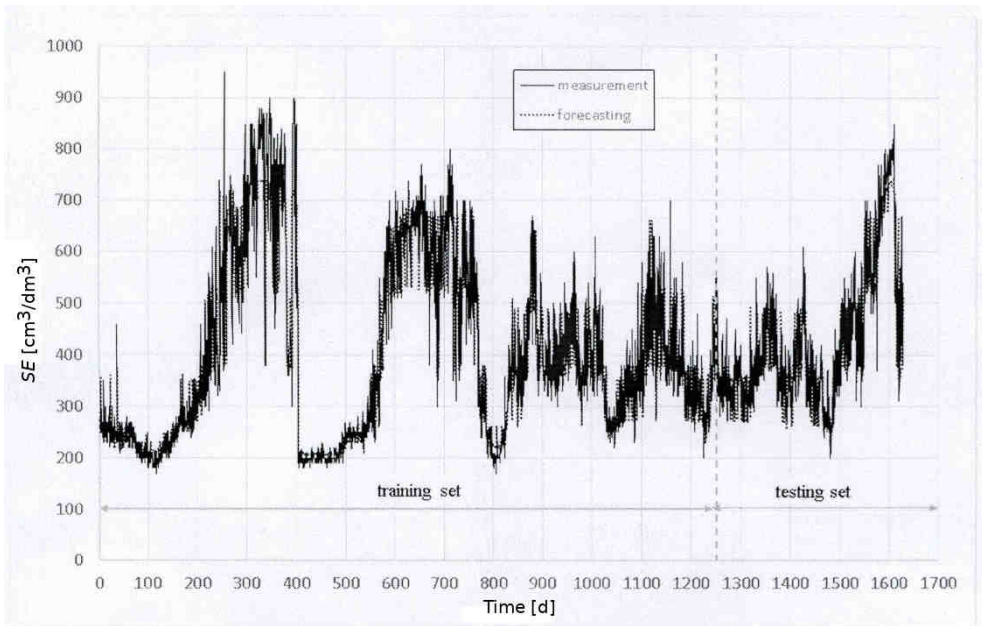

Fig. 1. Comparison of the results of settleability calculations by the modified method of random forests with the measurement ones

Analysis of the data given in Table 1, obtained on the basis of simulations carried out showed that the best ability of predicting the values of $SE(t)$ have models, in which results of the previous measurement of settleability have been taken into account. Forecasting variable $SE(t)$ based on $SE(t\text{-}1)$ error values for absolute and relative statistical models

developed using the method MARS, random forests, and the modified model of RF are respectively equal *MAE* = 43.1 cm^3/dm^3 and *MAPE* = 10.0%, *MAE* = 42.0 cm^3/dm^3 and *MAPE* = 10.1% and *MAE* = 35.9 cm^3/dm^3 and *MAPE* = 8.8%. Furthermore, based on the performed simulation, it was found that the methods RF and MARS inclusion in the model based on the *SE*(*t*) with the additional values of *T*(*t*-1) leads to a slight improvement in the accuracy of predictions settleability values. In order to visualize the results of calculations obtained using the modified random forests method compared them with the measurement ones (Fig. 1).

Conclusions

It was found that the methods MARS and random forests (RF) may be used for mathematical modelling of the settleability process. Carried out calculations showed that introduction of the additional variable of a classification type, based on the self organizing neural networks, into the random forest model leads to a significant improvement in the predictive ability of the such improved RF method. It is confirmed by the values given in Table 1. On the other hand, comparable results of the settleability process forecast were obtained using the methods MARS and RF. Calculations have shown that the best results in the settleability process predicting were obtained when the values of the settleability and temperature specified in the previous measurement were the data input to the model.

References

[1] Giokas DL, Daigger GT, Sperling M, Kim Y, Paraskevas PA. Comparison and evaluation of empirical zone settling velocity parameters based on sludge volume index using a unified settling characteristics database. Water Res. 2006;37(16):3821-3836. DOI: 10.1016/s0043-1354(03)00298-7.

[2] Dellana SA, West D. Predictive modeling for wastewater applications: Linear and nonlinear approaches. Environ Modell Software. 2009;24(1):96-106. DOI: 10.1016/j.envsoft.2008.06.002.

[3] Zhang R, Hu X. Effluent quality prediction of wastewater treatment system based on small-world. Ann J Computers. 2012;7(9):2136-2143. DOI: 10.4304/jcp.7.9.2136-2143.

[4] Lou I, Zhao Y. Sludge bulking prediction using principle component regression and artificial neural network. Mathemat Probl in Eng. 2012;2012(2012): DOI: 10.1155/2012/237693.

[5] Poutiainen H, Niska H, Heinonen-Tanski H, Kolehmainen M. Use of sewer on-line total solids data in wastewater treatment plant modelling. Water Sci Technol. 2010;62(4):743-750. DOI: 10.2166/wst.2010.317.

[6] Verma A, Wei X, Kusiak A. Predicting the total suspended solids in wastewater: A data-mining approach. Engineering Applications of Artificial Intelligence. 2013;26(4):1366-1372. DOI: 10.1016/j.engappai.2012.08.015.

[7] Kusiak A, Wei X. A data-driven model for maximization of methane production in a wastewater treatment plant. Water Sci Technol. 2012;65(6):1116-1122. DOI: 10.2166/wst.2012.953.

[8] Grieu S, Thiéry F, Traoré A, Nguyen TP, Barreau M, Polit M. KSOM and MLP neural networks for on-line estimating the efficiency of an activated sludge proces. Chem Eng J. 2006;1116(1):1-11. DOI: 10.1016/j/cej.2005/10.004.

[9] Andres JD, Lorca P, de Cos Juez FJ, Sánchez-Lasheras F. Bankruptcy forecasting: a hybrid approach using Fuzzy c-means clustering and Multivariate Adaptive Regression Splines (MARS). Expert Systems Appl. 2010;38:1866-1875. DOI: 10.1016/j.eswa.2010.07.117.

[10] Han HG, Qiao JF. Prediction of actived sludge bulking based on a self-organizing RBF neural network. J Process Control. 2012;22(6):1103-1112. DOI: 10.1016/j.jprocont.2012.04.002.

[11] Mwale FD, Adeloye AJ, Rustum R. Application of self-organising maps and multi-layer perceptron-artificial neural networks for streamflow and water level forecasting in data-poor catchments: The case of the Lower Shire floodplain, Malawi. Nordic Hydrol. 2014;45(6):838-854. DOI: 10.2166/nh.2014.168.

[12] Rustum R, Adeloye A. Improved modelling of wastewater treatment primary clarifier using hybrid anns. Int J Computer Sci Artificial Intelligen. 2012;2(4):14-22. DOI: 10.5963/IJCSA/0204002.

[13] Belanche L, Valdes J, Comas J, Roda I, Poch M. Prediction of the bulking phenomenon in wastewater treatment plants. Artificial Intellige in Eng. 2000;14(4);307-317. DOI: 10.1016/S0954-1810(00)00012-1.

[14] Friedman J. Multivariate adaptive regression splines. Annals Statistics. 1991;19:1-141. http://projecteuclid.org/download/pdf_1/euclid.aos/1176347963.
[15] Breiman L. Random forests. J Machine Learning. 2000;45(1):5-32. DOI: 10.1023/A:1010933404324.
[16] Kohonen T. Self-organized formation of topologically correct feature maps. Biol Cybernetics. 1982;43:59-69. DOI: 10.1007/BF00337288.
[17] Han HG, Chen QL, Qiao JF. An efficient self-organizing RBF neural network for water quality prediction. Neural Network. 2011;24(7):717-725. DOI: 10.1016/j.neunet.2011.04.006.
[18] Han HG, Ying L, Guo YN, Qiao JF. A soft computing method to predict sludge volume index based on a recurrent self-organizing neural network. J Appl Soft Computing. 2016;38(C):477-486. DOI: 10.1016/j.asoc.2015.09.051.
[19] Martins AMP, Heijnen JJ, van Loosdrecht MCM. Bulking sludge in biological nutrient removal systems. Biotechnol Bioeng. 2004;86(2):125-135. DOI: 10.1002/bit.20029.