Study of energy consumption in a wastewater treatment plant using logistic regression

C Boncescu*1, L D Robescu1, D A Bondrea1 and M E Măcinic1

1University POLITEHNICA of Bucharest, Department of Hydraulics, Hydraulic Machinery and Environmental Engineering, 313 Splaiul Independentei, Sector 6, 060042, Bucharest, Romania
Email: corina.boncescu@gmail.com

Abstract. The aim of this paper is to predict the energy consumption of a wastewater treatment plant from Romania, taking into account the flowrate, concentration of BOD, TSS, COD and the energy consumption. For the mathematical model the logistic regression was applied. The input data used were from a waste treatment plant in Romania, for a period of 2 years 2015 and 2017, a total of random 403 dataset. The treatment technologies of WWTP consist of advanced biological treatment SBR (nitrification, denitrification, and phosphorus removal), aerobic sludge stabilization, dewatering, storage and chemical disinfection. Octave software was used to build the model. The answer of the model refers to the fact that for a given situation there will be high energy consumption or low energy consumption in the wastewater treatment plant (WWTP). Performance of the model was compared with real value.

1. Introduction
Water resources are the basis of human life and economic development and are closely linked to health and the environment. Accurate water quality prediction is the key to improving water management and pollution control. In general, the water quality of a treatment plant is sensitive to parameters such as pH, temperature, concentrations of specific substrates and contaminants. Two hybrid models are used to predict six indicators of water quality, including water temperature, dissolved oxygen, pH value, specific conductance, turbidity and dissolved fluorescent organic matter. The results show that short-term water quality may be affected by external factors, so the data may vary over time and have a high degree of nonlinear characteristics, which undoubtedly increases the difficulty of prediction [1]. An essential advantage of prediction models for evaluating the performance of treatment plants is that they can directly predict the output values only after the training and validation stage. Machine learning models could be a reliable method for predicting water quality; early warning of water quality control for wastewater treatment can eliminate many risks [2-4].

Most of the researchers focused on the application of machine learning models to determine effluent quality parameters: total nitrogen biological oxygen demand, chemical oxygen demand, and total suspended solids, total phosphorous taking into account flowrate quality parameters [5]. Biochemical oxygen demand (BOD), chemical oxygen demand (COD) and total suspended solids (TSS) are the most frequently regulated parameters for wastewater effluents. Measurement and prediction of these parameters are essential for performance evaluation and modernization of wastewater treatment plants [6]. Thus, the researchers performed an analysis of linear and nonlinear
models. For each and every parameter analysed, 144 different simple and compound linear models are presented. The combined models have led to significant improvements and outstanding performance. The proposed non-linear and hybrid models have a great potential to replace conventional models in order to predict wastewater quality parameters and to improve the operation and design of current wastewater installations [7].

Neural networks models have been applied both for water quality analysis and for sludge properties. Most models were based on the analysis of organic load in water. For an improved analysis of BOD using artificial intelligence techniques, it has been proposed models that have included the impact of sludge quality [8]. A model based on a deep learning network has been developed to predict the performance of a biofilm system. The model considers concentrations of chemical oxygen demand (COD), ammonia (NH4 + -N) and total nitrogen (TN). The proposed model could serve as a tool to predict the performance of the biofilm system [9].

Most machine learning techniques are based on regression models. Multiple regression models were applied for the analysis of water quality control in the treatment plants, offering accuracy to the models between 83 and 94% [10]. Logistic regression models were also applied for the analysis of the quality of activated sludge regarding the presence or absence of filamentous bacteria in which case the model was successfully validated [11]. A regression tree-based regression approach was used on a sample of 106 US stations to analyse the amounts of ammonia in the effluent. Validation of the model showed that regression trees can provide a median classification accuracy > 70% [6]. The most recent research in this field has been carried out to determine the water quality indicator and the volume of sludge index, respectively, for a treatment plant in Poland. The results showed an accuracy of 95% [12].

Energy efficiency has been taken into account after the energy crisis from 70’s [13]. It has also been shown that energy consumption worldwide is increasing rapidly [14]. Energy consumptions related to water processes account for 2-3% of the world's electricity thus the energy management of wastewater treatment plants (WWTPs) is considered a current topic [15]. There are numerous methods to determine the energy efficiency but few researchers have taken into account the energy cost. In 2018, the first energy consumption study was conducted using AI technology that links the parameters of quality and energy cost [16]. Current research on reduction energy consumption from the wastewater treatment plants are aimed at applying the artificial intelligence (AI) technology. Using AI algorithms can be made a prediction in time and an optimization of the process [17]. Since 2015, these technologies have been applied to 317 treatment plants in north-western Europe, mainly in Belgium, the Netherlands, Denmark, France, Luxembourg, Austria and Germany [18].

Currently, energy cost models are generated using logarithmic, exponential or linear functions, which become complex when the relationship between variables is extremely complex and nonlinear. Several cost models have been produced to assess the relationship between the most relevant process variables and treatment costs using traditional regression. The literature review allows the identification of three dominant approaches to data-based cost modelling: linear, exponential and logistics [19-20]. Using these approaches, many authors have obtained interesting results and useful management information by calibrating these function forms for databases and their specific purpose [21-24].

In this article we will present a method for predicting energy consumption taking into account the influent flow, the concentration of BOD, TSS, COD, energy consumption and energy cost using the prediction method with logistic regression. In the case of the study, daily data were taken into account for a period of 2 years for a wastewater treatment plant in Romania that uses SBR technology.

2. Material and methods
In this article, the logistic regression was implemented to analyse the probability of energy consumption depending on the influent flow and BOD, TSS and COD concentrations. The daily energy consumption of the treatment plant was also taken into account. The treatment technology of wastewater treatment plant analysed consist of advanced biological treatment SBR (nitrification,
denitrification, and phosphorus removal), aerobic sludge stabilization, dewatering, storage and chemical disinfection. Octave software was used to build the model.

To identify the optimal solution of the mathematical model, the cost function was used and the function gradient was determined for a training data set \((x, y)\).

\[
J(\theta) = \frac{1}{m} \sum_{i=1}^{m} \left[ -y_i \log(h_\theta(x^i)) - (1 - y_i) \log(1 - h_\theta(x^i)) \right]
\]

(1)

\[
\frac{\partial J(\theta)}{\partial \theta} = \frac{1}{m} \sum_{i=1}^{m} \left( h_\theta(x^i) - y^i \right) x^i
\]

(2)

Logistic regression was defined as:

\[
h_\theta(x) = g(\theta^T x)
\]

(3)

where, \(g(z)\) is the sigmoid function defined as:

\[
g(z) = \frac{1}{1 + e^{-z}}
\]

(4)

where, \(m\) - is the number of dataset analysed, \((x, y)\) - is training dataset

In the mathematical model were analysed logistic regressions between the influence flow and the concentrations of BOD, TSS and COD. In order to classify the analysed points, the energy consumption related to the influent flow was taken into account. It was calculated the average energy consumption with eq. 5, 6 and for \(c_i < c_{med}\) were considered a low energy consumption and for \(c_i > c_{med}\) were considered high energy consumption.

\[
c_i = \frac{E_i}{Q_{inf,i}}
\]

(5)

\[
c_{med} = \frac{1}{m} \sum_{i} c_i
\]

(6)

where, \(c_{med}\) - is the specific average energy consumption \([\text{kWh/m}^3]\), \(c_i\) - is the specific energy consumption on day \(i\) \([\text{kWh/m}^3]\), \(E_i\) - is the energy consumption on day \(i\) \([\text{kWh}]\), \(Q_{inf,i}\) - is the influential flow from day \(i\),

For the realization of the model, 403 random dataset were considered during two years 2015 and 2017 for the COD and TSS concentrations, out of which for 263 dataset the BOD concentration is also known. For testing the model were considered a number of 15 random dataset from 2018, for which are known the influent flow, BOD, TSS and COD concentrations and the energy consumptions, respectively. To validate the model the result was compared with real data.

3. Results and Discussion

The first step was to determine the training data set. In order to select the data, it was taken into account that they should be as different as possible from all four seasons. The minimum, maximum and average values were identified for flow, energy consumption and BOD, COD and TSS concentrations. The results obtained are presented in tables 1 and 2.

### Table 1. The values of training data set for BOD

| Training data set | max       | min       | med       |
|-------------------|-----------|-----------|-----------|
| Flow \([\text{m}^3]\)       | 39577     | 4716      | 10403.844 |
| Energy consumption\[\text{kWh}\] | 8722      | 1311      | 3359.762  |
| BOD\[\text{mg/l}\]         | 482       | 42        | 169.238   |
| \(c\) \[\text{kWh/m}^3\]   | 0.465     | 0.143     | 0.318     |
Table 2. The values of training data set for COD and TSS concentration

| Training data set       | max  | min  | med  |
|-------------------------|------|------|------|
| Flow [m³]               | 39577| 4716 | 10630.07 |
| Energy consumption [kWh]| 8722 | 1311 | 3258 |
| COD [mg/l]              | 830  | 73   | 296.8177 |
| TSS [mg/l]              | 1015 | 57   | 287.7687 |
| c [kWh/m³]              | 0.476| 0.143| 0.314 |

The ratio between energy consumption and influent flow are presented in figures 1 and 2. It was identified that the average value in the first case was $c_{med}=0.314$ [kWh/day] and for the second case was $c_{med}=0.318$ [kWh/day]. Depending on these values, the measured points were divided into two classes: high energy consumption, class zero, and low energy consumption, one class.

![Figure 1](image1.png)

**Figure 1.** Variation of specific energy consumption over time [kWh/m³*day] for COD and TSS training dataset

![Figure 2](image2.png)

**Figure 2.** Variation of specific energy consumption over time [kWh/m³*day] for BOD training dataset
The next step was to realize the logistic regression model for the three situations. The results are presented in figures 3, 4 and 5. The values symbolized by the yellow circle fall into the class zero representing high energy consumption. The values represented by the plus sign fall into class one representing low energy consumption. The blue line represents the decision boundary between the two classes obtained using the regression model.

**Figure 3.** Logistic regression for BOD training dataset

**Figure 4.** Logistic regression for TSS training dataset
In the case of the mathematical model applied for the case of BOD concentration, the accuracy of the model was obtained by 80%. For the cases of TSS and COD training dataset, the accuracy of the model was obtained 84% and 87%, respectively.

Table 3. The probability of tested data set for TSS, COD and BOD concentration

| Nr. crt. | Flow [m$^3$] | TSS [mg/l] | COD [mg/l] | BOD [mg/l] | Probability |
|----------|--------------|------------|------------|------------|-------------|
|          |              |            | COD        | BOD        | TSS | COD | BOD |
| 1        | 9235         | 93.2       | 142        | 82.39      | 0.98 | 0.98 | 0.88 |
| 2        | 8975         | 96.8       | 150        | 86.98      | 0.98 | 0.97 | 0.87 |
| 3        | 8645         | 99.2       | 142        | 82.3       | 0.98 | 0.97 | 0.88 |
| 4        | 8370         | 60.8       | 116        | 67.32      | 0.99 | 0.98 | 0.9  |
| 5        | 7752         | 86         | 134        | 77.97      | 0.98 | 0.98 | 0.88 |
| 6        | 9769         | 113        | 158        | 91.94      | 0.97 | 0.97 | 0.87 |
| 7        | 14803        | 84.4       | 136        | 78.71      | 0.98 | 0.99 | 0.92 |
| 8        | 10014        | 115        | 162        | 93.9       | 0.97 | 0.97 | 0.86 |
| 9        | 21592        | 330        | 386        | 224        | 0.47 | 0.42 | 0.52 |
| 10       | 16605        | 96.8       | 154        | 89.19      | 0.98 | 0.99 | 0.91 |
| 11       | 11134        | 66.8       | 124        | 71.72      | 0.99 | 0.99 | 0.91 |
| 12       | 14636        | 168        | 194        | 112        | 0.94 | 0.96 | 0.86 |
| 13       | 13455        | 215        | 230        | 133        | 0.86 | 0.91 | 0.79 |
| 14       | 18786        | 541        | 480        | 279        | 0.01 | 0.05 | 0.24 |
| 15       | 17492        | 403        | 457        | 265        | 0.16 | 0.07 | 0.38 |
The last step of the research was to apply the mathematical model for the 15 selected dataset for which flow rate, BOD, COD, TSS concentration and energy consumption are known. The selected data are different from those used in model. It was analyzed the probability that the dataset will be found in class one, to have a low cost of energy consumption. The results of the tested dataset are presented in table 3.

Analyzing the results from case 15 with case 10, the flow rate is approximately equal and the BOD concentration is four times higher and the TSS and COD concentrations are three times higher. The probability of having a low cost in case 15 is on average 10% and in case 10 is 90%. If we analyze case 10 with case 6 where the flow is much lower we observe that the probability decreases because the concentrations are much higher in case 6. In conclusion, the energy cost depends primarily on the values of water concentrations. Comparing the real values with the results obtained from the mathematical model, 10 of them were validated.

4. Conclusions
The mathematical model of logistic regression was developed in the Octave program. The analysis was performed for 403 training dataset and 15 testing set for a wastewater treatment plant operating with SBR technology. For the model, the influent flow, BOD, TSS, COD concentration and energy consumption were taken into account. The resulting model has an accuracy of 80%, 84% and 87%, respectively. This shows that a large volume of data is needed for better accuracy. Regarding the tested dataset, they provided a good result in proportion of 70%.

Research will continue to improve the model and increase its efficiency. For this, will be increase in the number of training dataset and more water quality parameters will be taken into account.

The final goal is to obtain a model based on machine learning to optimize energy consumption in a wastewater treatment plant.

Acknowledgements
The work has been funded by the Operational Programme Human Capital of the Ministry of European Funds through the Financial Agreement 51675/09.07.2019, SMIS code 125125.

References
[1] H. Lu, X. Ma, Chemosphere, 249, 126-169 (2020);
[2] H. Liu, Y. Zhang, H. Zhang, Process Biochem, 97, 72-79 (2020);
[3] A. Sharafati, S.B.H.S. Asadollah, M. Hosseinzadehd, Process Saf. Environ. Prot., 140, 68-78, (2020);
[4] F. Harroua, D. Abdelkader, S. Ying, S. Mohamed, J. Environ. Manage., 223, 807-814, (2018);
[5] D. Messaoud, L. Imed, A. Hellal, IWA, (2018);
[6] N. Khatri, K. K. Khatri, A. Sharma, J. Water Process. Eng., 37, 101477 (2020);
[7] K. Lotfi, H. Bonakdari, I. Ebtihaj, F. S. Mjallic, M. Zeynoddin, R. Delatolla, B. Gharabaghi, J. Environ. Manage., 240, 463 - 474, (2019);
[8] S. Jouanneau, E. Grangé, M. J. Durand, G. Thouand, Water Research, 166, 115079, (2019)
[9] S. Shi, G. Xu, Chem. Eng. J., 347, 280-290, (2018);
[10] S.P. Lee, S.Y. Min, J.S. Kim, J.U. Park, M.S. Kim, Enviro. Eng. Res., 19, 31-33,(2014);
[11] N. Deepnarain, S. Kumari, J. Ramjith; F. M. Swalaha, V. Tandoi, K. Pillay, F. Bux, Water Sci. Technol., 72(3), 391–405, (2015);
[12] K. Chmielowski, D. Bedla, E. Dacewicz , L. Jurik, Pol. J. Environ. Stud. 29, 1101-1110 (2020),
[13] Ç, Oluklulu A research on the photovoltaic modules that are being used actively in utilizing solar energy, sizing of the modules and architectural using means of the modules. MSc, Gazi University, Ankara, Turkey, (2001);
[14] L. Pérez-Lombard, J. Ortiz, C. Pout, Energy Build., 40(3), 394-398. (2008);
[15] O Gustaf. Water and Energy: Threats and Opportunities, IWA, (2012);
[16] Torregrossa D., Leopold U, Hermández-Sancho F., Hansen J. J. Environ. Manage., 223, 1061–1067, (2018);
[17] Mosavi A., Bahmani A. Energy consumption prediction using machine learning; a review, www.preprints.org, (2019);
[18] H. Guo, K. Jeong, J Lim, J. Jo, K. Young Mo, P. Jong-pyo, K. Joon Ha, C. Kyung Hwa, Journal of Environ Sci., 32, 90 – 101. (2015);
[19] Y. Yu, Z. Zou, S. Wang, J. Environ. Sci., 75, 201-208 (2019);
[20] X. Yang, J. Wei, G. Ye, Y. Zhao, Z. Li, G. Qiu, F. Li, C. Wei, Sci. Total Environ., 714, 136655, (2020);
[21] H. W. Chen, N. B. Chang, J. Environ. Manage, 65(4), 383–409, (2002);
[22] E. Friedler, E. Pisanty.. Water Res., 40(20), 3751–3758, (2006);
[23] G. Iordache, M. Dordescu, A. Constantin, L. Rosu, V. Cusnerenco, J. OUACSCE, 1 (12), 367, (2010);
[24] F. H. Sancho, M. M. Senante, R. S. Garrido. Sci. Total Environ., 409(14), 2693–2699, (2011).