Pol. J. Environ. Stud. Vol. 29, No. 2 (2020), 1101-1110
DOI: *10.15244/pjoes/99902* ONLINE PUBLICATION DATE: 2019-10-30

Original Research

Effect of Parametric Uncertainty of Selected Classification Models and Simulations of Wastewater Quality Indicators on Predicting the Sludge Volume Index

Krzysztof Chmielowski[1]*, Dawid Bedla[2], Ewa Dacewicz[1], Lubos Jurik[3]

[1]University of Agriculture in Cracow, Department of Sanitary Engineering and Water Management, Kraków, Poland
[2]University of Agriculture in Cracow, Department of Ecology Climatology and Air Protection, Kraków, Poland
[3]Slovak University of Agriculture in Nitra, Department of Water Resources and Environmental Engineering, Nitra, Slovakia

Received: 21 May 2018
Accepted: 21 November 2018

Abstract

This article presents a method for assessing the impact of the predictive uncertainty of selected wastewater quality indicators and the parametric uncertainty of classification models on the forecast results of simulating activated sludge sedimentation using classification models. The data for the calculations were obtained from monitoring carried out at a municipal wastewater treatment plant with a capacity of 72,000 m^3/d^1, located in the Sitkówka-Nowiny commune. The treatment plant receives wastewater, mostly from Kielce city. In the article the possibility of modeling the sedimentation of activated sludge at a wastewater treatment plant using logistic regression and Gompertz models was presented. The included values of the variables (i.e., sewage quality indicators) have been predicted by black-box methods (support vectors and k-nearest neighbor). This approach can be used to improve the operational efficiency of the bioreactor when continuous measurements of sewage quality are not available.

Keywords: sludge volume index, wastewater, simulations

Introduction

Commonly used models for predicting natural phenomena include classification models. These may be black box methods (neural networks, support vector machines, random forests, etc.) or models giving an explicit dependence enabling conclusive assessment of the influence of the explanatory variables on the result of calculations. However, because in this model the result is a specific discrete value, the result often proves unsatisfactory. In the case of classification models, the problem is of a different nature, because only allocation to the appropriate class is modelled, which in many cases makes it possible to identify technological parameters with a much smaller number of input variables than in

*e-mail: k.chmielowski@ur.krakow.pl

the case of a regression model. The predictive ability of regression models is defined on the basis of commonly used measures of fit, such as correlation coefficient, relative error, etc., while in the case of classification models, the accuracy (number of events that have been correctly identified) is determined. In addition to these parameters, the standard deviations of the coefficients determined in the models are to some extent a measure of their accuracy. In classification models, however, these errors are not taken into account in the calculation results, which may lead to certain discrepancies in the results, as confirmed by the work of [1, 2].

An area where black box methods can be used successfully is wastewater treatment plants, in which one of the key parameters determining the operation of the biological reactor is the sludge volume index. However, in these models (regression and classification), the relationships cannot be conclusively identified, which means that analysis of the results is relatively complicated and requires additional calculations to assess the suitability of the model. In the case of classification models, such as logistic regression, linear discriminant analysis, etc., the calculations and identification of dependencies between explanatory and explained variables are much less complex than in black box models. The models mentioned are explicit dependencies, which, as shown in numerous studies, can be used to forecast complex phenomena such as bulking of activated sludge at wastewater treatment plants. In addition to problems associated with assessing the correctness of the model, limitations in their use may appear at the operational stage. This is because sewage quality parameters have a significant influence on sedimentation capacity, but due to the time and cost of determining them, it may be difficult to obtain continuous measurements. In practice, this is a significant limitation on the ability to identify sludge sedimentation capacity and to correct it by changing the operating parameters of the biological reactor. The wastewater quality indicators included in the mathematical model can in fact be determined by mathematical modelling. Here an important question arises as to how the forecast errors of individual indicators included in the model will affect the results of analyses obtained by the mathematical model for predicting sedimentation capacity. At the same time, doubt arises as to whether the sludge sedimentation capacity results obtained in this manner will be reliable and can be used in practice by the plant operator to control the process if the analyses also considers the error in estimating the parameters of the mathematical model.

In response to these questions, the article considers the possibility of forecasting the sedimentation capacity of activated sludge using classification models in which wastewater quality indicators have been calculated using mathematical models. As the models have limited predictive capacity, in order to assess the suitability of the proposed methodology at the stage of operation of the wastewater treatment plant, the impact of the forecast error of individual wastewater quality indicators and estimation of coefficients in the classification models on the accuracy of simulation of the activated sludge volume index was analysed in detail.

Experimental

Study Site

The data for the calculations were obtained from monitoring carried out at a municipal wastewater treatment plant with a capacity of 72,000 m^3/d, located in the Sitkówka-Nowiny commune. The plant receives wastewater from the sanitary sewage system in Kielce, the Sitkówka-Nowiny commune and part of the Masłów commune. First, the wastewater is pre-treated mechanically (step screens and an aerated grit chamber with a separate grease trap), and then passes to the primary sedimentation tanks, from which it is directed to the biological part. Organic, nitrogen and phosphorus compounds are removed from the wastewater in a 5-stage Bardenpho system with a separate pre-denitrification tank. Following biological treatment, the sewage flows to four secondary sedimentation tanks, where it is separated from the activated sludge and enters the Bobrza River.

Since 2012 the monitoring carried out at the Sitkówka-Nowiny wastewater treatment plant has included measurements of the amount of wastewater (Q) and quality indicators, i.e., biochemical oxygen demand (BOD_5), chemical oxygen demand (COD), suspended solids (TSS), total nitrogen (TN), ammonium nitrogen ($N\text{-}NH_4$), nitrate nitrogen ($N\text{-}NO_3$), total phosphorus (TP), and chloride (Cl), as well as operating parameters of the biological reactor: pH, temperature (T_{sl}), activated sludge concentration (MLSS), recirculated sludge concentration (REC), sludge recirculation ratio (RAS), dose of chemical coagulant (PIX), and oxygen concentration in nitrification tanks (DO).

Methods

The article presents a method for assessing the impact of the predictive uncertainty of selected wastewater quality indicators and the parametric uncertainty of classification models on the forecast results of simulating activated sludge sedimentation using selected classification models. The steps of the algorithm are shown in Fig. 1.

The diagram in Fig. 1 shows that the first step is to develop classification models to identify sludge sedimentation. The purpose of this is to determine the explanatory variables (x_i) of the modelled value of the sludge volume index. In this study, in order to compare the prediction capabilities of classification models, we considered the use of the logistic regression method and the Gompertz model. These methods are commonly

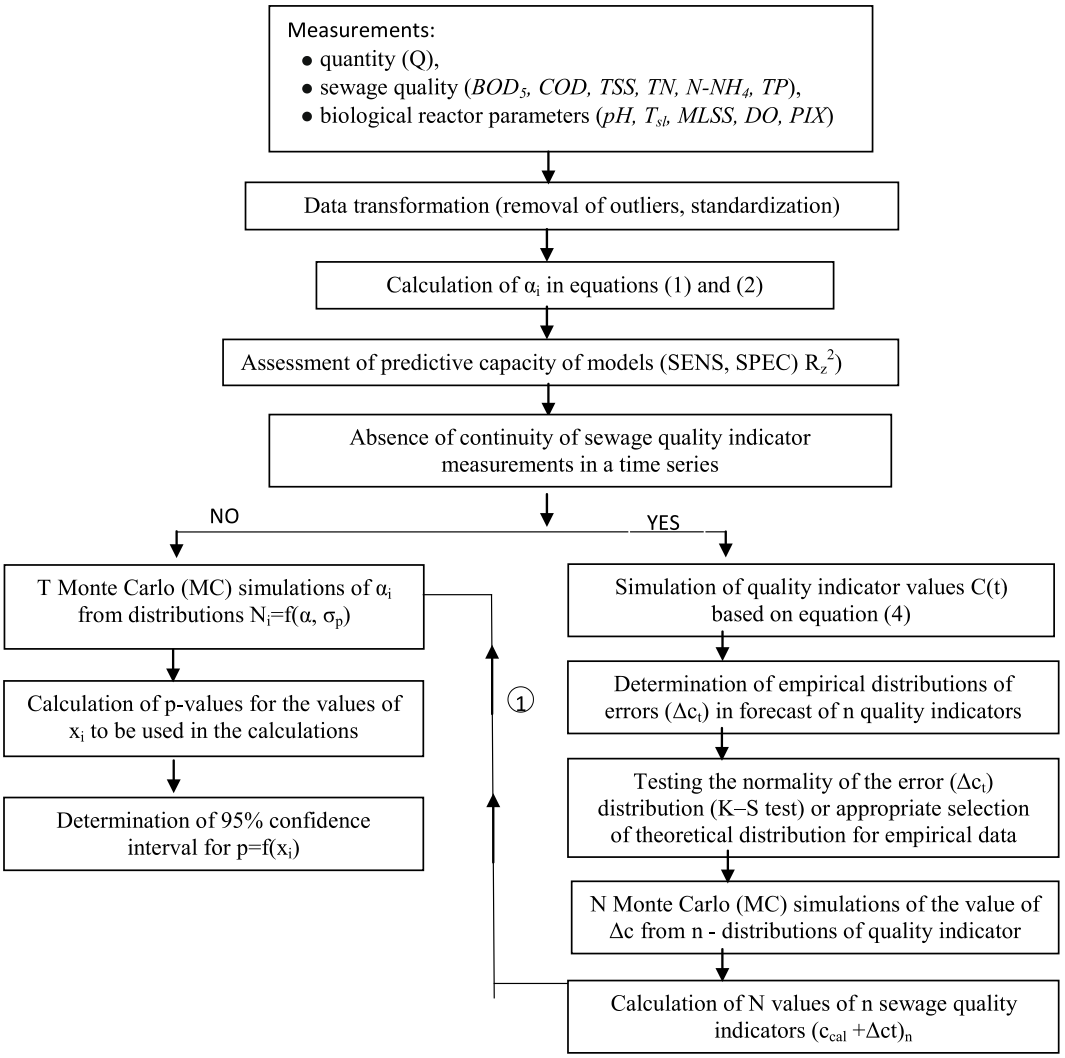

Fig. 1. Calculation diagram of the influence of parametric uncertainty and uncertainty of forecast of wastewater quality indicators on SVI simulation by classification models.

used in medicine, microbiology and the social sciences, and a review of the literature [1, 3, 4] shows that the Gompertz model has shown much greater capacity to classify both economic events and disease cases than logistic regression. Therefore, analyses should be performed to compare the capabilities of these models to predict sedimentation and to identify the model whose use in everyday practice will assist the plant operator in making appropriate decisions, thereby improving the efficiency of the facility's operation. The classification models discussed in the paper are described by the following relationships [5]:

Logistic regression:

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

Gomperzt model:

$$p(X) = exp\left(-exp(-X)\right) \tag{2}$$

…where X is the vector constituting a linear combination of independent variables (x_i), i.e., the quantity and quality of wastewater and the operating parameters of the biological reactor described as:

$$X = \alpha_0 + \sum_{i=1}^{m} \alpha_i \cdot x_i \tag{3}$$

Only explanatory variables (x_i) that were statistically significant at the adopted confidence level (α = 0.95) were included in the classification models. In these analyses, the criterion used to assess the impact of the sewage quality indicators and operational parameters of activated sludge chambers was the value of the sludge volume index, SVI_{lim} = 150 ml/g [4]. Coefficients of correlation (McFadden R^2, Cox–Snell R^2 and Nagelkerke R^2), as well as the values for sensitivity (SENS), specificity (SPEC) and calculation error (R_z^2) [5] were used to assess the predictive power of the classification models.

In the case of classification models [5]:
Sensitivity (SENS):

$$SENS = 100 \cdot \frac{TP}{TP+FN}$$

Specificity (SPEC):

$$SPEC = 100 \cdot \frac{TN}{FP+FN}$$

Calculation error (R_z^2):

$$R_z^2 = 100 \cdot \frac{TN+TP}{TP+TN+FP+FN}$$

The number of indications classified accordingly, where:
– *FN* – false negative
– *TN* – truly negative
– *TP* – truly positive
– *FP* – false negative

In the next step, if the explanatory variables in models (1) or (2) are values of wastewater quality indicators and there are problems with obtaining continuity of measurements, a black box method can be used to supply the missing values. In this study, wastewater quality was predicted using the support vector machine (SVM) and k-nearest neighbour (k-NN) methods, which have previously been used for this purpose by other researchers, such as [6-8]. It was assumed that n wastewater quality indicators will be predicted based on the values of sewage flow and temperature at time $(t - m)_n$ and $(t - k)_n$, which can be written as follows [8, 9]:

$$C(t) = f(Q(t-1), Q(t-2), Q(t-m), \ldots, T_{in}(t-1), T_{in}(t-2), T_{in}(t-k))_n \quad (4)$$

…where:
C(t) - concentrations of selected indicators
Q(t-m) - inflow of sewage to the watertreatment plant at the moment (t-m)
Tin(t-k) - temperature of inflowing sewage at the moment (t-k)

In practice, this makes it possible to select in advance the appropriate values for the reactor operating parameters in order to obtain the required technological effect of the activated sludge chambers. The boosted trees method was used to identify explanatory variables, as discussed in detail by [10].

The k-nearest neighbour method used in the study is one of the simplest nonparametric methods, in which the value of the explained variable is determined according to the following relationship [11]:

$$\hat{y} = \frac{1}{K} \cdot \sum_{m=1}^{l} y_i \cdot J\left(x_i, x_j\right) \quad (5)$$

…where x_i is one of K nearest neighbours of x_j when the distance $d(x_i,x_j)$ is one of the smallest distances between observations; l is the number of observations; and $J(x_i,x_j)$ is a function assuming the following values: 1 when x_i is one of K nearest neighbours of x_j, and 0 when the first condition is not met. The Mahalanobis distance, which is the distance between two points in an n-dimensional space differentiating the contribution of individual components, was used in the calculations. In the above case, the number of nearest neighbours K was determined by the method of successive approximations to obtain the minimum mean absolute error (MAE) and mean absolute percentage error (MAPE), which is common practice.

Support vector machines (SVM), whose model structure resembles a neural network, have much greater model complexity than the k-NN method, and thus its predictions have been shown to be more accurate in numerous studies [8]. In this method, a kernel function is used to transform an N-dimensional non-linear space to a higher K-dimensional linear space, and simulation results are determined based on the following formula [12]:

$$y = \sum_{i=1}^{N_{sv}} (\alpha_i - \alpha_i^{'}) \cdot K(x, x_i) + w_0 \quad (6)$$

…in which: N_{sv} is the number of support vectors corresponding to the number of non-zero Lagrange coefficients dependent on C and ε; α_i is Lagrange coefficients; and $K(x,x_i)$ is kernel function. The values of coefficients α_i and the number of support vectors (N_{sv}) are determined by minimizing the following expression:

$$\sum_{i=1}^{Z} \frac{C}{l} \cdot |y_i - f(x_i)|_\varepsilon + \frac{1}{2} \cdot \|f_k\|^2 \quad (7)$$

…where $|y_i - f(x_i)|\varepsilon = max\{0, |y_i - f(x_i)| - \varepsilon |\}$, ε – acceptable error, l – number of observations in the training set, C – constant dependent on ε, $\|f_k\|$ – norm of f in the Hilbert space.

It follows from formulas (6) and (7) that the prediction capabilities in the SVM method are influenced by three parameters: capacity (C), the kernel function, and the insensitivity threshold ε = 0.01, determined based on the work of Burges [13] and Rutkowski [14]. A sigmoid function was used to predict quality indicators in equation (3), which is common practice in this type of application. The optimal C and kernel function (γ = 0.2 ÷ 1.0) were found by the method of successive approximations until the minimum MAE and MAPE values were obtained. STATISTICA 10 software was used to develop the models described above for predicting selected wastewater quality indicators by the k-NN and SVM methods.

Based on the computation results (classification models for simulation of sludge sedimentation and prediction of selected sewage quality indicators), an analysis was carried out to assess the impact of sewage quality prediction uncertainty and the parametric uncertainty of the classification models on the results of

calculations of the probability of exceeding the sludge volume index limit.

The diagram in Fig. 1 shows that the calculations of the impact of the forecast error of individual wastewater quality indicators on simulation of p is based on the following steps:

A) Identification of n error distributions of the forecast of quality indicators $\Delta c_t = f(\mu,\sigma)_n$ and fit of the empirical data to the theoretical distributions.

B) Calculations of Spearman correlation coefficients (R) between pairs of indicators; where the R value indicates a significant correlation between variables, appropriate data sampling methods should be used, such as the Iman-Conover method.

C) T Monte Carlo simulations of the indicator forecast errors Δc_n from n theoretical distributions.

D) Calculation of the lth value of n quality indicators in T samples as $(c_{cal} + \Delta c)_n$.

E) T calculations of the lth p values from formulas (1) and (2) with the values assumed for the remaining explanatory variables in the classification models.

F) Determining confidence intervals for individual p values corresponding to the values assumed for x_i.

In the case where the input data (i.e., quality indicators), constitute a continuous series of measurements, then according to the calculation diagram (Fig. 1), the influence of the parametric uncertainty of the model on the results of the calculations of the probability of SVI exceedance is analysed. The procedure is identical to the one described above, and the basis for identification of i theoretical distributions is the mean values obtained for coefficients α_i and standard deviations (σ_i), which are parameters of normal distributions N (α_i, σ_i).

In addition, the diagram in Fig. 1 shows that it is possible to simultaneously assess the impact of both parametric uncertainty and uncertainty of the forecast of wastewater quality indicators on the results of calculations of p from formulas (1) and (2). In this case, first T Monte Carlo simulations of forecast errors of the lth values of n quality indicators are performed as described above, and then the values α_i and σ_i are additionally simulated N times, and on this basis the values for the confidence intervals for individual p values are determined based on N·T calculations.

Results and Discussion

The results of measurements of the quantity and quality of sewage and the operating parameters of the biological reactor (Table 1) indicated that the data vary widely, which leads to significant changes in the activated sludge volume index.

During the research period, there were problems with sedimentation of activated sludge at the wastewater

Table 1. Range of variation in the values of parameters describing the quantity and quality of sewage and the operational parameters of activated sludge chambers [3].

Variable	Unit	Minimum	Mean	Maximum	1st quantile	5th quantile
Q	m^3/d	32,564	40,698	86,592	37,205	47,095
T_{in}	^0C	10.6	16.3	20.9	14.3	18.5
T_{sl}	^0C	10.0	15.9	23.0	12.88	18.00
pH	-	7.2	7.7	7.8	7.50	7.70
MLSS	kg/m^3	1.98	4.26	6.59	3.85	5.86
REC	kg/m^3	6.54	8.61	9.84	8.22	9.11
RAS	%	44.6	91.4	167.5	72.54	115.25
F/M	kgBOD/kgMLSS·d	0.030	0.070	0.150	0.051	0.086
PIX	m^3/d	0	0.8	1.93	0.48	0.85
DO	mg/dm^3	0.55	2.56	5.78	1.90	2.26
SVI	cm^3/g	95	166	320	149.63	199.6
BOD_5	mg/dm^3	127	309	557	253.5	394.25
COD	mg/dm^3	384	791	1250	660.5	885.5
TSS	mg/dm^3	126	329	572	274.0	350.5
$N\text{-}NH_4^+$	mg/dm^3	24.4	49.4	65.9	43.8	55.0
TN	mg/dm^3	39.9	77.7	124.1	68.63	83.25
$N\text{-}NO_3$	mg/dm^3	0.05	0.17	1.20	0.05	0.16
TP	mg/dm^3	4.30	7.80	12.60	6.70	8.40

treatment plant, as confirmed by both the mean and maximum value of the sludge volume index. In practice, this leads to problems at the stage of operation of the treatment plant, and in order to eliminate them a mathematical model must be developed to predict the sedimentation capacity of the activated sludge. Based on measurements of the quantity and quality of inflowing sewage and the operational parameters of the reactor, we used STATISTICA software to develop classification models in which a vector constituting a linear combination of variables is described by the following formula (7):

$$X = \alpha_1 \cdot \frac{BOD_5}{TN} + \alpha_2 \cdot \frac{BOD_5}{TP} + \alpha_3 \cdot L_{N-NH_4} + \alpha_4 \cdot MLSS + \alpha_5 \cdot T_{sl} + \alpha_6 \cdot m_{PIX} + \alpha_7 \cdot DO + \alpha_e \cdot REC + \alpha_0 \quad (7)$$

…where:
BOD_5 – biological oxygen demand
TN – total nitrogen
TP – total phosphorus
MLSS – activated sludge concentration
Tsl – sludge temperature
mPIX – number of PIX dosing
DO – oxygen concentration
REC – recirculated sludge concentration

The numerical values of coefficients α_i in the logit and Gompertz models and parameters of the fit of the measurement results to the calculations are presented in Table 2.

Analysis of the data in Table 2 shows that the classification models have satisfactory predictive capabilities, but the calculation results were better fitted to the measurements in the Gompertz model than in the logit model, as indicated by the SENS, SPEC and R_z^2 values. The data in Table 2 show that the models have identical event identification capability when SVI<150 ml/g, as confirmed by the value of SPEC = 0.9886. However, the Gompertz model was more capable than the logit model of identifying cases where SVI>150 ml/g, as indicated by the SENS values, which in these cases were 0.9971 and 0.8795. The functional dependencies $p = f(x_i)$ are confirmed by analyses performed by [4, 15, 16] at municipal wastewater treatment plants with a flow-through system.

In the next stage, based on the classification dependencies obtained (Table 2), models were developed to forecast the flow rate and sewage quality indicators, i.e., BOD_5, TN, TP and $N\text{-}NH_4$. The variables explaining

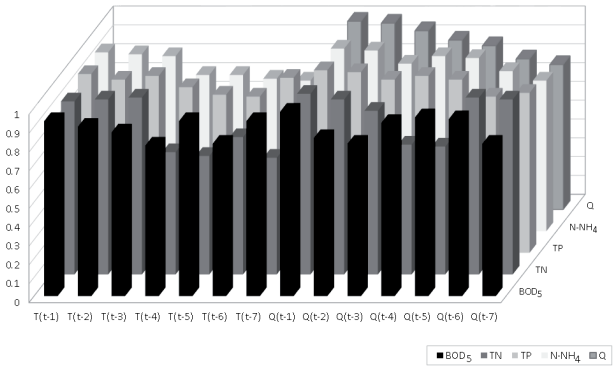

Fig. 2. Importance values (IMP) for individual variables explaining sewage quality indicators (BOD_5, TN, TP and $N\text{-}NH_4$) and flow rate.

Table 2. Numerical values for coefficients β_i in the logit and Gompertz models and parameters of the fit of measurement data to calculation results.

Variable	Logit			Gompertz		
	α_i	Standard deviation	p	α_i	Standard deviation	p
BOD_5/TN	0.0530	0.002	0.001	0.0530	0.002	0.001
BOD_5/TP	0.0140	0.001	0.039	0.015	0.001	0.042
L_{N-NH4}	0.0011	0.0001	0.035	0.0010	0.0001	0.037
T_{sl}	-0.517	0.030	0.032	-0.517	0.024	0.016
MLSS	-2.543	0.150	0.019	-2.184	0.080	0.019
REC	1.339	0.290	0.031	1.198	0.150	0.022
DO	-1.644	0.280	0.022	-1.656	0.160	0.015
PIX	-0.723	0.210	0.026	-0.849	0.090	0.016
Absolute term	9.965	0.058	0.023	9.563	0.060	0.001
	$R^2_{McFadden} = 0.781$	SENS	0.8795	$R^2_{McFadden} = 0.958$	SENS	0.9971
	$R^2_{Cox\text{-}Snell} = 0.599$	SPEC	0.9886	$R^2_{Cox\text{-}Snell} = 0.834$	SPEC	0.9886
	$R^2_{Negelkerke} = 0.868$	R^2_z	0.9261	$R^2_{Negelkerke} = 0.919$	R^2_z	0.9891
	AIC = 37.787	SBC	58.882	AIC = 30.822	SBC	51.916

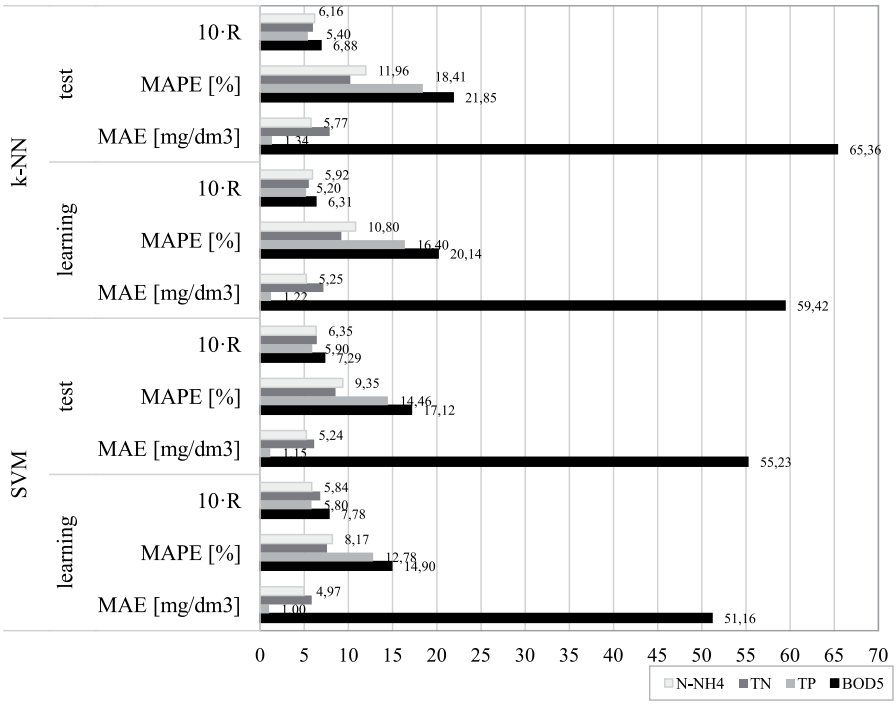

Fig. 3. Parameters (MAE, MAPE, 10·R) of the fit of the results of calculations by SVM and k-NN to the measurements of wastewater quality indicators.

individual indicators were determined based on the importance values (IMP) as calculated by the boosted trees method (Fig. 2).

The results shown in Fig. 2 are confirmed by the results of calculations made by the author in previous studies [9, 17], which means that the quality of the sewage quality indicators is influenced by the flow rate and temperature of the incoming sewage to the treatment plant. This indicates that the quality of wastewater in the sewage system is influenced by the degree of wastewater dilution and the biochemical reactions taking place in the wastewater flowing through the sewers. Based on the data in Fig. 2, models were determined to forecast the flow rate and sewage quality indicators using the SVM and k-NN methods; parameters of the fit of calculation results to measurements for the training and test sets are shown in Fig. 3.

In the case of the statistical models for predicting sewage quality obtained by the SVM method, the value of C was 5 ÷ 12 and $\gamma = 0.25 \div 0.65$, and for simulation of Q they were C = 7 and $\gamma = 0.50$. The number of neighbors (K) in the k-NN method ranged from 7 to 10. Analysis of the results (Fig. 4) indicates that the SVM method resulted in smaller errors in fitting the calculation results to the measurements than the k-NN method. The R value for the model for predicting BOD_5 using the SVM method is greater than that determined by [18] (R = 0.83) using the ANN method, where the temperature and pH of inflowing sewage and the TSS concentration were used to forecast the quality index. Our results are also confirmed by analyses carried out by [17, 19] at wastewater treatment plants. Furthermore, analysis of the error values obtained in the case of TP, TN and $N\text{-}NH_4$ shows that [6] obtained a better fit using the k-NN method to predict the indicators, optimizing the number of neighbours K accordingly.

In the case of the model simulating the rate of sewage inflow to the WWTP by SVM and k-NN, the forecast error values were MAE = 1923 m^3/d, MAPE = 4.57% and R = 0.92 for the first method and MAE = 3446 m^3/d, MAPE = 6.54% and R = 0.86 for the second. Based on the results, we calculated forecast errors for the values of discrete quality indicators and flow rate, prepared empirical distributions of errors, and tested the normality of the error distributions using the Kolmogorov–Smirnov test. The analyses showed no grounds to reject the hypothesis that the distributions were normal, which is confirmed by the p values (Table 3). On this basis, the parameters of the

Table 3. P-values of the Kolmogorov–Smirnov test and parameters of statistical distributions of errors in the forecast of sewage quality indicators by the SVM and k-NN methods.

Quality indicator	SVM		k-NN	
	Standard deviation	p	Standard deviation	p
TP	1.44	0.18	1.62	0.23
BOD_5	64.80	0.12	70.15	0.16
TN	7.15	0.11	7.50	0.18
$N\text{-}NH_4$	6.01	0.20	6.15	0.24
Q	4291	0.06	5396	0.07

distributions were determined, and these indicated that mean values were equal to 0; the standard deviations for individual variables are presented in Table 3. At the same time, the calculations showed that the values of the correlation coefficient R between pairs of individual variables Δc_i (Table 3) are less than R = 0.26, which indicates a weak correlation, and therefore the relationship between Δc_i was omitted from further analyses.

Based on the results of the calculations, Fig. 4 presents curves for p = f (MLSS, DO, REC) taking into account the inaccuracy of the forecast of wastewater quality indicators. Fig. 6 presents example logit curves taking into account the uncertainty of the estimated parameters (α_i) in equations (1) and (2).

The results (Figs 4, 5) indicate that among the parameters considered, the uncertainty of the estimated parameters in Eq. (2) has the greatest influence on the accuracy of the forecast of the probability of SVI_{lim} exceedance, as confirmed by the 95% confidence interval. The uncertainty of SVI forecast (SVI = 150 ml/g) is less influenced by the accuracy of the forecast of selected sewage quality indicators (BOD_5, TN, TP and $N\text{-}NH_4$) and the flow rate, as illustrated in Fig. 5.

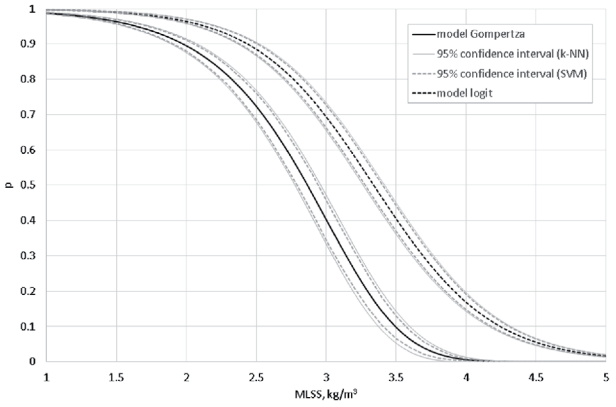

Fig. 4. Example of logit and Gompertz curves taking into account the influence of uncertainty in the forecast of wastewater quality indicators by the k-NN and SVM methods.

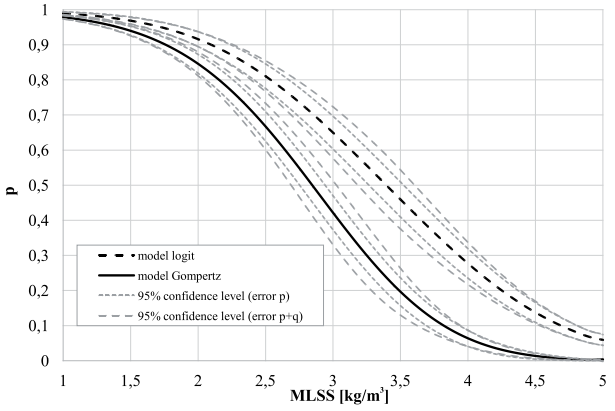

Fig. 5. Example logit and Gompertz curves taking into account the influence of the uncertainty of estimating parameters α_i in the model and of the forecast of wastewater quality indicators by the k-NN method.

Of the two methods considered for modelling wastewater quality, lower forecast uncertainty of the probability of SVI_{lim} exceedance was obtained using the SVM method than the k-NN method (Fig. 4), which is also indicated by the parameters of the fit of calculation results to measurements [3, 20]. Fig. 5 shows examples of logit and Gompertz curves taking into account the parametric uncertainty of the models and the error in predicting indicators using only the k-nearest neighbour method, based on the calculation diagram presented in Fig. 1. The calculations shown in Fig. 5 indicate that simultaneous consideration of the forecast errors (of quality indicators and model parameters) indicates an increase in the uncertainty of the prediction of the probability of SVI exceedance compared to the case where these uncertainties were considered individually. In practical considerations, at the stage of modelling and controlling the reactor parameters, this is crucial for the correct operation of the biological reactor.

In the analyzed case, the considered variables in the classification models are independent. Based on this, simulators of their values were developed using the Monte Carlo method, omitting their correlation. In other cases, the correlation between variables should be taken into account [20]. Independent variables describing the quality of wastewater are forecasted using statistical models. By substituting the quality classification parameters with the error of the forecasting model, the result obtained allows for the optimal selection of the bioreactor's operating parameters. Taking into account the error of the forecast of wastewater quality indicators as well as the estimation error of the coefficients in the model, we increase the safety of the wastewater treatment plant operation and limit the wrong technological decisions.

Conclusions

Mathematical models are a tool used to analyse the relationships between an explained variable and explanatory variables, an example of which is their use for predicting and thus controlling the sedimentation of activated sludge at wastewater treatment plants. From a practical point of view, analysis of the agreement between the modelled and measured course of the process is based on the parameters of the fit of the results to the calculations. However, when the explanatory variables included in the model are difficult to determine and there are problems with their continuity, they can be predicted, resulting in calculation results that are subject to uncertainty. Our paper presents the possibility of modelling the sedimentation of activated

sludge at a wastewater treatment plant using logistic regression and Gompertz models, in which values of the variables included, i.e., sewage quality indicators, have been predicted by black-box methods (support vectors and k-nearest neighbour). The approach presented in the paper is innovative. It makes it possible to assess the reliability of the results of simulation of activated sludge sedimentation capacity using mathematical models in which the values of measured quality indicators are replaced with the results of calculations, with the parametric uncertainty of the model taken into account as well. This approach can be used to improve the operational efficiency of the bioreactor when continuous measurements of sewage quality are not available, but also takes into account the accuracy of the model's forecast of sludge sedimentation, which most previous studies [20-22] have failed to consider.

The calculations showed that both logistic regression and the Gompertz model can be used to identify activated sludge sedimentation. However, it should be noted that the results of sludge sedimentation forecasts using the Gompertz model show a better fit of the measurement results to the calculations. The statistical analyses performed in the study revealed a smaller error in the sludge sedimentation forecast when the quality indices were modelled using the support vector machine method than when the k-nearest neighbor method was used. The calculations indicated greater uncertainty of identification of activated sludge sedimentation when the statistical models included parameter identification error rather than forecast errors of individual wastewater quality indicators. In addition, the statistical analyses show that errors in both sewage sludge quality forecast and parameter estimation play an important role in predicting activated sludge sedimentation, which has a significant impact on the ability to simulate sedimentation capacity as well as the ability to control reactor parameters.

Based on the results, it can be concluded that classification models can be used to simulate sedimentation of activated sludge. Among the considered models (probit, logit, G-tz), the best fit was obtained using the logit model.

Conflict of Interest

The authors declare no conflict of interest.

References

1. GRANIERO P.A., PRICE J.S. Distribution of bog and heath in a Newfoundland blanket bog complex: Topographic limits on the hydrological processes governing blanket bog development. Hydrology and Earth System Sciences, **3**, 223, **1999**.
2. HUSMEIER D., DYBOWSKI R., ROBERTS S. (ed.). Probabilistic modeling in bioinformatics and medical informatics. Springer Science & Business Media, 495, **2006**.
3. SZELĄG B., SIWICKI P. Application of the selected classification models to the analysis of the settling capacity of the activated sludge – case study. E3S Web of Conferences **17**, 8, **2017**.
4. BAYO J., ANGOSTO J. M., SERRANO-ANIORTE J. Evaluation of physicochemical parameters influencing bulking episodes in a municipal wastewater treatment plant. Water Pollution VIII: Modelling, Monitoring and Management. **95**, 531, **2008**.
5. HARRELL F. Regression modeling strategies with application to linear models, logistic regression and survival analysis. Springer Verlag, New York, **2001**.
6. MINSOO K., YEJIN K., HYOSOO K., WENHUA P., CHANGWON K. Evaluation of the k – nearest neighbour method for forecasting the influent characteristics of wastewater treatment plant. Frontiers of Environmental Science and Engineering. **10** (2), 299, **2016**.
7. KUSIAK A., VERMA A., WEI X. A data-mining approach to predict influent quality. Environmental Monitoring and Assessment. **185**, 2197, **2013**.
8. SZELĄG B., BARTKIEWICZ L., STUDZIŃSKI J., BARBUSIŃSKI K. Evaluation of the impact of explanatory variables on the accuracy of prediction of daily inflow to the sewage treatment plant by selected models nonlinear. Archives of Environmental Protection, **2017** [In press].
9. LUBOS J., KALETOVA T., SEDMAKOVA M., BALAZOVA P., CERVENANSKA A. Comparison of service characteristics of two town's WWTP. Journal of Ecological Engineering. **18** (3), 61, **2017**.
10. VERMA A., WEI X., KUSIAK A. Predicting the total suspended solids in wastewater: data-mining approach. EngineeringApplications of Artificial Intelligence, **26** (4) 1366, **2013**.
11. PIOTROWSKI A., OSUCH M., NAPIÓRKOWSKI M.J., ROWIŃSKI P.M. NAPIÓRKOWSKI J.J. Comparing large number of metaheurestics for artifi cial neural networks training to predict water temperature in a natural river, Computers & Geosciences, **64**, 136, **2014**.
12. VAPNIK V. Statistical Learning Theory. John Wiley and Sons. New York, **1998**.
13. BURGES, CHRISTOPHER JC. A tutorial on support vector machines for pattern recognition. Data mining and knowledge discovery, **2.2**, 121, **1998**.
14. RUTKOWSKI L. Flexible neuro-fuzzy systems: structures, learning and performance evaluation. Springer Science & Business Media. Kluwer Academic Publisher, **2004**.
15. LUO I., ZHAO Y. Sludge bulking prediction using principle component regression and artificial neural network. Mathematical Problems in Engineering, **2012**.
16. BEZAK-MAZUR E., STOIŃSKA R., SZELĄG B. Rating of impact of operational parameters and occurrence of filamentous bacteria on the volumetric active sediment index - case study Rocznik Ochrona Środowiska. **18**, 487, **2016**.
17. SZELĄG B., STUDZIŃSKI J. A data mining approach to the prediction of food-to-mass ratio and mixed liquor suspended solids. Pol. J. Environ. Stud. **26** (5), 2231, **2017**.
18. ABYANEH H.Z. Evaluation of multivariate linear regression and artificial neural networks in prediction of water quality parameters. Journal of Environmental Health Science and Engineering. **12** (40), 1, **2014**.
19. DOGAN E., ATES A., YILMAZ E.C., EREN B. Application of artificial neural networks to estimate wastewater treatment plant inlet biochemical oxygen demand. Environmental progress. **27** (4), 439, **2008**.

20. SZELĄG B., GAWDZIK J. Application of selected methods of artificial intelligence to activated sludge settleability predictions. Pol. J. Environ. Stud. **25** (4), 1709, **2016**.

21. SZELĄG B., GAWDZIK J., GAWDZIK A. Application of selected methods of black box for modelling the settleability process in wastewater treatment plant. Ecological Chemistry and Engineering. **24** (1), 119, **2017**.