Desalination and Water Treatment
www.deswater.com

doi: 10.5004/dwt.2023.29715

301 (2023) 106–114
July

Sequential algorithm of building the regression-classification model for total nitrogen simulation: application of machine learning

Krzysztof Barbusiński[a], Bartosz Szeląg[b,*], Anita Białek[c], Ewa Łazuka[d], Emilia Popławska[d], Joanna Szulżyk-Cieplak[d], Roman Babko[e], Grzegorz Łagód[d]

[a]*Silesian University of Technology, Konarskiego 18, 44-100 Gliwice, Poland, Tel.: +48 32237-11-94; email: krzysztof.barbusinski@polsl.pl (K. Barbusiński)*
[b]*Warsaw University of Life Sciences, Nowoursynowska 166, 02-787 Warsaw, Poland, Tel.: +48 22 59 310 00; email: bartoszszelag@op.pl (B. Szeląg)*
[c]*Kielce University of Technology, Tysiąclecia Państwa Polskiego 7, 25-314 Kielce, Poland, Tel.: +48 41342-47-35; email: anita_bialek@interia.eu (A. Białek)*
[d]*Lublin University of Technology, Nadbystrzycka 38D, 20-618 Lublin, Poland, Tel.: +48 81538-43-22; emails: e.lazuka@pollub.pl (E. Łazuka), e.poplawska@pollub.pl (E. Popławska), j.szulzyk-cieplak@pollub.pl (J. Szulżyk-Cieplak), g.lagod@pollub.pl (G. Łagód)*
[e]*Schmalhausen Institute of Zoology NAS of Ukraine, 01030 Kyiv, Ukraine, Tel.: +38 044235-10-70; email: rbabko@ukr.net (R. Babko)*

Received 11 November 2022; Accepted 19 June 2023

ABSTRACT

Total nitrogen (TN) concentration is one of important indications of wastewater quality and also a parameter important for wastewater treatment plant performance evaluation. Since the variability of total nitrogen in the effluent from the wastewater treatment plant is the result of the processes taking place in the bioreactor, the processes can be described by mechanistic models, for example, activated sludge models. However, calibration of many parameters is required in such models, and can leads to problems in identifying their proper numerical values. The paper proposes a novel way to deal with this problem by presenting a methodology for building a model for simulating TN, based on sequential structure. In the applied approach, regression models for simulation of TN are first created using Extreme Gradient Boosting (XGBoost), and random forest (RF) methods. In the case of unsatisfactory predictive ability, a division of the dependent variable into a classifier form is made. In the next stage, classification models are created by RF and XGBoost methods and sensitivity analysis is performed by calculating Shapley indices. Two classification models were built that allow for the identification of TN_{eff} variability ranges. The new approach using two models instead of one is preferable because it allows control and optimization of the bioreactor operation.

Keywords: Total nitrogen simulation; Wastewater parameters; Operating and control of WWTPs; Machine learning; Extreme Gradient Boosting (XGBoost); Random forest (RF); Regression and classification models

1. Introduction

Total nitrogen (TN) is one of the important wastewater quality indicators (WQI) for evaluating the performance of wastewater treatment plants (WWTPs). The concentration of total nitrogen in the effluent is regulated by legal acts [1]. The removal of total nitrogen in bioreactors is a complex process that is sensitive to the quality of influent

* Corresponding author.

Presented at the 15th Scientific Conference on Micropollutants in the Human Environment, 14–16 September 2022, Częstochowa, Poland

wastewater and meteorological conditions [2–5]. In many cases quality measurements at the effluent during operation show exceedances of the permissible values of total nitrogen, while the other WQIs (BOD, COD, TSS, TP) are below the maximum ones. To minimalize the above-mentioned problem, bioreactor models are being developed [6–8]. Activated sludge models have been widely used for this purpose. However, due to the complex dynamics of biochemical processes, they are usually over-parameterized, which leads to problems with their calibration [9–11]. Therefore, machine learning models (ML) were also used to model bioreactors [12–15]. In this approach, the problem of model calibration was eliminated. In these models, the basis for creating a simulation tool is measurement data, which was used to identify the model structure and validate its predictive capabilities [16–18]. Regression tree methods, artificial neural networks and their modifications were used to model TN. These models took into account the quality of the effluent at the influent, but due to the limited range of input data, they made it impossible to assess the impact of parameters in the bioreactor on the efficiency of WWTP operation. This issue was rectified by Wang et al. [19] who applied Shapley indices to a sensitivity analysis that determined the effects of quantity, effluent quality at the influent and process parameters on the results of total nitrogen simulations.

Despite numerous works in the field of modeling total nitrogen for a bioreactor using ML, the reliability of simulation results is still not validated by independent simulation tools. This is important from the point of view of using the simulation results obtained as a basis for modifying bioreactor settings to achieve the intended technological effect.

This paper presents a methodology for creating a model for simulating total nitrogen, based on sequential structure. In the adopted approach, regression models for simulation of total nitrogen are first created using Extreme Gradient Boosting (XGBoost), random forest (RF) methods. In the case of unsatisfactory predictive ability, a division of the dependent variable into a classifier form is made. In the next stage, classification models are created by RF and XGBoost methods and sensitivity analysis is performed by calculating Shapley indices. The above-mentioned calculation procedure is presented for the data from the period 2008–2022 from the Sitkówka-Nowiny wastewater treatment plant.

2. Research object

The analysis was based on the Sitkówka-Nowiny wastewater treatment plant, which has a nominal capacity of 72,000 m^3/d, that is, 275,000 PE. The wastewater from the city of Kielce and surrounding municipalities flows into the WWTP. Influent wastewater is treated mechanically on step screens and aerated settling tank. The treated wastewater flows into a bioreactor designed in the BARDENPHO system, where biogenic compounds are removed. The treated wastewater flows into four secondary settling tanks, where the clarification process takes place. Next, the wastewater is discharged to the Bobrza River.

3. Calculation methodology

A calculation methodology that includes two stages of model development was proposed (Fig. 1). In the adopted methodology, on the basis of measurement data collected at the site, the so-called pre-processing of data processing is performed, which is a preliminary statistical analysis of the data.

3.1. Measurement data (step 1)

At the WWTP, as part of continuous monitoring, the quantity and quality of wastewater at the inflow, in the bioreactor and at the WWTP outlet have been measured since 2008. At the inlet and outlet, measurements of wastewater quality indicators are made once a month, including: BOD (biochemical oxygen demand), COD (chemical oxygen demand), TSS (total suspended solids), TN (total nitrogen), NH_4–N (ammonia), NO_3–N (nitrate), NO_2–N (nitrite), TKN (total Kjeldahl nitrogen), TP (total phosphorus). In addition, continuous measurements (1 h resolution) of flow rate are carried out at the inlet. In the bioreactor, continuous measurements (1 h resolution) of the following indicators of wastewater quality, including NH_4–N, NO_3–N, PO_4–P (phosphates), are carried out with analyzers. Operational

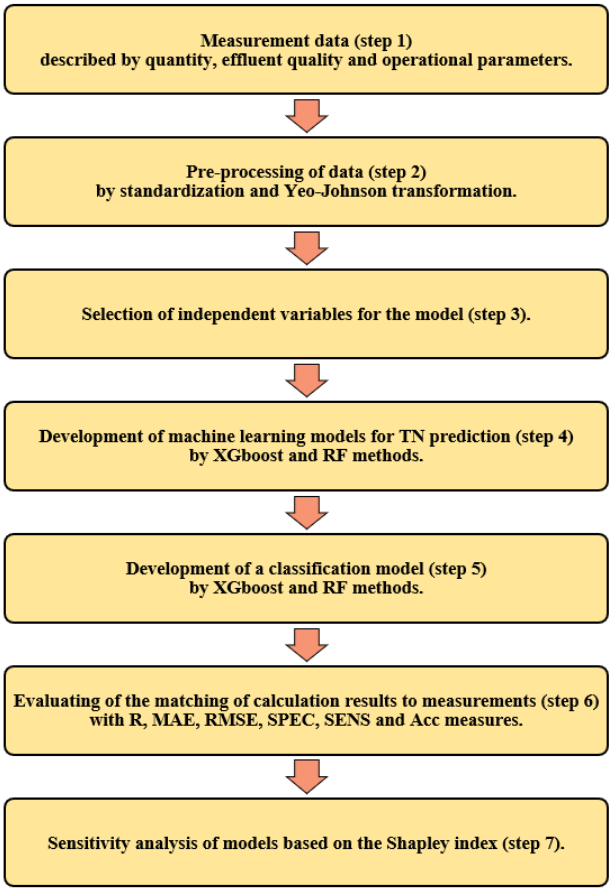

Fig. 1. Computational algorithm used to create a model simulating general nitrogen using machine learning methods.

parameters, that is, recirculation, sludge concentration, oxygen concentration, and sludge temperature are monitored with a resolution of 30 min. During the operation stage, the data on the amount of dosed methanol and PIX coagulant are also collected with a daily resolution. TN was determined using a total organic carbon (TOC) analyzer (TOC-VCSH) coupled with a TN module (TNM-1). The concentrations of COD, inorganic N forms (NH_4–N, NO_3–N) and TP were determined using a Xion 500 spectrophotometer (Dr. Lange GmbH, Berlin, Germany). The analytical procedures, which were adopted by Dr. Lange and SHIMADZU Corporation, followed Standard Methods for the Examination of Water and Wastewater [20]. TSS were measured by the gravimetric methods in accordance with Standard Methods [20].

3.2. Pre-processing of data (step 2)

Prior to the calculations, the data was subjected to preliminary analysis. Data was transformed to facilitate the comparison of different variables, reduce the skewness of the distribution of the variables and the impact of outliers. Data standardization (scaling and centering) was performed, which consists in transforming the data into the form:

$$x_i^* = \frac{x_i - \mu_i}{\sigma_i} \tag{1}$$

where x_i – the values of i-th variable; μ_i – average value of i-th independent variable; σ_i – standard deviation of i-th independent variable.

The data was then transformed using the Yeo-Johnson transform to resemble a normal distribution:

$$\psi(\lambda, x) = \begin{cases} \dfrac{(x+1)^{\lambda}+1}{\lambda} & \text{if } \lambda \neq 0, x \geq 0 \\ \log(x+1) & \text{if } \lambda = 0, x \geq 0 \\ \dfrac{-\left[(-x+1)^{2-\lambda}-1\right]}{2-\lambda} & \text{if } \lambda \neq 2, x < 0 \\ -\log(-x+1) & \text{if } \lambda = 2, x < 0 \end{cases} \tag{2}$$

where x, y – independent and dependent variables, respectively; λ – scaling factor. On the basis of the transformed data using the Yeo-Johnson transformation, the next computational steps were performed (calculation of correlations, determination of the regression and classification model).

3.3. Selection of independent variables (step 3)

The selection of independent variables for the model to simulate total nitrogen was made on the basis of the results of calculating Pearson correlation coefficients and literature data [13].

3.4. Development of models for TN prediction by XGBoost, RF methods (step 4)

In this study, the RF and XGBoost methods were used to model general nitrogen. Each tree is constructed on a random sample of n observations drawn with return from the learning set (bootstrap sample). The independent variables selected in the construction of each tree are also chosen randomly. During the construction of the tree at each node, partitioning is done by drawing without returning m out of p attributes ($m \leq p$). The m parameter is usually determined as follows: $=\sqrt{p}$, as suggested in the literature [21]. Forecasting based on the random forest model involves determining the forecasts for each tree included in the forest and determining the arithmetic mean of these individual forecasts as the forecast of the entire model.

The XGBoost method was developed by Chen and Guestrin [22] based on a regression tree model, which uses an advanced boosting algorithm that considers so-called "regularization" to prevent overfitting. In XGBoost models, each successive tree learns to predict a value of the residual obtained in the previous iterative step. The process of learning the model is based on the minimization of the objective (loss) function enriched with a part causing the regularization of the model.

$$l(x_1, x_2, \ldots x_i) = L(x_1, x_2, \ldots x_i) + \Omega(x_1, x_2, \ldots, x_i) \tag{3}$$

where Ω – a regularization term; $L(x_1, x_2, \ldots, x_i)$ – loss function for regression task described as $\sum_i (y_i - \hat{y}_i)^2$; y_i – observed value of dependent variable; $\hat{y}_i$ – predicted value; x_1, x_2, …, x_i – independent variables.

The selection of optimal hyperparameters for the model was made using the grid search. The machine chose the best *m*try parameter in the range of integers 2 and 14, as well as the parameter specifying the number of *n*tree trees for the random forest. For the XGBoost model, the parameters max_depth, eta, subsample were obtained. The selection of hyperparameters was made based on the R^2, mean absolute error (MAE), root mean square error (RMSE) values. With these and other improvements to the underlying gradient boosting algorithm, XGBoost dominates the machine learning industry this day.

3.5. Development of a classification model (step 5)

In the next stage of calculations, in order to verify the obtained simulation results with XGBoost and RF regression models, classification models were made. This approach can be adopted to control the process of nitrogen removal from wastewater for the case of unsatisfactory simulation results for calculations with ML models.

In order to build a classification model, there is a need to transform TN_{eff} data to binary. Two thresholds for dividing the data into binary forms were adopted in the study:

$$Z_i = \begin{cases} \text{when TN} > \text{TN}_{\text{eff}(0.5p)} \text{ then } Z_i = 1 \text{ and} \\ \quad \text{TN} < \text{TN}_{\text{eff}(0.5p)} \text{ then } Z_i = 0 \\ \text{when TN} > \text{TN}_{\text{eff}(0.75p)} \text{ then } Z_i = 1 \text{ and} \\ \quad \text{TN} < \text{TN}_{\text{eff}(0.75p)} \text{ then } Z_i = 0 \end{cases} \tag{4}$$

where Z_i – variable describing the value of TN in binary form; $\text{TN}_{\text{eff(0.5p)}}$ –50% percentile value of TN determined

from measured data; $TN_{eff(0.75p)}$ – 75% percentile value of TN determined from measurements. The advantage of the adopted approach is the ability to identify total nitrogen in three ranges:

- $TN < TN_{eff(0.5p)}$,
- $TN \hat{I} [TN_{eff(0.5p)}; TN_{eff(0.75p)}]$,
- $TN > TN_{eff(0.75p)}$.

3.6. Evaluating of the matching calculation of calculation results to measurements (step 6)

The following measures of matching simulation results to measurements were used to assess the predictive ability of the regression model:

- coefficient of correlation (*R*):

$$R = \frac{\sum_{i=1}^{N}\left(y_{i,\mathrm{mes}} - \overline{y_{\mathrm{mes}}}\right)\cdot\left(y_{i,\mathrm{sim}} - \overline{y_{\mathrm{sim}}}\right)}{\sqrt{\sum_{i=1}^{N}\left(y_{i,\mathrm{mes}} - \overline{y_{\mathrm{mes}}}\right)^2}\cdot\sqrt{\sum_{i=1}^{N}\left(y_{i,\mathrm{sim}} - \overline{y_{\mathrm{sim}}}\right)^2}} \tag{5}$$

- mean absolute error (MAE):

$$\mathrm{MAE} = \frac{1}{N}\cdot\sum_{i=1}^{N}\left(y_{\mathrm{mes}} - y_{\mathrm{sim}}\right) \tag{6}$$

- root mean square error (RMSE):

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\cdot\sum_{i=1}^{N}\left(y_{\mathrm{mes}} - y_{\mathrm{sim}}\right)^2} \tag{7}$$

where: *N* – number of measurements; y_{mes} – results of total nitrogen measurements; y_{sim} – results of total nitrogen simulations.

The following measures were used to evaluate the matching of the calculation results to the classification model measurements:

- sensitivity (SENS):

$$\mathrm{SENS} = 100\cdot\frac{\mathrm{TPos}}{\mathrm{TPos} + \mathrm{FNeg}} \tag{8}$$

- specificity (SPEC):

$$\mathrm{SPEC} = 100\cdot\frac{\mathrm{TPeg}}{\mathrm{FPos} + \mathrm{TNeg}} \tag{9}$$

- accuracy (Acc):

$$\mathrm{Acc} = 100\cdot\frac{\mathrm{TPos} + \mathrm{TNeg}}{\mathrm{TPos} + \mathrm{TNeg} + \mathrm{FPos} + \mathrm{FNeg}} \tag{10}$$

where: TPos, TNeg, FPos, FNeg – classification results based on RF, XGBoost models (Table S1).

3.7. Sensitivity analysis of models based on the Shapley index (step 7)

The concept was designed to allocate the total profit/reward among players according to the relative importance of their contribution to the final outcome of the game. An importance value is assigned to every feature representing the influence on the model prediction of including this feature. To calculate this effect, $f_{S\cup\{i\}}$ model with the presence of this feature and the f_s model with its omission were analyzed. The SHAPE method requires the calculation of the determined model on all subsets of $S \subseteq M \subseteq \{i\}$, features, where: *M* – the set of all features. The simulation of two Shapley values is evaluated on the basis of the difference $f_{S\cup\{i\}}\left(x_{S\cup\{i\}}\right) - f_S\left(x_S\right)$, where x_s represents the input feature values in the set *S*. Because the effect of withholding a feature is dependent upon other model features, the preceding differences are calculated on all possible differences $f_{S\cup\{i\}}\left(x_{S\cup\{i\}}\right) - f_S\left(x_S\right)$, for all possible subsets $S \subseteq M / \{i\}$. Shapley values were determined as a weighted average of all these differences based on the formula:

$$\Phi_i = \sum_{S\subseteq M/\{i\}} \frac{|S|!\left(|M| - |S| - 1\right)!}{M!}\left(f_{S\cup\{i\}}\left(x_{S\cup\{i\}}\right) - f_S\left(x_S\right)\right) \tag{11}$$

where f_s is the output of the ML model to be explained by means of a set *S* of features, and *M* is the full set of all features. As part of the presented analyses, the *R* – cran software using the Shapley value package was used to calculate SHAP values in RF, XGBoost models and evaluate the impact of individual independent variables.

4. Results

4.1. Correlation matrix

Pearson correlation coefficients were calculated (Table 1) according to the methodology (step 3). The highest value of correlation between total nitrogen (TN_{eff}) and independent variables (quantity, wastewater quality, operational parameters) was found for recirculation (REC; *R* = 0.58) and mixed liquor suspended solid (MLSS; *R* = 0.53). The value of the correlation between TN_{eff} – met was only *R* = 0.08, but this does not exclude the existence of a relationship between variables, due to the fact that it may be non-linear. Among the WQI values on the tributary, the highest correlation value with TN_{eff} was found for NO_2–N (*R* = 0.45), after mechanical treatment the highest correlation was found with COD_m (*R* = 0.23). A high correlation was established between REC – MLSS (*R* = 0.74), but this value is less than the cut-off value of *R* = 0.90 indicating multi-correlation, indicating that both independent variables can be included in the model. A high correlation (*R* = 0.58–0.71) was found between TSS_m and BOD_m, COD_m. The correlation of total nitrogen with quantity, quality of wastewater (BOD, COD, TSS, TN, NH_4–N, NO_3–N) confirms the influence of the amount of organic matter for the course of biochemical processes by microorganisms to remove pollutants in the influent wastewater.

The correlation of total nitrogen with activated sludge temperature confirms that the rate of metabolic processes is a seasonal factor that depends on the season. During the period of reduced temperature, there is a decrease in the dynamics of biochemical processes, resulting in an increase in total nitrogen at the outflow. Operational parameters

Table 1
Pearson correlation coefficients between individual variables

	BOD	COD	TSS	TN	NO_2	NO_3	NH_4	BOD_m	COD_m	TSS_m	T_{as}	MLSS	F/M	REC	meth	TN_{eff}
Q	0.21	0.20	0.19	0.28	0.36	0.06	0.34	0.19	0.32	0.19	0.02	0.28	0.48	0.49	0.08	0.32
BOD	1.00	0.51	0.58	0.33	0.39	0.31	0.33	0.37	0.17	0.12	0.11	0.30	0.01	0.30	0.01	0.20
COD		1.00	0.59	0.40	0.36	0.35	0.30	0.16	0.18	0.07	0.29	0.12	0.09	0.12	0.03	0.12
TSS			1.00	0.38	0.34	0.35	0.35	0.19	0.12	0.17	0.22	0.15	0.09	0.24	0.06	0.25
TN				1.00	0.37	0.24	0.71	0.13	0.14	0.05	0.34	0.02	0.03	0.04	0.05	0.03
NO2					1.00	0.22	0.42	0.04	0.06	0.03	0.36	0.38	0.41	0.45	0.14	0.45
NO3						1.00	0.14	0.10	0.06	0.03	0.33	0.02	0.00	0.13	0.08	0.04
NH4							1.00	0.10	0.13	0.02	0.35	0.17	0.14	0.16	0.17	0.13
BOD_m								1.00	0.72	0.71	0.18	0.31	0.40	0.22	0.24	0.15
COD_m									1.00	0.70	0.11	0.30	0.19	0.23	0.23	0.26
TSS_m										1.00	0.18	0.31	0.19	0.24	0.29	0.22
T_{as}											1.00	0.16	0.03	0.25	0.07	0.15
MLSS												1.00	0.49	0.72	0.13	0.53
F/M													1.00	0.54	0.04	0.41
REC														1.00	0.04	0.58
meth															1.00	0.08
TN_{eff}																1.00

(MLSS, REC, F/M) determine the pollutant load of the bioreactor, retention time and dynamics of multiplication, as well as death of microorganisms in activated sludge [7]. With the above-mentioned considerations in mind, the total nitrogen model can be written as a general relationship:

$$TN_{eff} = f\begin{pmatrix} Q, BOD, COD, TSS, NH_4 - N, \\ NO_3 - N, T_{as}, MLSS, REC, met \end{pmatrix} \quad (12)$$

where Q – flow rate (m^3/d); BOD – biochemical oxygen demand (mg/L); COD – chemical oxygen demand (mg/L); TSS – total suspended solids (mg/L); NH_4–N – ammonium (mg/L); NO_3–N – nitrate (mg/L); T_{as} – temperature in activated sludge chambers (°C); MLSS – mixed liquor suspended solids (kg/m^3); REC – recirculation (%); met-daily dose of methanol (m^3/d).

4.2. Development of models for TN prediction by XGBoost, RF methods

The dataset contained 144 observations. The data was divided into a teaching set and a testing set. The testing set accounted for 30% of the observations of the original dataset. The dependent variable in the regression models was TN. The independent variables are listed in Section 4.1 – Correlation matrix.

The performed calculations showed that the best results of TN_{eff} calculations using the RF method were obtained for 200 trees. For the XGBoost model, it was found that the greatest correspondence of simulation results to measurements for depth tree equal to 10. The simulations carried out showed that the RF model only predicts the average values of TN_{eff} with satisfactory accuracy, at the same time the minimum and maximum values are underestimated (Figs. 2 and 3b). The XGBoost model showed an improvement in the predictive ability of TN_{eff} compared to the RF model (Fig. 2). It was shown that the minimum and maximum values of TN_{eff} were modeled with adequate accuracy (Fig. 2 and 3a).

4.3. Evaluating of the matching calculation of calculation results to measurements (R^2, MAE, RMSE)

On the basis of the simulation results, measures of matching of the calculation results to the TN_{eff} simulation were determined by determining the values of R, MAE, RMSE for the teaching and testing set, respectively (Table 2).

On the basis of the calculations performed, it was found that the RF model for the teaching and testing sets has similar predictive capabilities. This is confirmed by the determined values of R, MAE, and RMSE. For the XGBoost model, a very good fit was found for the learning set, as confirmed by R = 0.96, MAE = 0.44 mg/L and RMSE = 0.53 mg/L. The results of XGBoost model calculations for the learner and test set may indicate overfitting of the model and limited generalization capabilities.

For the XGBoost and RF models, importance values (Imp) were determined for each independent variable, which are given in Table 3. The calculations performed with the XGBoost model showed that the values of determined importance ranged from 0.59–1.00, while with the RF model they were equal to 0.30–1.00. The simulations performed with the XGBoost model showed that MLSS, TSS, Q have the greatest influence on total nitrogen in the outflow, and NO_3–N has the least influence. For the RF model, REC and F/M were found to have a key effect on TN_{eff}.

The RF model calculations performed showed that among the independent variables considered, T_{as} has the least influence. A significant effect of the external carbon source on the total nitrogen content of the outflow was also found in the XGBoost and RF models.

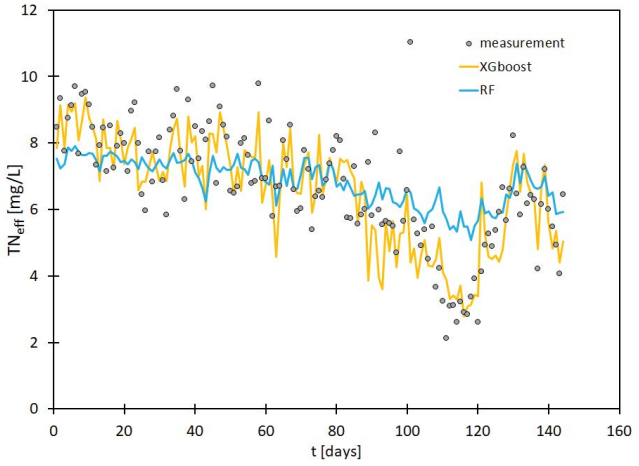

Fig. 2. Comparison of TN_{eff} measurement results for simulation with RF and XGBoost methods.

Table 2
Comparison of fit measures for the determined models (XGBoost, RF) for the teaching and testing set

	Teaching			Testing		
Method	R	MAE	RMSE	R	MAE	RMSE
XGBoost	0.96	0.44	0.53	0.51	1.30	1.72
RF	0.75	1.10	1.38	0.76	1.00	0.78

4.4. Classification models (XGBoost, RF)

Considering the limited predictive ability, as indicated by the R, MAE, RMSE values for the test set, classification

Table 3
Validity values for individual independent variables in the XGBoost and RF models

XGBoost		RF	
Variables	Imp	Variables	Imp
MLSS	1.00	REC	1.00
TSS	0.89	F/M	0.85
Q	0.82	NO_2–N	0.73
F/M	0.81	MLSS	0.72
TN	0.81	Methanol	0.51
REC	0.79	Q	0.48
BOD	0.76	TN	0.47
NO_2–N	0.76	TSS	0.46
NH_4–N	0.70	BOD	0.40
COD	0.69	NH_4–N	0.39
T_{as}	0.65	NO_3–N	0.35
Methanol	0.59	COD	0.34
NO_3–N	0.47	T_{as}	0.30

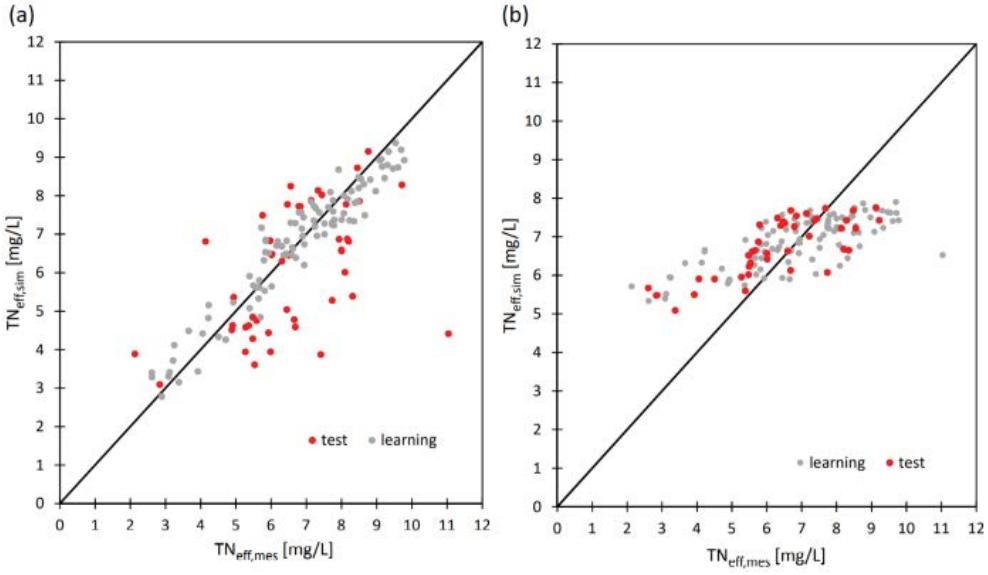

Fig. 3. Comparison of TN_{eff} measurements and calculations for the learning and test set by methods: (a) XGBoost and (b) RF.

models were performed using the XGBoost and RF methods. A 50% percentile value of 6.59 mg/L and a 75% percentile value of 8.17 mg/L were used to divide the TN_{eff} values. Thus, zero-one variables were partitioned, which formed the basis for developing computational models. The determined values of the measures of matching of the simulation results to the calculations (SPEC, SENS, Acc) are given in Table 4. Meanwhile, the determined values of the importance (Imp) of the individual independent variables for the XGBoost and RF models are given in Table 5. On the basis of the data in Table 4, it was concluded that the obtained classification models have satisfactory predictive capabilities and can be used to identify TN_{eff} in the ranges TN < 6.59 mg/L, TN = 6.59–8.17 mg/L, TN > 8.17 mg/L.

The calculations performed showed that in the XGBoost model, the validity values for identifying TN_{eff} corresponding to the 50% percentile, are 0.26–1.00 and for predicting TN_{eff} corresponding to the 75% percentile, they change in the range of 0.76–1.00. In the RF method, the validities were found to be 0.40–1.00 and 0.49–1.00 in the models for classifying TN_{eff} (50%, 75% percentile). It was found that in the XGBoost model, MLSS, REC, NO_2–N have the greatest influence for identifying the 50% percentile of TN_{eff} and the least influence is the COD. It was found that in the XGBoost model, T_{as}, F/M and TN have the greatest influence for identifying the 75% percentile of TN_{eff}, and the least influence corresponds to NO_3–N, NO_2–N. It was found that TN_{eff} is also strongly influenced by Q, TSS, MLSS, and the amount of methanol dosed for which Imp values > 0.90 were obtained. In the RF model, REC, F/M and MLSS have the greatest influence for identifying the 50% percentile of TN_{eff} and NO_3–N has the least influence. In the RF model, MLSS, REC (Imp > 0.90) have a key influence for identifying the 75% percentile of TN_{eff}, and the least effect corresponds to the amount of dosed methanol.

Table 4
Comparison of the values of the measures of matching of measurements to calculation results of classification models obtained with the XGBoost and RF methods

Indices	XGBoost		RF	
	50% percentile	75% percentile	50% percentile	75% percentile
SENS	0.75	0.81	0.80	0.82
SPEC	0.78	0.76	0.78	0.80
Acc	0.77	0.78	0.79	0.81

5. Discussion

Based on the data in Table 6, it can be stated that the proposed methodology of the model building, compared to those developed so far, includes two stages, creating a regression model and building a classification model.

The advantage of the adopted approach compared to the others is the fact that in the case of unsatisfactory predictive abilities of the designated regression models, classification models are built. The application of a single classification model has limited possibilities of its use in the control and optimization of bioreactor operation, but based on two classification models (which was adopted in the paper), it allows for the identification of TN_{eff} variability ranges and thus the correction of the bioreactor settings, if necessary. The models proposed in this paper take into account both the quantity and quality of wastewater, as well as the operational parameters of the bioreactor; a similar approach was adopted in the study by Szeląg et al. [29], Hvala and Kocijan et al. [30], Lee et al. [26], Luo et al. [25]. The proposed solution is important in terms of the possibility of controlling the settings of the bioreactor, which can be used at the stage of the WWTP operation. It should be noted that the model developed in this study takes into account the

Table 5
Validity validities for individual independent variables for the classification models determined by the XGBoost and RF methods

XGBoost				RF			
50% percentile		75% percentile		50% percentile		75% percentile	
Variables	Imp	Variables	Imp	Variables	Imp	Variables	Imp
MLSS	1.00	T_{as}	1.00	REC	1.00	MLSS	1.00
REC	0.90	F/M	0.97	F/M	0.81	REC	0.93
NO_2–N	0.85	TN	0.97	MLSS	0.73	T_{as}	0.87
F/M	0.61	Q	0.96	NO_2–N	0.69	F/M	0.73
TSS	0.44	TSS	0.95	TN	0.58	TSS	0.70
BOD	0.38	MLSS	0.94	BOD	0.47	BOD	0.68
TN	0.34	Methanol	0.93	NH_4–N	0.46	TN	0.64
Q	0.34	COD	0.89	COD	0.46	Q	0.59
T_{as}	0.30	REC	0.89	Methanol	0.45	NH_4–N	0.53
NH_4–N	0.29	BOD	0.81	T_{as}	0.41	NO_3–N	0.49
NO_3–N	0.29	NH_4–N	0.79	TSS	0.40	NO_2–N	0.49
Methanol	0.26	NO_3–N	0.76	Q	0.40	COD	0.49
COD	0.16	NO_2–N	0.76	NO_3–N	0.33	Methanol	0.48

Table 6
Comparison of the obtained model for the TN_{eff} simulation with those determined by other researchers

Study	WWTP	Input	Method	Fitting
Ráduly et al. [23]	Full-scale	*Q*, *T*, TSS, Xi	MLP	na
Lee et al. [24]	Full-scale	*Q*, COD, NH_4–N, TN, OP, DO	PLS + MLP	$R = 0.92$
Luo et al. [25]	Full-scale	TN, OP, DO, NH_4–N	MLP + FR	$R = 0.938$
Lee et al. [26]	Full-scale	*Q*, *T*, pH, COD, TN, TP, MLSS, SVI, DO	PLS + MLP	$R = 0.992$
Guo et al. [27]	Full-scale		SVM, MLP	$R = 0.68$
Yaqub et al. [28]	Full-scale		LSTM	MSE = 0.015
Szeląg et al. [29]	Full-scale	*Q*, *T*, DO, TN, MLSS, BOD, REC, WAS	CNN, SVM, BT	$R = 0.95$ RMSE = 0.82 (CNN) $R = 0.91$ RMSE = 0.71 (SVM) $R = 0.82$ RMSE = 1.52 (BT)
Hvala and Kocijan [30]	Full-scale		Hybrid (MC + ML)	$R = 0.90$
Bagherzadeh et al. [31]	Full-scale	NH_4–N, COD, BOD, DO	MLP, RF, BT	$R = 0.76$ (BT)
This study	Full-scale	*Q*, BOD, COD, TSS, NH_4–N, NO_3–N, T_{as}, MLSS, REC, met	XGBoost, RF	$R = 0.51$ RMSE = 1.72 (XGBoost) $R = 0.76$ RMSE = 0.78 (RF) Acc = 0.77–0.78 (XGBoost) Acc = 0.79–0.81 (RF)

amount of methanol (external carbon source), which allows the process to be controlled with a wide range of variability of the quality of wastewater at the inflow to the treatment plant, even with high dilution of wastewater and unfavorable C/N, C/P values. The model provided by Bagherzadeh et al. [31] has limited applicability in the operational phase as it only considers the quality of the wastewater. This is the factors that make it impossible to control and optimize the bioreactor operation, but only to identify its operation in a continuous time. The performed calculations showed that the data collected in the WWTP on the basis of monitoring (measurement once a month) may be useful for the analysis of the operation of the treatment plant. Studies by other authors [23,25,30] have shown that the models developed on the basis of data collected by means of continuous (on-line) monitoring are more accurate. The data collected continuously in the form of time series (constant resolution) deliver a lot of information, including the identification of the assessment of the impact of the object inertia on the changing quantity, quality of wastewater on the inflow and meteorological conditions. It should be remembered that these data also provide valuable information about changes taking place in the activated sludge, which only after a certain period of time may manifest themselves in the wastewater treatment plant and deterioration of operating conditions.

6. Summary and conclusion

Paper presents a methodology for building a model for simulating total nitrogen, based on sequential structure. In the applied approach, regression models for simulation of total nitrogen are first created using XGBoost and random forest methods. In the case of unsatisfactory predictive ability, a division of the dependent variable into a classifier form is made. In the next stage, classification models are created by random forest and XGBoost methods and sensitivity analysis is performed by calculating Shapley indices. The application of a single classification model has limited possibilities of its use in the control and optimization of bioreactor operation, but based on two classification models (adopted in the paper) allows for the identification of TN_{eff} variability ranges and thus the correction of the bioreactor settings, if necessary.

The models proposed in this paper take into account both the quantity and quality of wastewater, as well as the operational parameters of the bioreactor. Mentioned models covers also the amount of external carbon source, which allows to evaluate the process with a wide range of variability of the wastewater quality at the inflow to the wastewater treatment plant, for example, in case of high dilution of wastewater or unfavorable C/N, C/P values.

Acknowledgements

Publication was partially supported under the rector's pro-quality grant: Silesian University of Technology, Poland, grant number: 08/040/RGJ22/0164 and partially by the Polish Ministry of Education and Science within the grant FD-20/IS-6/999.

References

[1] Ministry of Maritime Economy and Inland Navigation, Regulation of the Minister of Maritime Economy and Inland Navigation from July 2019 on Substances Particularly Harmful to the Aquatic Environment and the Conditions to be Met When Discharging Sewage Into Waters or Ground, as Well as When Discharging Rainwater or Meltwater Into Waters or Into Devices Water, Official Gazette of the Republic of Poland, Poland, 2019.
[2] F. Hernández-del-Olmo, E. Gaudioso, R. Dormido, N. Duro, Energy and environmental efficiency for the N-ammonia removal process in wastewater treatment plants by means of reinforcement learning, Energies (Basel), 9 (2016) 755, doi: 10.3390/en9090755.

[3] J.-J. Zhu, L. Kang, P.R. Anderson, Predicting influent biochemical oxygen demand: balancing energy demand and risk management, Water Res., 128 (2018) 304–313.
[4] M.-J. Mehrani, J. Drewnowski, M. Majewska, G. Lagód, S. Kumari, F. Bux, B. Szęląg, Assessment of wastewater quality indicators for wastewater treatment influent using an advanced logistic regression model, Desal. Water Treat., 232 (2021) 421–432.
[5] B. Szeląg, L. Bartkiewicz, J. Studziński, K. Barbusiński, Evaluation of the impact of explanatory variables on the accuracy of prediction of daily inflow to the sewage treatment plant by selected models nonlinear, Arch. Environ. Prot., 43 (2017) 74–81.
[6] M. Henze, W. Gujer, T. Mino, M. van Loosedrecht, Activated Sludge Models ASM1, ASM2, ASM2D and ASM3, Water Intelligence Online, IAWPRC Scientific and Technical Reports No. 9, IAWPRC Publisher: IWA Publishing, ISBN: 9781780402369, 2000, doi: 10.2166/9781780402369.
[7] J. Drewnowski, J. Mąkinia, A. Szaja, G. Łagód, Ł. Kopeć, J.A. Aguilar, Comparative study of balancing SRT by using modified ASM2d in control and operation strategy at full-scale WWTP, Water (Basel), 11 (2019) 485, doi: 10.3390/w11030485.
[8] H. Hauduc, I. Takács, S. Smith, A. Szabo, S. Murthy, G.T. Daigger, M. Spérandio, A dynamic physicochemical model for chemical phosphorus removal, Water Res., 73 (2015) 157–170.
[9] B. Petersen, P.A. Vanrolleghem, K. Gernaey, M. Henze, Evaluation of an ASM1 model calibration procedure on a municipal–industrial wastewater treatment plant, J. Hydroinf., 4 (2002) 15–38.
[10] G. Mannina, A. Cosenza, P.A. Vanrolleghem, G. Viviani, A practical protocol for calibration of nutrient removal wastewater treatment models, J. Hydroinf., 13 (2011) 575–595.
[11] R. Vitanza, I. Colussi, A. Cortesi, V. Gallo, Implementing a respirometry-based model into BioWin software to simulate wastewater treatment plant operations, J. Water Process Eng., 9 (2015) 267–275.
[12] H. Haimi, M. Mulas, F. Corona, R. Vahala, Data-derived soft-sensors for biological wastewater treatment plants: an overview, Environ. Modell. Software, 47 (2013) 88–107.
[13] J. Fernandez de Canete, P. Del Saz-Orozco, R. Baratti, M. Mulas, A. Ruano, A. Garcia-Cerezo, Soft-sensing estimation of plant effluent concentrations in a biological wastewater treatment plant using an optimal neural network, Expert Syst. Appl., 63 (2016) 8–19.
[14] B. Szeląg, J. Drewnowski, G. Łagód, D. Majerek, E. Dacewicz, F. Fatone, Soft sensor application in identification of the activated sludge bulking considering the technological and economical aspects of smart systems functioning, Sensors, 20 (2020) 1941, doi: 10.3390/s20071941.
[15] J. Fernandez de Canete, P. del Saz-Orozco, J. Gómez-de-Gabriel, R. Baratti, A. Ruano, I. Rivas-Blanco, Control and soft sensing strategies for a wastewater treatment plant using a neuro-genetic approach, Comput. Chem. Eng., 144 (2021) 107146, doi: 10.1016/j.compchemeng.2020.107146.
[16] T.Y. Pai, P.Y. Yang, S.C. Wang, M.H. Lo, C.F. Chiang, J.L. Kuo, H.H. Chu, H.C. Su, L.F. Yu, H.C. Hu, Y.H. Chang, Predicting effluent from the wastewater treatment plant of industrial park based on fuzzy network and influent quality, Appl. Math. Modell., 35 (2011) 3674–3684.
[17] H. Guo, K. Jeong, J. Lim, J. Jo, Y.M. Kim, J. Park, J.H. Kim, K.H. Cho, Prediction of effluent concentration in a wastewater treatment plant using machine learning models, J. Environ. Sci., 32 (2015) 90–101.
[18] 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.
[19] Y. Zhang, C. Li, H. Duan, K. Yan, J. Wang, W. Wang, Deep learning based data-driven model for detecting time-delay water quality indicators of wastewater treatment plant influent, Chem. Eng. J., 467 (2023) 143483, doi: 10.1016/j.cej.2023.143483.
[20] American Public Health Association, Standard Methods for the Examination of Water and Wastewater, 21st ed., Washington D.C., 2005.
[21] T. Hastie, R. Tibshirani, J. Friedman, Random Forests, In: The Elements of Statistical Learning, Springer Series in Statistics, Springer, New York, NY, 2009. Available at: https://doi.org/10.1007/978-0-387-84858-7_15
[22] T. Chen, C. Guestrin, XGBoost: A Scalable Tree Boosting System, KDD'16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794. Available at: https://doi.org/10.1145/2939672.2939785
[23] B. Ráduly, K.V. Gernaey, A.G. Capodaglio, P.S. Mikkelsen, M. Henze, Artificial neural networks for rapid WWTP performance evaluation: methodology and case study, Environ. Modell. Software, 22 (2007) 1208–1216.
[24] D.S. Lee, M.W. Lee, S.H. Woo, Y.J. Kim, J.M. Park, Nonlinear dynamic partial least squares modeling of a full-scale biological wastewater treatment plant, Process Biochem., 41 (2006) 2050–2057.
[25] F. Luo, R.H. Yu, Y.G. Xu, Y. Li, Effluent Quality Prediction of Wastewater Treatment Plant Based on Fuzzy-Rough Sets and Artificial Neural Networks, 6th International Conference on Fuzzy Systems and Knowledge Discovery, FSKD 2009, 2009, pp. 47–51. Available at: https://doi.org/10.1109/FSKD.2009.494
[26] H.W. Lee, M.W. Lee, J.M. Park, Multi-scale extension of PLS algorithm for advanced on-line process monitoring, Chemom. Intell. Lab. Syst., 98 (2009) 201–212.
[27] H. Guo, K. Jeong, J. Lim, J. Jo, Y.M. Kim, J. pyo Park, J.H. Kim, K.H. Cho, Prediction of effluent concentration in a wastewater treatment plant using machine learning models, J. Environ. Sci. (China), 32 (2015) 90–101.
[28] M. Yaqub, H. Asif, S. Kim, W. Lee, Modeling of a full-scale sewage treatment plant to predict the nutrient removal efficiency using a long short-term memory (LSTM) neural network, J. Water Process Eng., 37 (2020), doi: 10.1016/j.jwpe.2020.101388.
[29] B. Szeląg, K. Barbusiński, J. Studziński, Activated sludge process modelling using selected machine learning techniques, Desal. Water Treat., 117 (2018) 78–87.
[30] N. Hvala, J. Kocijan, Design of a hybrid mechanistic/Gaussian process model to predict full-scale wastewater treatment plant effluent, Comput. Chem. Eng., 140 (2020), doi: 10.1016/j.compchemeng.2020.106934.
[31] F. Bagherzadeh, M.-J. Mehrani, M. Basirifard, J. Roostaei, Comparative study on total nitrogen prediction in wastewater treatment plant and effect of various feature selection methods on machine learning algorithms performance, J. Water Process Eng., 41 (2021) 102033, doi: 10.1016/j.jwpe.2021.102033.

Supplementary information

Table S1
Classification matrix

Classification		Forecast decisions	
		Positive	Negative
Observed	Positive	True positive (TPos)	False negative (FNeg)
Decisions	Negative	False positive (FPos)	True negative (TNeg)