Desalination and Water Treatment
www.deswater.com

doi: 10.5004/dwt.2020.25944

196 (2020) 58–66
August

Prediction of BOD_5 content of the inflow to the treatment plant using different methods of black box – the case study

Alicja Gawdzik[a,*], Jarosław Gawdzik[b], Barbara Gawdzik[c], Andrzej Gawdzik[d], Marcin Rybotycki[e]

[a]*Department of Process Engineering, University of Opole, ul. Dmowskiego 7/9, 45-365 Opole, Poland, Tel. +48774016700; email: gawdzika@yahoo.com (A. Gawdzik)*
[b]*Division of Waste Management, Faculty of Civil and Environmental Engineering, University of Technology, Al. 1000-lecia Państwa Polskiego 7, 25-314 Kielce, Poland, email: jgawdzik@gmail.com (J. Gawdzik)*
[c]*Institute of Chemistry, Faculty of Mathematics and Science, Jan Kochanowski University, ul. Świętokrzyska 15, 25-406 Kielce, Poland, email: b.gawdzik@ujk.edu.pl (B. Gawdzik)*
[d]*Department of Process Engineering, University of Opole, ul. Dmowskiego 7/9, 45-365 Opole, Poland, Tel. +48774016700; email: kip@uni.opole.pl (A. Gawdzik)*
[e]*Department of Process Engineering, University of Opole, ul. Dmowskiego 7/9, 45-365 Opole, Poland, Tel. +48774016700; email: kip@uni.opole.pl (M. Rybotycki)*

Received 24 September 2018; Accepted 2 March 2020

ABSTRACT

The publication presents the possibility of modeling in a 1 d advance of the content of organic compounds in the influent wastewater to the treatment plant, where the content of these compounds is determined by both the biochemical and chemical oxygen demand. To predict the quality of the wastewater at the inflow a set of indicators where used to make measurements on a daily basis. In order to develop statistical models 3 methods where used, namely: multivariate adaptive regression splines (MARS), boosted trees (BT), and genetic programming (GP). The carried-out calculations showed that, to calculate the BOD_5 there can only be used models developed on the basis of the value of daily wastewater flow rate to the wastewater treatment plant with 1- and 2-d lags compared to the predicted value. However, in the forecast model of the COD, better wastewater quality index was obtained when as the explanatory variables were measured with COD values of 1- and 3-d lags to the modeled quantity than the daily flow rate referred to the last two measurements.

Keywords: Multivariate adaptive regression splines; Boosted trees; Genetic programming; Organic compounds; BOD_5; COD; Wastewater treatment plant (WWTP)

1. Introduction

Operation of a sewage treatment plant is a complex process requiring the maintenance of multiple processes at the appropriate level in order to achieve a minimum reduction effect of pollution conditioned by applicable laws. However, due to the stochastic nature of the supply of both the quantity as well as quality of sewage, there exists an unevenness between the daily, weekly, and monthly cycles. In practice, this leads to disturbances in the operation of the object, which can translate into inappropriate decisions made by the technologist responsible for the process of wastewater treatment [1–3]. To avoid this, it is appropriate to do a mathematical modeling of both the quantity and the quality of the wastewater, because it gives the opportunity to prepare the object for the optimal tuning devices purification plant in order to maintain the quality of the effluent treatment plant at the appropriate level [4,5].

The suitable operating parameters of the biological reactor can be determined on the basis of calculations carried out by means of a physical model describing the kinetics of biochemical compounds of nitrogen, carbon, and phosphorus in the various blocks of the reactor [6–8]. In the case of these models, special attention should be paid

* Corresponding author.

to the amount of carbon contained in the influent wastewater, as it determines the processes occurring in the biological reactor. They have a significant impact on the sediment load expressed as the ratio of BOD_5 load to the content of the biomass in the reactor, and thus affecting the sludge age, the amount of removed excess sludge and the operating parameters chambers of the activated sludge [9].

Forecasting the quality of the wastewater in the inflow to the treatment plant is off significant importance from the engineering point of view, because it gives the opportunity to identify abnormal events that may lead to disturbance in the operation of sewage treatment plants. Anticipating this type of events allows facilities to be prepared in advance so that the balance between microorganisms in the activated sludge is not disturbed. In practice, this could lead to a deterioration in the quality of the wastewater at the outlet. The application of mathematical models to forecast the quality of the wastewater eliminates potential problems in operations of the reactor, which could result from the decrease in the load of sewage flowing into the treatment plant and gives the possibility to predict settings in individual wastewater treatment plant (WWTP) devices, which leads to economic benefits.

To describe the quality and quantity of wastewater flowing into municipal sewage treatment plants, there is the method of black box [10–12]. Based on the performed review of the literature it can be said that the authors of the papers [13,14] used it successfully to simulate the chemical oxygen demand (COD), concentration of total suspended solids (TSS), total nitrogen (TN), and phosphorus (TP) in wastewater flowing into the treatment plant type models ARIMA (autoregressive integrated moving average), and artificial neural networks (ANN) [15] to forecast the total concentrations of the suspended solids a number of models where used such as, the method of support vector machine (SVM) [16], forests random (RF) [17], multivariate adaptive regression splines (MARS) and *k*-nearest neighbor (k-NN) to give the slight differences in the values of simulated and measured. Moreover, the analysis done by Minsoo et al. [18] has confirmed the possibility of using the method of *k*-NN for modeling both the quality and quantity of the wastewater. The above-mentioned models calculate influent quality in the supplies carried out solely on the basis of the indicators measured in the past made measurements, and only focuses on the prediction of COD, while not predicting biochemical oxygen demand (BOD), which also plays an important role in the biological processes in the bioreactor. Despite the fact that the above methods were used to simulate sewage quality indicators at the treatment plant inflow, including BOD values, the independent variables included in them and their quantity were not simple to be implemented at the operational stage. Based on the literature data [19], it was found that BOD values were modeled on one side based on the values of the same indicator determined in previous measurements. An alternative solution is an approach where measurements of other sewage quality indicators such as total nitrogen, total phosphorus, suspension are used. It is true that the time to perform their determinations is shorter than BOD, but the cost of the measurement is not low, which generates technical problems at the stage of using the model in technical conditions. Thus, there is a need to look for cheap solutions in which the costs of determining sewage quality indicators will be limited.

In view of the above comments, the publication presents the possibility of modeling in 1 d advance of the content of organic compounds in the influent wastewater to the treatment plant. Where the content of these compounds is the biochemical and COD. To predict the quality of the wastewater at the inflow a set of indicators where used to make measurements on a daily basis. In order to develop statistical models 3 methods of black box where used (MARS, BT, and GP). The analysis used the results of 3 y of measurements of the flow and quality of wastewater that flows to the big sewage treatment plant located in Podkarpackie province.

1.1. Object of Investigations

The object of the research is the sewage treatment plant in Poland built in the 70-ties of the XX century, which was later repeatedly modernized. The designed average daily capacity of the object is $Q_{d\text{avg}}$ = 62,500 m^3/d and $Q_{d\max}$ = 75,000 m^3/d and the population equivalent is equal to 400,000 RLM. The technology of the wastewater treatment involves mechanical–biological wastewater treatments with integrated removal of nutrients, based on a conventional multi-phase activated sludge with denitrification and nitrification ahead of the circulation system. The annual rainfall is 597–857 mm, and the number of days with rainfall is 165–286. The average annual air temperature varies from 7.1°C to 8.6°C, while the number of days with snowfall is 46–94.

The considered sewage treatment plant collects sewage from the entirety of the city, which basically has a distribution sewage system, that consists of sanitary and rainwater sewage systems. In some areas of the city there is no rainwater drainage system. More than 50% of the collector's length are concrete channels; about 65% of the sewer network exceeds the service life by more than 20 y. The level of groundwater in the city strongly depends on the water level in the Wisłok River and on the water levels in its tributaries. In practice, with high groundwater levels (locally approximately 0.3 m below ground level) and a large amount of precipitation water, an increase in the collector fillings is observed.

2. Materials and methods

In this paper, we estimate the quality of wastewater at the inflow to the treatment plant for two cases. In the first of them to forecast biological and COD based on the results of research on indicators of wastewater quality measured in the last conducted measurement. In the second case, the calculation of BOD_5 and COD based on the results of the measured flow rate value. Three statistical models were used for analysis, that is, the method of reinforced trees classified as the so-called black box method and the MARS and GP (genetic programming) methods, where the obtained result has the character of an explicit dependence having the appropriate interpretation. Prior to the analysis the data was standardized by converting minimum–maximum.

Method MARS is one of the many methods used to solve problems of a regression [20–23] and is an extension of the classical input data capture in the developed mathematical

models. Besides the overall recognition of the explanatory variables, as is in the classic regression model, the method MARS ranges of the variability taken into account, predictors are divided into compartments in which the analyzed variables can have a different impact on the phenomenon. The lines of separation are established on the basis of threshold values (t), which means that, depending on whether the variable is below or above the value t, the predictor can be considered to be included in the statistical model with a different weighting or another sign. The separation of the analyzed variables for smaller and larger than the threshold t is based on the function of the base form:

$$h(X) = \alpha_i \cdot \left(\max\left(0, X - t\right)\right) \tag{1}$$

where $h(X)$ is the vector of the basis functions for individual variables (x and y) for which the following relationship is satisfied:

$$x_i - t_i = \begin{cases} x_i - t_i; & \text{for } x_i > t_i \\ 0; & \text{for } x_i \le t_i \end{cases} \tag{2}$$

The regression relationship in the MARS method makes the spline function obtained from the linear combination of the product of the basis functions with corresponding weights stored as follows:

$$f\left(X\right) = \alpha_0 + \sum_{m=1}^{M} \alpha_m \cdot h_m\left(X\right) \tag{3}$$

where $X = [x_1, x_2, ..., x_M]$ is the vector input, α_m is the weight, h_m is the basis functions.

To calculate the parameters of the model, the algorithm that gives the ability to search the space of observation to determine the threshold values (nodes) has been prepared. The algorithm is based on the method of recursive division of the feature space and is composed of the following two stages alternately until the stop criterion is reached representing a generalized value of the error in the evaluation of the cross [24]. In the first stage, there is an increased complexity of the model which is obtained by adding base functions until it reaches the maximum number of functions declared by the user. In the next stage there is an activated elimination procedure (i.e., cutting) that is the least important feature of the base model, the removal of which leads to the smallest decline in predictive ability of the model. The calculation procedure makes it possible to reduce independent variables in the model having little effect on the simulation results [25]. This is important from the point of view of the model's complexity and elimination of variables (sewage quality indicators) that lead to excessive model expansion.

Method BT (boosted trees) is the implementation of the method of stochastic gradient amplification used in classification and regression problems [26]. The main idea of the method is to create a sequence of the decision trees, each of which will be used to determine the residues generated by the former. The calculations have shown that for some problems of estimation and prediction the prediction obtained with the aid of reinforced trees are much closer to the real values than those obtained with individual regression trees.

A special kind of evolutionary algorithm has been proposed in the paper [27], which is called genetic programming (GP) being a further development of a genetic algorithm (GA) used to create programs in the form of so-called "arsive" trees (Fig. 1) solving the stated problem [28]. This model is based on a process of collective learning population dots called individuals. Individuals in GP present a tree of nodes randomly selected from two collections: the leaves of the tree defined from a set of terminal arguments (T) and the other nodes from the set of functions (F). Depending on the function of T, F the individual may be a Boolean expression or a mathematical function. The terminal collection may be considered independent by variables describing the phenomenon and constants, whereas the function set includes basic mathematical operators (+, –, /, sine, cosine, exp, etc.) those can be used at the stage of creating the model. The process of evolution begins with a random selection of n individuals in the population (selection) and in the next step four operations (generations) are repeated N is the times: reproduction, genetic operations, evaluation, and succession until it is fulfilled stopping the criterion of the algorithm. Stochastic reproduction operator involves the selection from the current population n of the parents with the help of which are generated child individuals. The parent's individuals are subjected to genetic operators designed to mix the information contained in them by crossing (Fig. 1a) and mutation (Fig. 1b). Thus, obtained specimens children are assessed, that is, values of optimized quality criteria called matching function are determined. Point mutation changes the function or terminal signs in the selected parts of the tree to another within the same tree, in turn, mutation called sub-tree (Fig. 1) randomly turns the whole sub-tree creating new ones. Crossing subtrees (Fig. 1a) is recognized as the most important provider of genetic programming, as it enables the creation of new trees by replacing the randomly chosen subtrees among existing.

In these analyzes, the criterion of matching data calculation results obtained using mathematical dependence based on the GP to measurement data were of mean absolute error (MAE) and relative (MAPE). To develop the model used to predict the biological and COD we use basic mathematical operators such +, –, /. In addition, it was assumed that the initial number of individuals is n = 200 and the amount of generations is N = 300. The mutation probability was established at P_m = 0.25 and the crossing at P_e = 0.65.

2.1. 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)

$$\text{MAE} = \frac{1}{n} \cdot \sum_{i=1}^{n} \left| y_{i,\text{obs}} - y_{i,\text{pred}} \right| \tag{4}$$

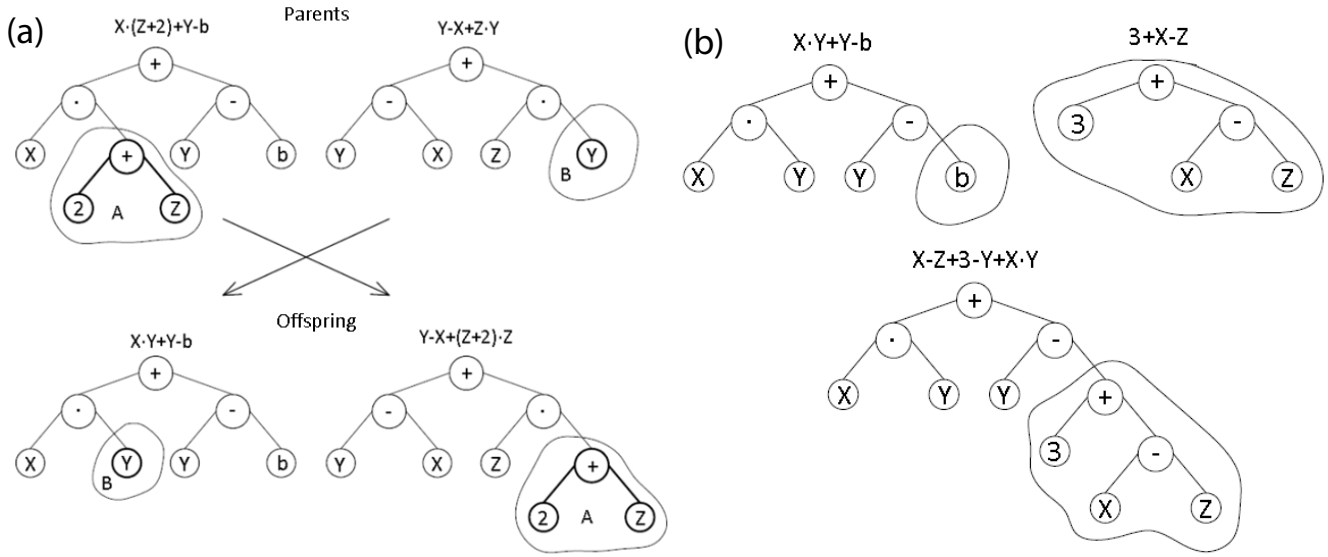

Fig. 1. Flowchart of genetic operators crossing (a) and mutation (b).

- mean percentage error (MAPE)

$$\mathrm{MAPE} = \frac{1}{n} \cdot \sum_{i=1}^{n} \left| \frac{y_{i,\mathrm{obs}} - y_{i,\mathrm{pred}}}{y_{i,\mathrm{obs}}} \right| \cdot 100\% \quad (5)$$

where the subscripts: obs is the refers to the measured values, pred is the refers to calculated values, n is the refers to number of elements in the set.

Measures used in the work to match the results of the calculations to the measurements (MAE, MAPE) are widely used in the discussed topic, which is confirmed by numerous articles in the field of modeling sewage treatment plants [11,12,17,29].

3. Results and discussion

Based on the results of measurements of the quality and amount of wastewater there were derived ranges of variation of the flow rate and the values of biochemical and COD (Table 1). From the data presented in Table 1, it can be concluded that the COD and BOD_5 have varied considerably. Due to the fact that said indicator of waste-water quality are all inputs to the model describing the kinetics of carbon compounds in the biological reactor and changes to a large extent there is a need of modeling their values to predict the performance of individual treatment plant inflow. For this reason, that within the considered working methods only the MARS method parameter estimation algorithm makes it possible to eliminate predictors that have a negligible effect on the dependent variable in the first place, the simulation of selected quality indicators of wastewater flowing into the treatment plant using this method was firstly performed [30].

Table 1
Range of variation of parameters describing the quantity and quality of wastewater flowing into the treatment plant

Parameter	Minimum	Maximum	Mean
COD (mg/dm^3)	159.0	2,510.0	927.1
BOD_5 (mg/dm^3)	38.1	788.0	374.0
Q (m^3/d)	26,973	66,773	38,658

On basis of the variables determined using the method MARS a forecast of values of the COD and BOD_5 was made using other methods. For this purpose, the wastewater quality parameters measured in the last measurement were taken. Next, the simulation of the sewage quality indicators of wastewater influent from sewage municipal WWTP was conducted on the basis of the flow rate measurements.

By following the algorithm for building the model discussed above, the MARS method first established independent variables that have a significant impact on the results of simulation of BOD values. This method is an implementation of the generalization of the technique described in Friedman [24] used to solve both regression and classification problems in which the goal is to find output (dependent) variables based on input (predictive) variables. The relationship between these variables is modeled in the discussed method using a set of coefficients and base functions determined solely from the data. The essence of the MARS method is to divide the entrance space into areas where separate regression or classification functions are determined for each of them. This approach makes this method particularly useful when we have more than two variables at the entrance to the system. The MARS method allows to generate very good models even in cases where the relationships are very complex, non-monotonic, and difficult to any parametric modeling.

Table 2 shows the determined predictors underlying the simulation of the analyzed indicators of the quality of the wastewater, while the values of the fitting parameters (MAE, MAPE) measurement data for the simulation results of the BOD_5 and COD obtained using models MARS and RF after 10-fold cross-validation are given in Tables 3 and 4.

Table 2
Summary of the explanatory variables in the models for prediction of BOD_5 and COD, obtained by MARS

L.p.	Parameter	Variable
1		BOD($t-1$), BOD($t-3$)
2	BOD_5	$Q(t-1)$, $Q(t-2)$, $Q(t-3)$, $Q(t-4)$, $Q(t-5)$
3		BOD($t-1$), BOD($t-3$), BOD($t-5$), $Q(t-1)$
1		COD($t-1$), COD($t-3$)
2	COD	$Q(t-1)$, $Q(t-2)$
3		COD($t-1$), COD($t-3$), $Q(t-1)$

Based on the data presented in Table 2, it can be stated that only the measurement results of the analyzed quality index measured with 1- and 3-d lags against the predicted value are sufficient for the BOD_5 and COD forecasts. Where biochemical and COD calculations are conducted on the flow basis it is sufficient to determine the value of the indicator in question in the previous Q measurements.

On the other hand, the determination of the BOD_5 value uses quality indicators and flow rates that then become necessary data for both the values of the BOD measured with 1- and 3-d lags and the flow rate measured with 1-d lag with respect to the modeled quality. In the case of predicting the COD value on basis of measurements of the quality indicator and the flow rate, it is sufficient to determine the value of the COD with 1- and 3-d lags and delayed flow with 1 d lag with respect to the modeled quantity. Based on the data presented in Table 3, it can be said that the errors MAE and MAPE of the BOD_5 predicted values obtained by MARS based on BOD_5 ($t-i$) and $Q(t-i)$ (where i = 1 or 3) varied slightly within the range MAE = 42.28 – 43.31 mg/dm^3 and MAPE = 12.23% – 12.87% which indicates a similar predictive ability of the obtained statistical models.

Because of the BOD_5 designation period practical applications of BOD ($t-i$) model is not possible. Only the model based on $Q(t-1)$, $Q(t-2)$, $Q(t-3)$, $Q(t-4)$, and $Q(t-5)$ can be implemented to determine the quality indicator described by following equation:

BOD = 3.63063159352875e+002 – 4.95024242975886e–003 × max(0; $Q(t-2)$ – 3.67360000000000e+004) – 2.19696285872868e–003 × max(0; 3.67360000000000e+004 – $Q(t-2)$) + 1.48458140184550e–003 × max(0; $Q(t-3)$ – 5.8222 0000000000e+004) + 1.75660547386152e–003 × max(0; 5.8222 0000000000e+004 – $Q(t-3)$) – 1.55013542803144e–003 × max (0;$Q(t-1)$–3.37380000000000e+004)–6.64142437474370e–003× max(0; 3.37380000000000e+004–$Q(t-1)$)–4.20651110579788e–003 × max(0; $Q(t-5)$ – 3.74890000000000e+004) – 2.38740791982347e–003 × max(0; 3.74890000000000e+004 – $Q(t-5)$) – 6.24453150979316e–003 × max(0; $Q(t-4)$ – 5.3277 0000000000e+004) + 7.32061592009373e–004 × max(0; 5.3277 0000000000e+004 – $Q(t-4)$) – 1.52724461630437e–002 × max (0; $Q(t-4)$ – 4.51010000000000e+004) + 1.87568007342769e–002 × max(0; $Q(t-4)$ – 4.94040000000000e+004) + 4.19569370260533e–003 × max(0; $Q(t-5)$ – 4.0004 0000000000e+004) + 5.62903217638442e–003 × max (0; $Q(t-2)$ – 5.82220000000000e+004) + 3.67944255530525e–003 × max(0; $Q(t-4)$ – 4.01090000000000e+004).

The relationship obtained is empirical in nature, as are the equations obtained using the GP method. Still, their implementation is much more complex than in the genetic programming method. This is of great importance when using the model at the stage of operation of the treatment plant in technical conditions.

For the MARS model used to calculate the COD the highest value of the prediction errors (MAE = 134.30 mg/dm^3 and MAPE = 16.05%) were obtained when the explanatory variables were the values of $Q(t-1)$ and $Q(t-2)$. When the predictors included COD ($t-1$) and COD ($t-3$) the errors were decreased by 16.5% and 18.4% than in the previous case. In addition, based on the data presented in Table 3 it can be stated that the higher values of the prediction errors of the BOD_5 and COD were obtained by the BT method than by the method MARS when the explanatory variables indicators of the quality were the values of Q and BOD_5 or COD. In the framework of these analyzes, the following regression correlations were also reported to determine the quality indicators obtained by the genetic programming method:

$$\mathrm{BOD}(t) = \frac{0.36 + \mathrm{COD}(t-2)}{1.60 + \dfrac{\mathrm{COD}(t-2)}{\mathrm{COD}(t-1)}} \tag{6}$$

for which MAE = 63.53 mg/dm^3 and MAPE = 17.45%.

$$\mathrm{BOD}(t) = \mathrm{BOD}(t-1) \cdot \frac{Q(t-2) + \mathrm{BOD}(t-1)}{Q(t-2)} \tag{7}$$

Table 3
Summary of fitting parameters (MAE, MAPE) models obtained with the MARS and BT for prediction BOD_5 and COD

L.p.	Parameter	MARS				BT			
		Learning		Test		Learning		Test	
		MAE (mg/dm^3)	MAPE (%)	MAE (mg/dm^3)	MAPE (%)	MAE (mg/dm^3)	MAPE (%)	MAE (mg/dm^3)	MAPE (%)
1		39.85	11.84	43.31	12.87	41.56	12.51	43.75	13.45
2	BOD_5	38.90	11.79	42.28	12.23	40.59	12.08	42.73	12.99
3		38.99	11.43	42.38	12.42	43.98	13.19	46.29	14.18
1		105.93	12.48	115.10	13.56	110.21	12.81	116.01	13.77
2	COD	123.57	14.77	134.30	16.05	125.75	14.77	132.37	15.88
3		109.52	13.07	119.04	14.21	122.24	14.54	128.67	15.63

for which MAE = 54.31 mg/dm^3 and MAPE = 15.35%.

$$\mathrm{BOD}(t)=\frac{\mathrm{COD}(t-1)\cdot\left\{\begin{matrix}Q(t-2)+Q(t-4)-\\ \mathrm{COD}(t-3)\end{matrix}\right\}}{\mathrm{COD}(t-4)+\mathrm{COD}(t-3)+\mathrm{COD}(t-2)+Q(t-2)+Q(t-4)+3\cdot Q(t-3)} \tag{8}$$

for which MAE = 66.66 mg/dm^3 and MAPE = 18.39%.

$$\mathrm{COD}(t)=0.98\cdot\mathrm{COD}(t-1)+\frac{\mathrm{COD}(t-1)}{\mathrm{COD}(t-4)} \tag{9}$$

for which MAE = 142.32 mg/dm^3 and MAPE = 15.89%.

The Eqs. (6)–(9) are simply based on regression to determine the quality indicators of wastewater (BOD_5, COD). It is noteworthy that the relationships obtained for the BOD forecast given above are simple empirical relationships that can be used to identify the BOD value at the inflow to the sewage treatment plant. This approach is much simpler than that given in [31], which used the COD, TSS, TN, TP measurements to calculate the BOD value. Despite the fact that the obtained value of matching the results of calculations to measurements was high (R = 0.96), one can doubt whether is it possible to measure at the operational stage with high resolution such a large number of indicators of sewage quality. In addition, the relation given in the paper is much less complex than that proposed in [32] in which BOD used TN, TP, and TSS measurements for calculations. In addition, the advantage of the model given in the work is its simple empirical form, the models for BOD forecasting so far were ANNs, which are not trivial in implementation in the SCADA system at the sewage treatment plant. The given empirical dependencies can be applied by a technologist at a sewage treatment plant without the need to implement complex calculation algorithms.

Although the genetic programming algorithm is not able to clearly determine a posteriori the appropriate solution (in this case there are designed mathematical proofs of the existence of a minimum of the objective function), the presented equations may be used during the operational phase of the object without the need to implement complex numerical algorithms. Figs. 2–5 show a visual comparison of the measurements and simulations of biochemical and COD obtained by use of different statistical models presented in this paper. Based on the data presented in the figures you can find that in majority of cases involved there exists an underestimation of the values of BOD_5 in relation to the measured values. As a result, this may hinder the implementation of the model at the stage of operation of the sewage treatment plant due to differences in the values obtained from calculations and measurements. Nevertheless, analyzing the obtained curves, it can be seen that the model given in the paper allows BOD value estimation – identification of the indicator value above average and below average, which may be important for the technologist in the operation of the sewage treatment plant and the selection of appropriate set values in the biological reactor.

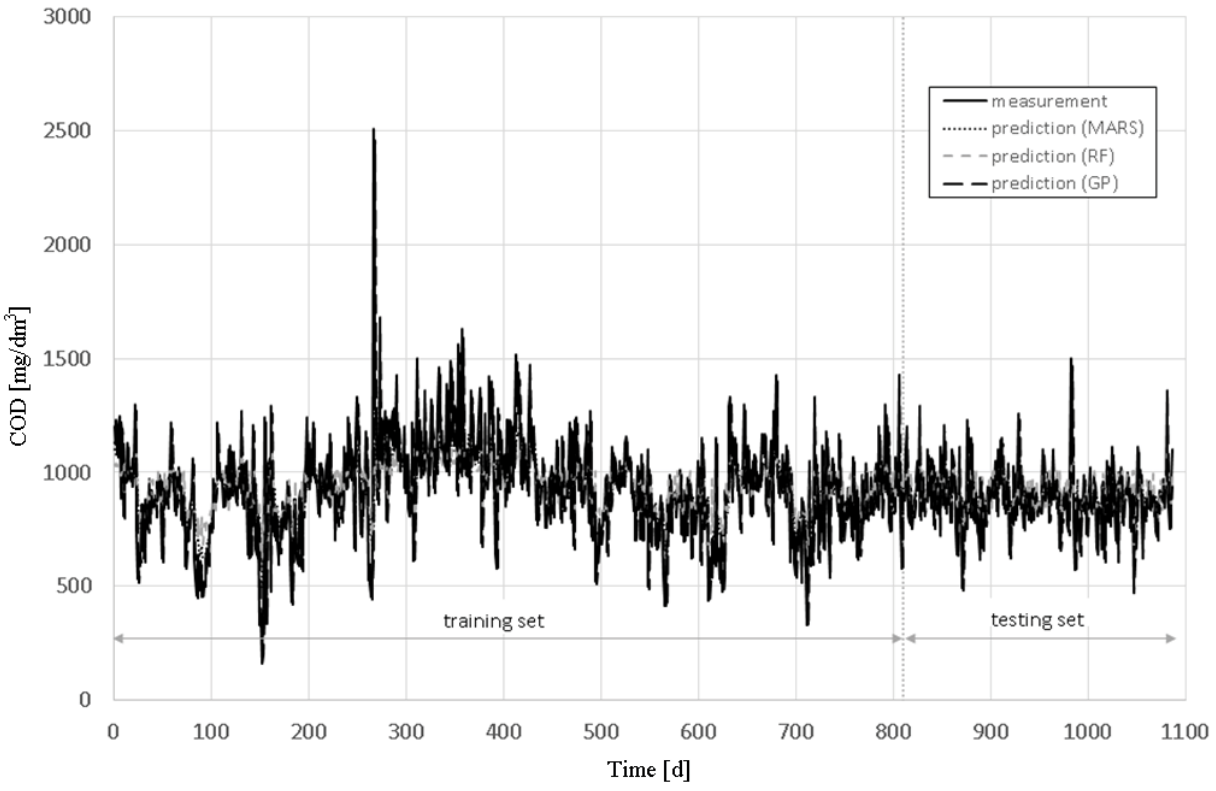

Fig. 2. Comparison of the results of measurements and calculations COD = f(COD(t – i)) methods MARS, BT, and GP.

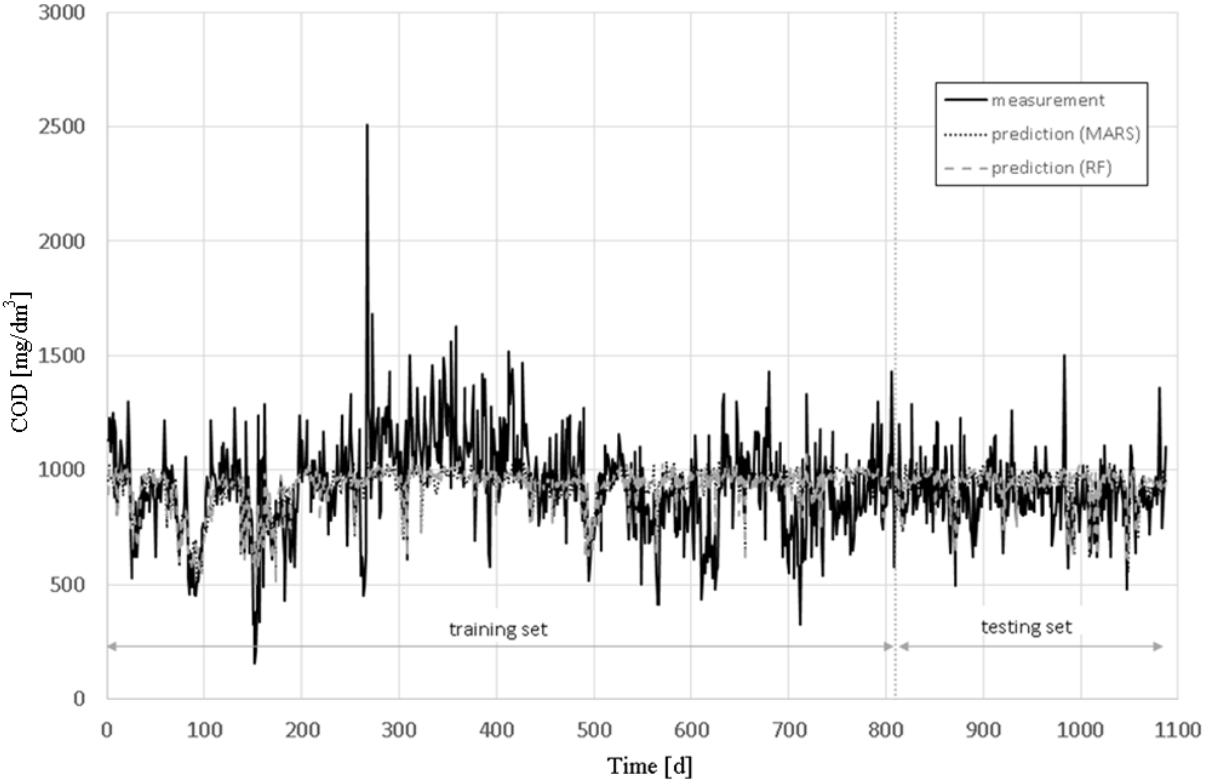

Fig. 3. Comparison of the results of measurements and calculations COD = $f(Q(t-i))$ methods MARS, BT, and GP.

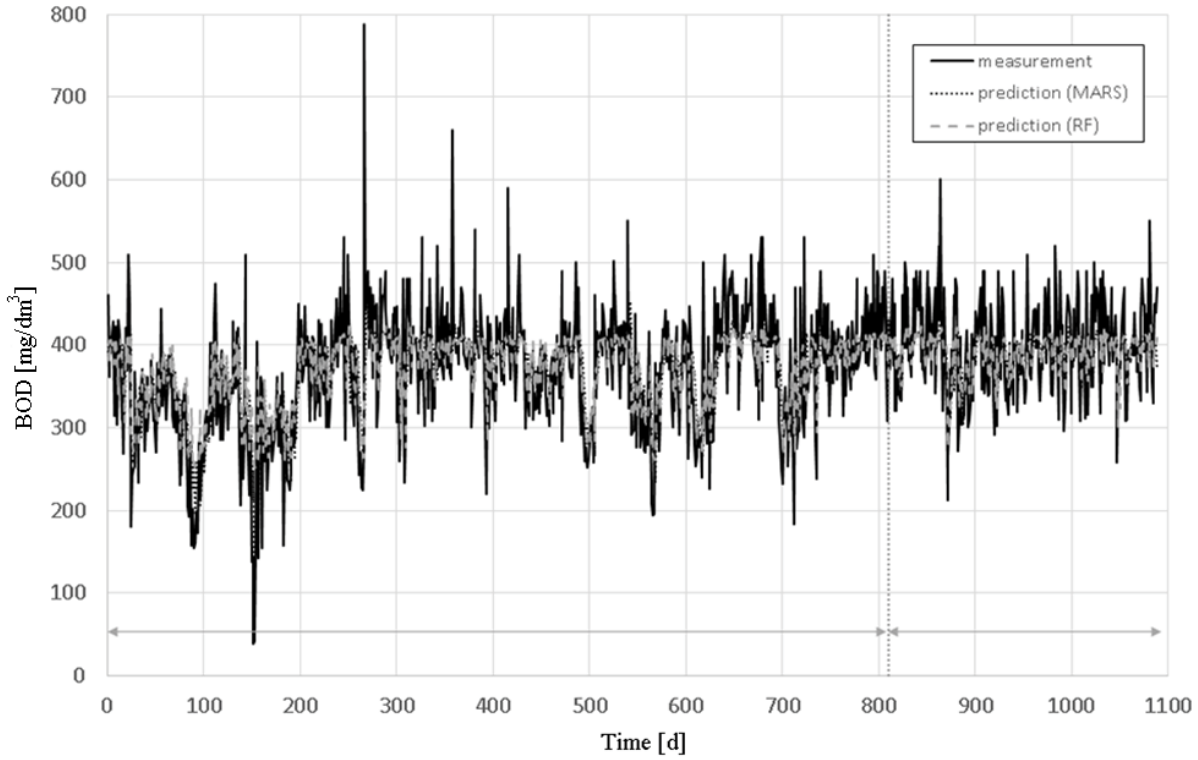

Fig. 4. Comparison of the results of measurements and calculations $BOD_5 = f(BOD(t-i))$ methods MARS and BT.

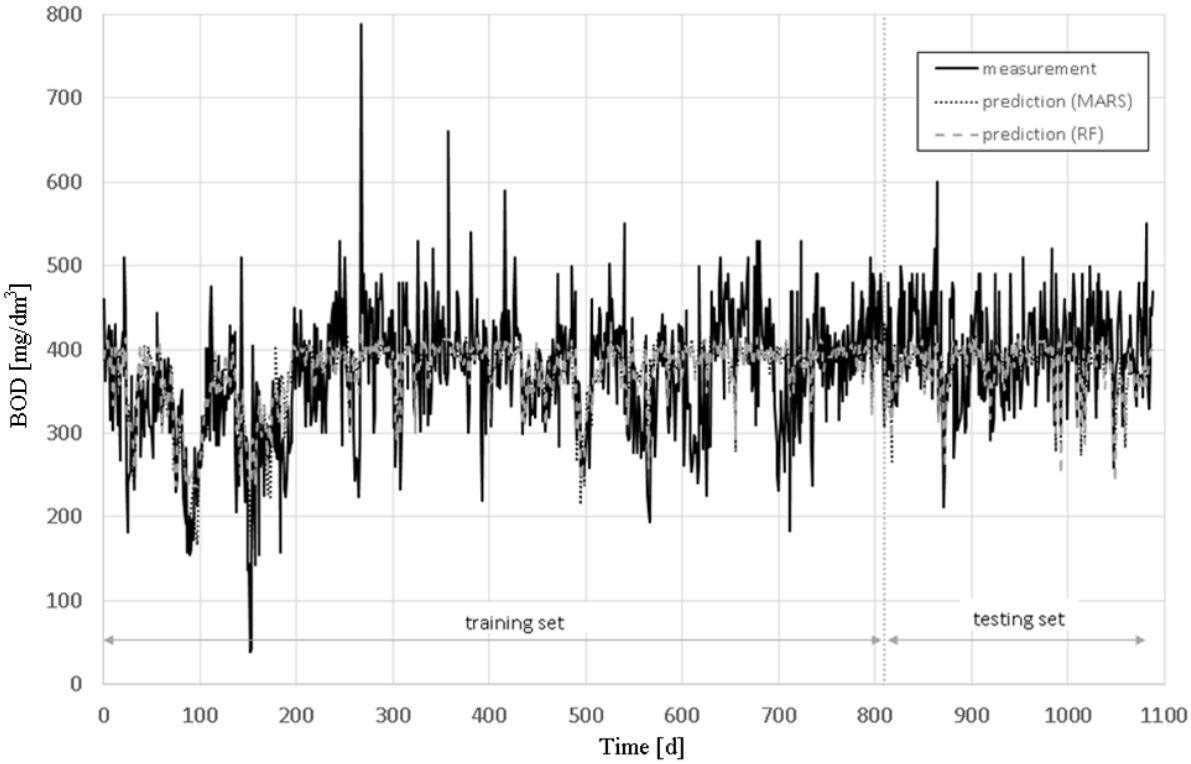

Fig. 5. Comparison of measured and calculated $BOD_5 = f(Q(t - i))$ methods MARS and BT.

4. Summary

Correct operation of the objects in the sewage treatment plant is a complex task because it requires determining the appropriate settings for the individual devices as a condition to obtain the required degree of pollution reduction. To this aim, the values of the indicators of the quality of the wastewater at the inlet for the cleaning should be determined in advance. Specifically, important are the carbon compounds due to fact that they determine the processes taking place in the bioreactor.

This paper presents the possibility of modeling both the biochemical and COD at the inlet to the treatment plant, both based on the rate flow as well as the above-mentioned indicators of wastewater quality. This analysis showed that, in the practical considerations for prediction the BOD_5 cannot be applied to models developed both by MARS, BT, and GP when the explanatory variables are of the values of the BOD. Therefore, to calculate the BOD_5 there can only be used models that were developed based on the value of the daily wastewater flow rate to the WWTP with 1- and 2-d lags rather than the predicted value. In the forecast model of the COD better wastewater quality index was obtained when the explanatory variables were measured with COD values of 1- and 3-d lags to the modeled quantity than the daily flow rate referred to the last two measurements. Within the considered methods (MARS, BT, and GP) similar error values of absolute and relative forecasts indicators of quality (BOD_5, COD) obtained for the models developed based on flow and content of carbon fixed in the last measurement made. Although larger value of errors are obtained by genetic programming (GP) then in the methods of MARS and BT, its depending regression can be used in the operation phase of WWTP by the service object and does not require additional equipment to implement treatment. Moreover, bearing in mind the fact that the developed statistical models for the prediction of COD lead to an underestimation of the value of the indicator, therefore expedient is further analysis in order to improve the predictive ability of the resulting mathematical model.

References

[1] G. Kaczor, T. Bergel, P. Bugajski, Impact of extraneous waters on the proportion of sewage pollution indices regarding its biological treatment, Infrastruct. Ecol. Rural Areas, 4 (2015) 1251–1260.

[2] B. Szeląg, P. Siwicki, Application of selected classification models to the analysis of the settling capacity of the activated sludge – case study, E3S Web Conf., 17 (2017) 1–8, doi: 10.1051/e3sconf/20171700089.

[3] E. Bezak-Mazur, R. Stoińska, B. Szeląg, Ocena wpływu parametrów operacyjnych i występowania bakterii nitkowatych na objętościowy indeks osadu czynnego – studium przypadku, Annu. Environ. Prot., 18 (2016) 487–498 [in Polish].

[4] K.V. Gernaey, M.C.M. Loosdrecht, M. Henze, M. Lind, SB. Jørgensen, Activated sludge wastewater treatment plant modelling and simulation: state of the art, Environ. Modell. Software, 19 (2004) 763–783.

[5] A. Vandekerckhove, W. Moerman, S.W.H. Hulle, Full-scale modelling of a food industry wastewater treatment plant in view of process upgrade, Chem. Eng. J., 135 (2008) 185–194.

[6] J. Cartensen, P. Harremoes, R. Strube, Software sensors based on the grey-box modeling approach, Water Sci. Technol., 33 (1996) 117–126.

[7] S. Al-Asheh, F.S. Mjalli, H.E. Alfadala, Forecasting influent-effluent wastewater treatment plant using time series analysis and artificial neural network techniques, Chem. Prod. Process Model., 2 (2007) 1–23.
[8] H. Poutiainen, H. Niska, H. Heinonen-Tanski, M. Kolehmainen, Use of sewer on-line total solids data in wastewater treatment plant modelling, Water Sci. Technol., 62 (2010) 743–750.
[9] J. Łomotowski, A. Szpindo, Nowoczesne Systemy Oczyszczania Ścieków, Wydawnictwo Arkady, Warszawa, 2002 (in Polish).
[10] H.Z. Abyaneh, Evaluation of multivariate linear regression and artificial neural networks in prediction of water quality parameters, J. Environ. Health Sci. Eng., 12 (2014) 1186–2052.
[11] B. Szeląg, L. Bartkiewicz, J. Studziński, Zastosowanie metod czarnej skrzynki do prognozowania wartości wybranych wskaźników jakości ścieków dopływających do oczyszczalni komunalnej, Environ. Prot., 38 (2016) 39–46 [in Polish].
[12] L. Bartkiewicz, B. Szeląg, J. Studziński, Ocena wpływu zmiennych wejściowych oraz struktury modelu sztucznej sieci neuronowej na prognozowanie dopływu ścieków komunalnych do oczyszczalni, Environ. Prot., 38 (2016) 29–36 (in Polish).
[13] A.G. El-Din, D.W. Smith, Modelling approach for high flow rate in wastewater treatment operation, J. Environ. Eng. Sci., 1 (2002) 275–291.
[14] S.A. Dellana, D. West, Predictive modeling for wastewater applications. Linear and nonlinear approaches, Environ. Modell. Software, 24 (2009) 96–106.
[15] A. Verma, X. Wei, A. Kusiak, Predicting the total suspended solids in wastewater: a data-mining approach, Eng. Appl. Artif. Intell., 26 (2013) 1366–1372.
[16] H. Guo, K. Jeong, J. Lim, J. Jo, Y.M. Kim, P. Jong, J.H. Kim, H.C. Kyung, Prediction of effluent concentration in a wastewater treatment plant using machine learning models, J. Environ. Sci., 32 (2015) 90–101.
[17] A. Kusiak, A. Verma, X. Wei, A data-mining approach to predict influent quality, Environ. Monit. Assess., 185 (2013) 2197–2210.
[18] K. Minsoo, K. Yein, K. Hyosoo, P. Wenhua, K. Changwon, Evaluation of the k-nearest neighbor method for forecasting the influent characteristics of wastewater treatment plant, Front. Environ. Sci. Eng., 10 (2015) 299–310.
[19] B. Szeląg, K. Barbusiński, J. Studziński, Activated sludge process modelling using selected machine learning techniques, Desal. Water Treat., 117 (2018) 78–87.
[20] J.D. Andres, P. Lorca, F.J. de Cos Juez, F. Sánchez-Lasheras, Bankruptcy forecasting: a hybrid approach using Fuzzy c-means clustering and multivariate adaptive regression splines (MARS), Expert Syst. Appl., 38 (2010) 1866–1875.
[21] R.D. De Veaux, D.C. Psichogios, L.H. Ungar, A Comparison of two nonparametric estimation schemes: MARS and neural networks, Comput. Chem. Eng., 17 (1993) 819–837.
[22] G. Gutiérrez, Á.S. Schnabel, J.F.L. Contador, Using and comparing two nonparametric methods (CART and MARS) to model the potential distribution of gullies, Ecol. Modell., 220 (2009) 3630–3637.
[23] B. Szeląg, A. Gawdzik, A. Gawdzik, Application of selected methods of black box for modelling the settleability proces in wastewater treatment plant, Ecol. Chem. Eng. S, 24 (2017) 119–127.
[24] J. Friedman, Multivariate adaptive regression splines, Annu. Stat., 19 (1991) 1–141.
[25] B. Szeląg, K. Barbusiński, J. Studziński, Application of the model of sludge volume index forecasting to assess reliability and improvement of wastewater treatment plant operating conditions, Desal. Water Treat., 140 (2019) 143–154.
[26] J.H. Friedman, Stochastic gradient boosted, Comput. Stat. Data Anal., 38 (2002) 367–378.
[27] J.R. Koza, Genetic Programming: On the Programming of Computers by Natural Selection, MIT Press, Cambridge, MA, 1992.
[28] X. Wei, A. Kusiak, H. Sadat, Prediction of influent flow rate: a data mining approach, J. Energy Eng., 139 (2013) 118–123.
[29] B. Szeląg, J. Studziński, A data mining approach to the prediction of food-to-mass ratio and mixed liquor suspended solids, Pol. J. Environ. Stud., 26 (2016) 2231–2238.
[30] M. Bunge, A general black-box theory, Philos. Sci., 30 (1963) 346–358.
[31] E. Dogan, A. Ates, E.C. Yilmaz, B. Grin, Application of artificial neural network to estimate wastewater treatment plan inlet biochemical oxygen demand, Environ. Prog., 27 (2008) 439–446.
[32] M. Ebrahimi, E.L. Gerber, T.D. Rockaway, Temporal performance assessment of wastewater treatment plants by using multivariate statistical analysis, J. Environ. Manage., 193 (2017) 234–246.