XV International Conference on Durability of Building Materials and Components
DBMC 2020, Barcelona
C. Serrat, J.R. Casas and V. Gibert (Eds)

Use of Machine Learning in the Function of Sustainability of Wastewater Treatment Plants

Goran Volf

Faculty for Civil Engineering, University of Rijeka, Radmile Matejčić 3, 51000 Rijeka, Croatia, goran.volf@uniri.hr

Abstract. *Wastewater treatment plants (WWTP) are complex and dynamic systems whose management and sustainability can be improved by using different modelling and prediction approaches of their work. A machine learning tool for development of model trees was used in this paper in order to develop a model for chemical oxygen demand (COD) in the wastewater effluent from the WWTP with activated sludge to increase its sustainability and helps in its management purposes. Measured data, both in influent and effluent of the WWTP were used for modelling. For the COD model, machine learning tool Weka and algorithm for development of model trees M5P were used. Obtained model has a high descriptive power and correlation coefficient and thus can be used for prediction and modelling purposes, which can help in management and sustainability of the WWTP. Also, the purpose of this paper is to show the benefits of using machine learning tools for developing WWTP models.*

Keywords: *Wastewater Treatment Plant, Machine Learning, Model Trees, Function of Sustainability, Management.*

1 Introduction

There is a need for wastewater treatment plants (WWTP) to adapt to a rise in water and energy demands, prolonged periods of droughts, climate variability, and resource scarcity (Cornejo 2015).

As population increases, minimizing the carbon and energy footprints of wastewater treatment, while properly managing nutrients (mainly nitrogen and phosphorous) is crucial to improving the sustainability of WWTP. Integrated resource recovery can also mitigate the environmental impact of wastewater treatment systems, however, mitigation potentially depends on various factors such as: a) treatment technology, b) resource recovery strategy, and c) size of system (Cornejo 2015).

Today, the biological treatment of wastewater with activated sludge is one of the most widely used technological processes in WWTP, because of its capabilities, economy and efficiency. The two main components of this process are aeration basin, e.g. the bioreactor and the secondary clarifier (Henze *et al.*, 2002).

Biological treatment of wastewater consists of complex physical, chemical and biological processes through which organic matter, nitrogen and phosphorus are removed from the wastewater. Successful wastewater treatment requires appropriate concentrations and conditions for the growth of microorganisms that must be achieved in the bioreactor (Henze *et al.*, 2002).

Due to the complexity and sensitivity of the treatment process, it is difficult to continuously maintain optimal operating conditions within the WWTP. Because of this,

modelling becomes very useful tool that is often used to simulate and control the operation of the WWTP. Mathematical models (e.g, Activated Sludge Models (ASM), IWA task group, 2000) are commonly used to model WWTP. In this paper, machine learning tool Weka (Witten and Frank, 2000), e.g. the algorithm M5P for induction of model trees, is used to model the WWTP e.g. chemical oxygen demand (COD) concentration in the effluent of the WWTP.

Today, various tools and methods are used to model WWTP, such as, for example, statistical models (Čurlin *et al*., 2008; Dürrenmatt and Gujer 2011; Razifa *et al*., 2014), expert systems (Dürrenmatt and Gujer 2011; Baeza *et al*., 1999; Roda *et al*., 1999), knowledge-based approaches (Comas *et al*., 2003), neural networks (Dürrenmatt and Gujer 2011; Belanche *et al*., 1999; Zhao *et al*., 1999; Hong *et al*., 2015; Mjalli *et al*., 2007), hybrid approaches (Sànches-Marré *et al*., 1996; Grieu *et al*., 2005; Picioreanu *et al*., 2004; Picioreanu *et al*., 2003) and various types of machine learning (Dürrenmatt and Gujer 2011; Hong *et al*., 2015; Comas *et al*., 2001; Atanasova and Kompare 2002; Kompare *et al*., 2006; Manu *et al*., 2017).

The purpose of the model obtained with the use of machine learning tools in this paper is to model the concentration of COD in the effluent of the WWTP. The value of the COD in the effluent is considered as the best indicator for operation quality for the WWTP, i.e. residual organic loads, which also indicated the efficiency of the treatment process (Čurlin *et al*., 2008; Henze *et al*., 2002; Tchobanoglous *et al*., 2003). For this reason, the variable specified is defined as the observed parameter, which best indicates the state of the process in the WWTP, and also, the variable whose dynamics the machine learning tool wants to explain and predict. Using the given model, it would be possible to predict the values of the COD concentration, and if it is greater than the limit value prescribed by Ordinance on emission of limit values for wastewaters (2016), a rapid response would be possible which would then ensure its reduction to an acceptable level (Čurlin *et al*., 2008).

Except the modelling the COD concentration in the effluent from the WWTP with machine learning, the aim of this paper is to demonstrate also some of the advantages and capabilities of machine learning tools.

2 Materials and Methods

2.1 Description of the WWTP

Modeled WWTP is a second stage WWTP, with size of 9,500 population equivalent (PE). The WWTP on the water line consists of mechanical pretreatment and second stage treatment of wastewater (Figure 1). Mechanical pretreatment consists of a coarse, fine screens and aerated grit chamber. The second stage of treatment consists of selectors, aeration basins and secondary clarifiers.

The flow of water on the WWTP is: first, the wastewater enters a channel in which are located coarse and fine screens, after which the wastewater goes to the aerated grit chamber. The water is then transported to the selector, where the selection of microorganisms (contact of biomass with wastewater) takes place. After selector water is then transported to aeration basins (bioreactors), where biological treatment with activated sludge takes place. Finally, the mixture of water and activated sludge is transported to secondary clarifiers, where the activated sludge floccules are deposited and treated water is discharged into the recipient (sea)

through a submarine outlet. Part of the deposited sludge from secondary clarifiers is returned back to the selector or aeration basin in order to maintain the required concentration of activated sludge for successful biological treatment of the process in aeration basins.

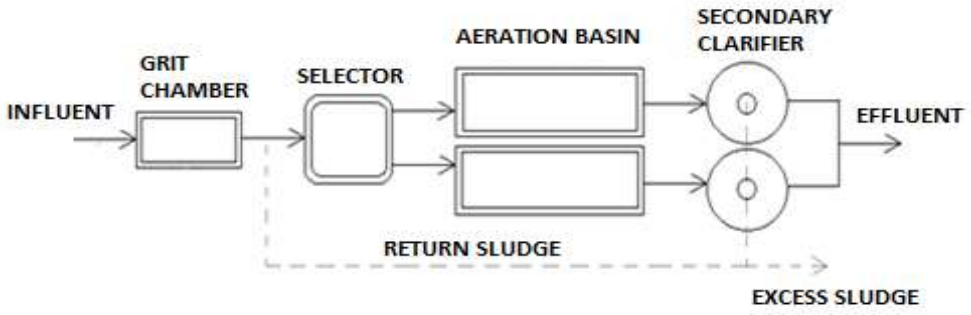

Figure 1. Water line for the WWTP.

2.2 Database

The data used for the modelling (see Table 1) were measured in influent and effluent of the WWTP. The data are presented as mean values through one day, that is, one record in the database represents the one-day situation of the WWTP operation. The database consists of total 718 situations (days).

To supplement the missing data in the total data set, the cubic spline method of interpolation between the measured values was used.

Table 1. Measured data at WWTP uded for modeling.

Data	Description	Unit
Q_{in}	Influent flow	m^3/s
Q_{out}	Effluent flow	m^3/s
T_{out}	Effluent temperature	^0C
COD_{in}	Influent Chemical Oxygen Demand	mg/l
COD_{out}	EffluentChemical Oxygen Demand	mg/l
NH_4-N_{in}	Influent Ammonium	mg/l

3 Machine Learning; Regression and Model Trees

While the simple linear regression calculates one equation (one weighing vector) for the entire data set, piecewise or tree-structured regression divides the data set into several subsets on which uniform class value or linear equation can be applied. The division to subsets is based on tests of the values of the input attributes which are put as nodes in a regression or model tree.

Thus, regression trees are hierarchical structures composed of nodes and branches, where the internal nodes contain tests on the input attributes. Each branch of an internal test corresponds to an outcome of the test and the predictions for the values of the target variable

(the class) are stored in the leaves which are the terminal nodes in the tree. If the leafs contain a single value for the class prediction, then we are talking about simple regression trees, while if a linear equation is used for prediction in the leaf, we are talking of model trees (Quinlan, 1992, Witten and Frank, 2000). Figure 2 illustrates the procedure of constructing regression and model trees.

For the experiment, a variation of the M5 algorithm was used, called M5P, implemented in the software package WEKA (Witten and Frank, 2000).

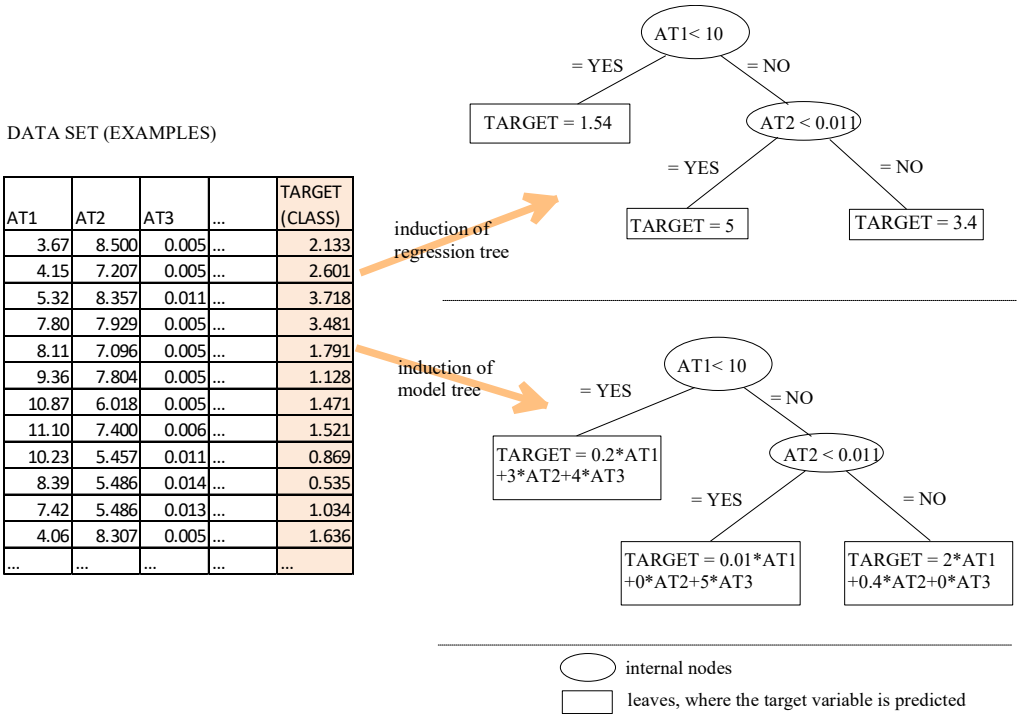

Figure 2. Induction of regression and model trees from given data set (examples).

4 Experiment Setup

The experiment was designed to produce a model of the COD concentration in the effluent of the WWTP. The COD concentration in the effluent was therefore set as the dependent variable, while the flow (Q), COD concentration, ammonium (NH_4-N) and total suspended solids (SS) in the influent of the WWTP were set as independent variables.

5 Results and Discussion

The purpose of the model obtained with machine learning tools, in this case model trees, is to predict the change in COD concentration in the effluent of the WWTP, using the measured variables in the influent of the WWTP. From the given data set (Table 1), a model of the COD concentration in the effluent of the WWTP was created (Figure 3). The model consists of a total 6 nodes and 7 leaves, which contains the values of the variables measured in the influent of the WWTP (see Table 1). Each leaf contains one equation to calculate the COD concentration in the effluent of the WWTP, depending on the structure of the tree itself. The

equations in the individual tree leaves are shown in Table 2. The correlation coefficient R for the obtained model using cross validation method is 0.64.

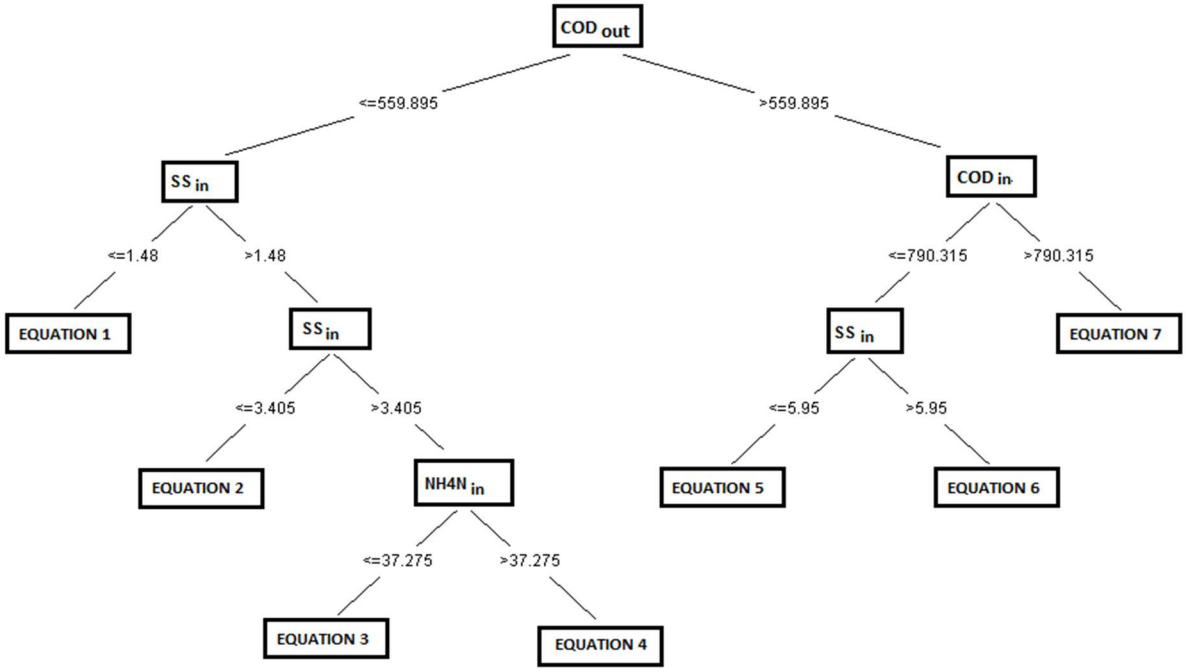

Figure 3. Obtained model for COD concentration.

Table 2. Model equations for COD concentration.

Equation number	Equation
1	COD_{out}=-0,0088*Q_{in} + 0,0954*COD_{in}-1,0529*NH_4N_{in}-0,0215*SS_{in}+45,669
2	COD_{out}=-0,0049*Q_{in}+0,038*COD_{in}-0,1132*NH_4N_{in}-0,0172*SS_{in}+20,0079
3	COD_{out}=-0,0006*Q_{in}+0,0016*COD_{in}-0,0016*NH_4N_{in}-0,0153*SS_{in}+20,1564
4	COD_{out}=-0,0006*Q_{in}+0,0016*COD_{in}-0,0034*NH_4N_{in}-0,0153*SS_{in}+23,1117
5	COD_{out}=-0,0009*Q_{in} +0,0046*COD_{in}+0,1559*NH_4N_{in}-0,0957*SS_{in}+24,0748
6	COD_{out} =-0,0009*Q_{in}+0,0046*COD_{in}+0,1559*NH_4N_{in}-0,0818*SS_{in}+17,0162
7	COD_{out}=-0,0009*Q_{in}+0,005*COD_{in}+1,3378*NH_4N_{in}-0,3479*SS_{in}-7,1299

To predict the COD concentration in the effluent of the WWTP, it is necessary to select from the model shown in Figure 3 the appropriate linear equation depending on the values of the individual attributes in the tree nodes. From Figure 3 can be seen that the COD concentration in the effluent depends mostly on the COD concentration at the inflow (initial

node), then SS and NH_4-N, while the flow (Q) do not appear in the tree at all, but only in single leaves equations. Interpreting the model results, the larger COD values in the effluent are given by the right side of the model tree, that is, the tree shown in Figure 3, and the smaller values by the left side of the model tree. Therefore, lower values of COD concentration in the effluent are associated with nodes in which SS and NH_4N are located, while the higher values are associated with nodes in which COD and SS are located.

A comparison of the time series of measured and modeled COD concentration values can be seen in Figure 5. From Figure 5, can also be seen a good adaptation of the measured and modeled COD concentration values, and also by visual inspection it can be concluded that the peak points values are satisfactorily matched.

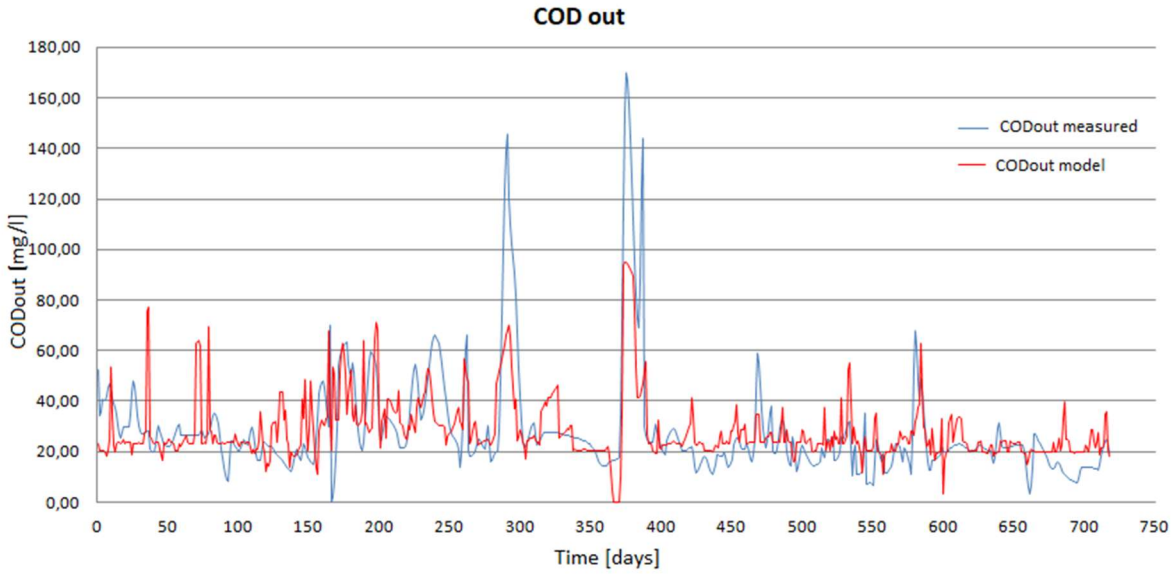

Figure 4. Time series comparison of measured and modeled values of COD concentrations in the effluent.

The results of the experiment show that it is useful to use different approaches when modeling WWTP. As for any method of modeling which use measured data, it is essential that the database consists of sufficiently different situations from which, in this case a machine learning algorithm can learn to predict a dependent variable. Also, for better model results, it would be useful to have more measured parameters in the influent of the WWTP affecting the dependent variable, such as water temperature, pH, chloride concentration (if it is a WWTP in the coastal area such as discussed in this paper), sludge age, sludge return, dissolved oxygen concentration in the aeration bioreactor, food to mass ratio (F/M), etc., which over a longer period can have a greater and significant impact on the COD concentrations (Henze *et al*., 2002). Therefore, as the database used in this paper contains relatively few input parameters, a more accurate prediction of the COD concentrations from the WWTP cannot be expected. However, the resulting model behaves as expected and produces satisfactory results.

7 Conclusions

Use of machine learning tools to create model from database, in this case model trees have been successfully applied to model WWTP, that is, the COD concentration in the effluent of the WWTP which can help in management and sustainability of the modeled WWTP. Obtained model is simple, understandable, and relatively accurate in predicting COD concentrations in the effluent of the WWTP. Before starting modelling procedure, it is important to note that the database contains the actual attribute values and that it has information about the time when the data was collected so that the dynamics of the system being modeled can be incorporated into the model.

Therefore, some of the advantages of using machine learning tools in modeling can be marked, firstly, the construction of descriptive, or white box models, which make it much easier to interpret the obtained models. Models are more comprehensible, thus providing insight into their functioning, that is, the functioning of the modeled system.

Also, it is especially important to emphasize the use of machine learning tools for simpler and more efficient management and sustainability of the WWTP, as shown in this paper.

Future work is recommended to focus on increasing the database so that model accuracy can be increased and other parameters such as nutrients, e.g. nitrogen and phosphorus can be modeled with the enlarged database, where properly managing of nutrients is crucial to improving the sustainability of the WWTP. Thus, new links and patterns among the data could be revealed.

Acknowledgements

This work has been supported by the University of Rijeka under the projects number 17.06.2.1.02 (River-Sea Interaction in the Context of Climate Change) and uniri-tehnic-18-129 5570 (Implementation of innovative methodologies, approaches and tools for sustainable river basin management). Also, this work is part of the project Influence of summer fire on soil and water quality founded by the Croatian Science Foundation.

ORCID

Goran Volf: https://orcid.org/0000-0002-7058-9012

References

Atanasova, N. and Kompare, B. (2002). Uporaba odločitvenih dreves pri modeliranju čistilne naprave za odpadno vodo. *Acta hydrotechnica*, 20, 33, 351-370.

Baeza, J., Gabriel, D. and Lafuente, J. (1999). An expert supervisory system for a pilot WWTP. *Environmental Modelling and Software,* 14, 383-390.

Belanche, L.I., Valdes, J.J., Comas, J., Roda, I.R. and Poch, M. (1999). Towards a Model of Input-Output Behaviour of Wastewater Treatment Plants using Soft Computing Techniques, *Environmental Modelling and Software,* 14, 409-419.

Cornejo, P.K. (2015). Environmental sustainabilita of wastewater treatment plants integrated with resource recovery: the impact of context and scale, *Graduate theses and disertations*, University of South Florida.

Comas, J., Dzeroski, S., Gibert, K., Roda, I.R. and Sanchez-Marre, M. (2001). Knowledge discovery by means of inductive methods in wastewater treatment data. *AI Communication*, 14, 45-62.

Comas, J., Rodríguez-Roda, I.R., Sànches-Marré, M., Cortés, U., Freixó, A., Arráez, J. and Poch, M. (2003). A knowledge-based approach to the defloculation problem: integrating on-line, off-line, and heuristic information, *Water Research,* 37, 2377-2387.

Čurlin, M., Bevetek, A., Ležajić, Z., Deverić-Meštrović, B. and Kurtanjek, Ž. (2008). Modeliranje procesa biološke obrade otpadne vode na komunalnom uređaju grada Velika Gorica. *Kemija u industriji,* 57, 2, 59-67.

Dürrenmatt, D.J. and Gujer, W. (2011). Data-driven modeling approaches to support wastewater treatment plant operation, *Environmental Modelling & Software*, 30, 47-56.

Grieu, S., Traoré, A., Polit, M. and Colprim, J. (2005). Prediction of parameters characterising the state of a pollution removal biologic process. *Engineering Applications of Artificial Intelligence*, 18, 559-573.

Henze, M., Herremoes, P., Jansen, JC. and Arvin E. (2002). *Wastewater Treatment-Biological and Chemical Processes*, Third edition. Springer. New York. US.

Hong, G., Kwanho, J., Jiyeon, L., Young M.K., Jong-pyo, P., Joon, H.K. and Kyung, H.C. (2015). Prediction of effluent concentration in a wastewater treatment plant using machine learning models, *Journal of Environmental Sciences*, 32, 90-101.

IWA Task group on mathematical modelling for design and operation of biological wastewater treatment, *Activated sludge models ASM1, ASM2, ASM2d and ASM3.* (2000). IWA Publishing. London. UK.

Kompare, B., Levstek, M. and Atanasova, N. (2006). Two approaches to wastewater treatment plant modelling. *Acta hydrotechnica,* 24, 40, 45-64.

Manu, D.S. and Thalla, A.K. (2017). Artificial intelligence models for predicting the performance of biological wastewater treatment plant in the removal of Kjeldahl Nitrogen from wastewater. *Applied Water Science*. doi 10.1007/s13201-017-0526-4.

Mjalli, F.S., Al-Asheh, S. and Alfadala, H.E. (2007). Use of artificial neural network black-box modeling for the prediction of wastewater treatment plants performance, *Journal of Environmental Management*, 83, 3, 329-338.

Ordinance on emission of limit values for wastewaters. Narodne novine br. 153/09, 63/11, 130/11 i 56/13, 2013, 80/13, 43/14, 27/15, 3/2016.

Picioreanu, C. and van Loosdrecht, M.C.M. (2003). *Use of mathematical modeling to study biofilm development and morphology.* IWA Publishing, University of Manchester (UK).

Picioreanu, C., Kreft J.U. and van Loosdrecht, M.C.M. (2004). Particle-based multidimensional multispecies biofilm model. *Microbiology,* 70, 5, 3024-3040.

Quinlan, J.R. (1992). *Learning with continuous classes.* Proc. AI'92, 5th Australian Joint Conference on Artificial Intelligence, Singapore. In: Adams & Sterling (Editors). World Scientific, 343-348.

Razifa, M., Bagyo Yanuwiadib, S., Rachmansyahb, A. and Belgiawanc, P.F. (2014). Implementation of Regression Linear Method to predict WWTP cost for EIA: case study of ten malls in Surabaya City, *Procedia Environmental Sciences*, 28, 158-165.

Roda, I.R., Comas, J., Sànches-Marré, M., Cortés, U., Lafuente, J. and Poch, M. (1999). Expert system development for a real wastewater treatment plant. *Chemical Industry and Environment III*, Proceedings. Kraków, Poland, 653-660.

Sànches-Marré, M., Cortés, U., Lafuente, J., Roda, I.R. and Poch, M. (1996). DAI-DEPUR: a distributed architecture for wastewater treatment plants supervision. *Artificial Intelligence in Engineering,* 10, 3, 379-423.

Tchobanoglous, G., Burton, FL. and Stensel, HD. (2003). *Wastewater Engineering-Treatment and Reuse.* Fourth edition. McGraw-Hill.

Witten, I.H. and Frank, E. (2000). *Data mining-Practical machine learning Tools and Techniques with Java implementations*. Academic Press.

Zhao, H., Hao, O.J. and McAvoy, T.J. (1999). Approaches to modeling nutrient dynamics: ASM2, simplified model and neural nets. *Water Science and Technology,* 39, 1, 227-234.