Prediction of wastewater quality indicators at the inflow to the wastewater treatment plant using data mining methods

Bartosz Szeląg[1,*], *Krzysztof* Barbusiński[2], *Jan* Studziński[3], *Lidia* Bartkiewicz[1]

[1]Kielce University of Technology, Faculty of Environmental, Geomatic and Energy Engineering, Tysiąclecia Państwa Polskiego Av. 7, Kielce, Poland
[2]Silesian University of Technology, Institute of Water and Wastewater Engineering, Konarskiego Street 18, Gliwice, Poland
[3]Systems Research Institute Polish Academy of Science, Newelska Street 6, Warszawa, Poland

Abstract. In the study, models developed using data mining methods are proposed for predicting wastewater quality indicators: biochemical and chemical oxygen demand, total suspended solids, total nitrogen and total phosphorus at the inflow to wastewater treatment plant (WWTP). The models are based on values measured in previous time steps and daily wastewater inflows. Also, independent prediction systems that can be used in case of monitoring devices malfunction are provided. Models of wastewater quality indicators were developed using MARS (multivariate adaptive regression spline) method, artificial neural networks (ANN) of the multilayer perceptron type combined with the classification model (SOM) and cascade neural networks (CNN). The lowest values of absolute and relative errors were obtained using ANN+SOM, whereas the MARS method produced the highest error values. It was shown that for the analysed WWTP it is possible to obtain continuous prediction of selected wastewater quality indicators using the two developed independent prediction systems. Such models can ensure reliable WWTP work when wastewater quality monitoring systems become inoperable, or are under maintenance.

1 Introduction

The operation of the wastewater treatment plant is a complex process which is meant to ensure effective and relatively stable purification of wastewater. That can prove problematic due to a high variation in the quantity and quality of raw wastewater conveyed by the sewer system to the treatment facility. To maintain adequate level of pollutant reduction in wastewater, it is necessary to develop tools for predicting both the quantity and quality of influent wastewater. That will make it possible to take steps to correct, in advance, the technological parameters in the treatment facilities, mainly in bioreactors.

* Corresponding author: bszelag@tu.kielce.pl

In modern wastewater treatment plants (WWTPs), the values of indicators that describe wastewater quality are measured with on-line systems according to a pre-set time step, or determined in laboratory tests. In both cases, however, problems may arise with obtaining time series of measurement data at a constant resolution. In practice, that leads to a situation in which the data providing an input to a mathematical model for the optimisation of the operation of WWTP individual components are not available. In order to ensure effective performance of WWTPs, it is advisable to develop autonomous systems for the prediction of influent wastewater quantity and quality.

Key indicators of wastewater quality, which at the same time provide a basis of WWTP efficiency assessment, include the following: biochemical oxygen demand (BOD_5), chemical oxygen demand (COD), total suspended solids (TSS), total nitrogen (TN) and total phosphorus (TP). To predict wastewater quantity and quality within separate and combined sewage systems, physical mathematical models, based on the differential equation systems were developed [1]. However, due to the fact that it is necessary to collect detailed information on the catchment and meteorological conditions, the problems with the model calibration may arise. Therefore, to model the quantity and quality of wastewater flowing from the catchment to the WWTP, data mining methods are used. On the basis of the literature data [2,3], it can be stated that to simulate total suspended solids (TSS), total nitrogen (TN) and total phosphorus (TP) in influent wastewater ARIMA-type models and Artificial Neural Networks (ANN) can be employed. To predict TSS, Verma et al. [4] used a number of models, such as Support Vector Machines (SVM), Random Forests (RF), Multivariate Adaptive Regression Splines (MARS) etc. The differences between simulated and measured values turned out to be small. Additionally, the analyses performed by Minsoo et al. [5] confirmed the applicability of the k-NN method for the prediction of quality indicators of influent wastewater. In addition to the above-mentioned methods for WWTP operation modelling hybrid models [6] are used. The originate from the combination of two statistical models (the classification and the regression ones, or a few regression-type models).

This study presents a proposal of the modelling of the values of selected quality indicators (BOD_5, COD, TSS, TN, and TP) of influent wastewater with the use of three data mining methods. The modelling is based on the results of last measurements that preceded the predicted quantity, and also on the measured values of daily inflow to the facility. The model could be applied when the systems measuring pollutant indicators become inoperable. In the analyses, 3-year daily time series of flow rate and wastewater quality indicators measured at the inflow to the wastewater treatment plant located in the city of Rzeszów were used.

2 The object of investigation

Rzeszów, an urban-industrial agglomeration, lies in the south-eastern Poland. The city has mainly the separate sewerage system. Municipal wastewater from the city and adjacent localities is conveyed to the mechanical-biological wastewater treatment plant with the design capacity of 62,500 m^3/d. Continuous monitoring conducted since 2013 provides measurements of parameters describing quantity and quality of influent wastewater and also of effluent at a daily resolution.

3 Methodology

In the study, simulations were conducted for two instances. In the first case, the possibility of predicting wastewater quality indicators (BOD_5, COD, TSS, TN and TP) was

considered, based on the results of last measurements of those indicators. In the other case, the possibility of predicting the values of the above-mentioned indicators on the basis of the flow rate measurements at the inflow to the plant was analysed. The advantage offered by this solution is that when the on-line measurement system fails, an alternative option makes it possible to predict selected quality indicators of influent wastewater on the basis of measured flows. Under service conditions of the plant, that allows maintaining continuity in predicting influent wastewater quality indicators, and thus ensures a stable operation of bioreactors by adequate selection of the technological parameters (e.g. oxygen concentration, TSS, the amount of PIX dosage, recirculation ratio, and others). The concept of the system described above is shown in Fig. 1.

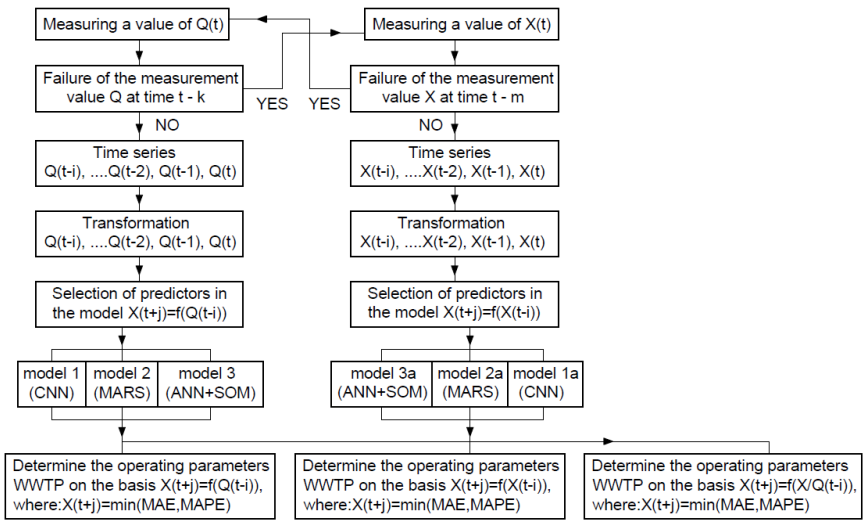

Fig. 1. Diagram of the autonomous measurement and prediction system for influent wastewater quality indicators intended to selected the settings of technological facilities

where: i – time shift between the predicted value and the previous values that describe the predicted value in the model for the prediction of a selected wastewater quality indicator; j – forward time step (in the study it was assumed j = 1) with which a quality indicator of concern is predicted; k, m – failure period in which the results of daily measurements of Q, X are not available.

Based on the concept shown above, statistical models were developed to predict wastewater quality indicators using the MARS method and cascade neural networks (CNN). Additionally, a hybrid model was designed, which is combination of neural network of the multilayer perceptron type and Kohonen self-organizing network. To ensure a correct training process and assessment of the mathematical model performance, 5-fold cross-validation was used. That was preceded by the dataset partitioning into the training set (75%), and the testing and validating set (25%). Prior to the development of statistical models, input and output data were standardised by means of the min-max transformation [7].

The MARS method is one of numerous tools used for solving regression problems [8]. In the classical approach, independent variables are treated uniformly, whereas in the MARS method, variation ranges of the predictors of concern are divided into subranges in which independent variables can have different impacts on the phenomenon investigated. The boundaries of subranges are determined on the basis of threshold values (t). The values of the predictors of concern are segregated into those lower and higher than threshold values t_i. That is performed by means of the basis function which has the following form:

$$h(\boldsymbol{X}) = \alpha_i \cdot (max(0, X - t)) \quad (2)$$

where: h(**X**) – vector of the basis functions for individual variables (x_i) for which the following condition is satisfied:

$$x_i - t_i = \begin{cases} x_i - t_i; for\ \ x_i > t_i \\ 0; for\ x_i \leq t_i \end{cases} \quad (3)$$

To calculate the values (α_i) of the model parameters, Friedman [8] developed a special algorithm that allows the search of observations to establish threshold values.

Multilayer unidirectional network known as multilayer perceptron [2,9,10] is one of the most widely used structures. It has found multiple applications to the modelling of the quantity and quality of influent wastewater and the operation of technological facilities. One of the modifications of multilayer perceptron–type ANN is a cascade network (CNN). It contains additional weighted connections from each layer input, and from each layer to a successive one. It should be noted that so far CNNs have solely been used to model the performance of wastewater treatment plants. They have never been applied to predict influent wastewater quality. In addition to classic neural networks with adequately modified architecture [6,11], hybrid models have been developed that result from the combination of the classification and regression models. Multilayer perceptron–type ANN provides an example of hybrid models. The most popular type of the network termed as self-organizing map (SOM) is Kohonen network [11]. Due to the algorithm used in a majority of SOMs, neurons representing similar classes are close in proximity to each other, thus producing an organised map. Consequently, it is possible to identify relations between the classes obtained. At the SOM training stage, the number of neurons making the topological layer and the number of training epochs are defined.

In this study, cascade neural network models were developed. In the models, three connections were assumed to join inputs to successive layers of the CNN [12,13]. In individual hidden layers, various activation functions were considered, namely hyperbolic tangent, linear, sine, exponential, logistic, and sigmoidal ones. For the output layer, the linear activation function was adopted. In addition, it was assumed in computations that the number neurons in the successive hidden layers is the same. The optimal number of neurons and values of weights were established for CNN by generating, every time, 500 different models for the adopted independent variables. Out of the trained networks, the one showing the lowest values of the absolute and relative errors in the prediction of wastewater quality indicators was selected. To obtain required generalizing abilities of CNN models, Bayesian regularisation algorithm was employed at the training stage [14]. In the case under consideration, for the hybrid model, Kohonen self-organizing neural networks with the topological layer 2x2 were used as a classifier, whereas the number of epochs during the training stage was 1000.

In the above analyses, to assess the predictive powers of the models predicting BOD_5, COD, TSS, TN, and TP values, the following were used: mean absolute error (MAE), mean relative error (MAPE) and coefficient of correlation (R).

4 Results

Based on the results of measurements of influent wastewater quantity and quality, the ranges of variation were established (Table 1). The data in Table 1 indicate substantial variation in BOD_5, COD, TSS, TN and TP values. Because the indicators of wastewater quality constitute input data for the models describing the kinetics of changes in carbon, nitrogen and phosphorus compounds in the bioreactors, and their values vary to a large

extent, it is necessary to predict those indicators. Out of the methods investigated in this study, only the parameter estimation algorithm developed for the MARS model offers the possibility of discarding those predictors the effect of which on the dependent variable is negligible. Therefore, the simulation, with the use of this method, of selected quality indicators of influent wastewater was performed first. On the basis of independent variables specified using the MARS method, predictions of BOD_5, COD, TSS, TN and TP values were made with the other methods.

Table 1. Range of variation in the values of parameters describing influent wastewater in the years 2013-2015

Parameter	Minimum	Maximum	Average
COD, mg/dm^3	159.0	2510.0	927.1
BOD_5, mg/dm^3	38.1	788.0	374.0
TSS, mg/dm^3	80.0	1140.0	430.0
TN, mg/dm^3	21.3	99.0	69.7
TP, mg/dm^3	3.4	37.5	12.45
Q, m^3/d	26973	66773	38658

The values of wastewater quality parameters obtained from the last measurements were used for that purpose. That was followed by simulations of influent wastewater quality indicators based on flow rate measurements.

The analyses demonstrate that in order to compute COD, TN and TP values, it is advisable to take into account two last measurements (i=1,2) of those indicators, which precede the predicted value. For TSS prediction, it is necessary to account for the values from TSS(t-1) to TSS(t-4), whereas for BOD_5 prediction, BOD_5(t-1) and BOD_5(t-3) values are sufficient. The analyses indicate that to predict COD, TSS, TN and TP, the values Q(t-1) and Q(t-2) could be used. To compute BOD_5, it is necessary to account for the values Q(t-1), Q(t-2) and Q(t-3). The results obtained with the use of the MARS method are confirmed by the investigations conducted by Szeląg et. al.[7] with the classification-tree method. Tables 2-4 show the values of parameters of fit (MAE and MAPE) of computations with MARS, CNN and ANN+SOM methods of selected wastewater quality indicators to the measurement results for the testing set. The number of neurons in individual hidden layers in the models ranged from 3 to 6.

As regards the models based on X(t-i), the best computational results were observed for the hyperbolic tangent function in the hidden layer, whereas in the modelling of the concentration of organic compounds, and of nitrogen and phosphorus, the logistic function produced the lowest values of MAE and MAPE.

In hybrid neural networks, the number of neurons in the hidden layer of the designed statistical models ranged 3-8 in separate classes, and the hyperbolic tangent function was most frequently used as the activation function. The data presented in Tables 2-4 show that better predictive powers of independent variables (lower MAE and MAPE values) are shown by the models developed on the basis of wastewater quality indicators, which is true for both MARS and CNN methods. The statistical models developed using ANN+SOM for computations of BOD_5, TSS and TN, based on Q(t-i), have better predictive abilities with respect to wastewater quality indicators.

Taking into account the time necessary for BOD_5 determination and the simulation results of wastewater quality indicators, it can be stated that autoregression models that take into account BOD_5(t-i) values cannot be used in practical applications. Additionally, error values for models relying on BOD_5(t-i) and Q(t-i) values range as follows: MAE = 30.16-41.12 mg/dm^3, MAPE = 8.75-12.39%, and MAE = 27.93-48.10 mg/dm^3, MAPE = 7.62-14.80%, respectively. In models for COD prediction, based on COD(t-i), and solely on Q(t-i), values of errors in the indicator prediction varied as follows: MAE = 93.29-117,54

mg/dm^3, MAPE = 10.26-14.06% and MAE = 92.50-133.02mg/dm^3, MAPE = 10.93-15.89%,respectively. In the models for TSS prediction that account for TSS(t-i), and only Q(t-i), the magnitude of simulation errors were MAE = 49.05-59.01 mg/dm^3, MAPE = 13.09-15.90% and MAE = 50.89-71.88 mg/dm^3, MAPE = 14.62-20.07%, respectively.

Table 2. Values of parameters of fit (MAE and MAPE) of computational results of wastewater quality indicators obtained with MARS method, determined on the basis of X(t-i) and Q(t-i), to measurement results

Indicators	X(t-i)			Q(t-i)		
	MAE	MAPE	R	MAE	MAPE	R
BOD_5	41.12	12.39	0.59	48.1	14.80	0.55
COD	117.54	14.06	0.56	133.02	15.89	0.52
TSS	59.01	15.9	0.55	71.88	20.07	0.38
TN	6.07	9.64	0.67	7.04	11.43	0.66
TP	1.61	13.97	0.65	2.16	19.03	0.45

Table 3. Values of parameters of fit (MAE and MAPE) of computational results of wastewater quality indicators obtained with CNN method, determined on the basis of X(t-i) and Q(t-i), to measurement results

Indicators	X(t-i)			Q(t-i)		
	MAE	MAPE	R	MAE	MAPE	R
BOD_5	33.07	9.67	0.79	37.29	11.29	0.72
COD	99.56	10.26	0.68	112.5	13.14	0.63
TSS	52	13.89	0.7	60.9	16.61	0.63
TN	5.45	8.53	0.74	5.76	8.96	0.72
TP	1.4	12.19	0.66	1.76	15.54	0.56

Table 4. Values of fit (MAE, MAPE) of computational results of selected wastewater quality indicators obtained with hybrid model (ANN+SOM) to measurement results

Indicators	X(t-i)			Q(t-i)		
	MAE	MAPE	R	MAE	MAPE	R
BOD_5	30.16	8.75	0.74	27.93	7.62	0.82
COD	93.29	11.34	0.73	92.5	10.93	0.72
TSS	49.05	13.09	0.75	50.89	14.62	0.73
TN	5.00	7.83	0.78	4.38	6.42	0.84
TP	1.72	15.11	0.62	1.79	15.75	0.55

As regards to TN prediction models based on TN(t-i), and solely on Q(t-i), the error values were similar, namely MAE = 5.00-6.07 mg/dm^3, MAPE = 7.83-9.64% and MAE = 4.38-7.04 mg/dm^3; MAPE = 6.42-11.42%. In addition, the data in Tables 2-4 show that for TP prediction models that involve TP(t-i) and solely Q(t-i), prediction errors range as follows: MAE = 1.40-1.72 mg/dm^3, MAPE = 12.19-15.11% and MAE = 1.76-2.16 mg/dm^3, MAPE =15.54-19.03%, respectively. Predictions of wastewater quality indicators (BOD_5, COD, TSS, TN and TP) listed in Tables 2-4 show that the lowest prediction errors were observed for the MARS method. The greatest congruence between computational results and measurements of BOD_5, COD, TSS, and TN was achieved using the ANN+SOM method. Only for TP, the lowest values of prediction errors were found when the CNN method was employed. Computations performed using the ANN+SOM method demonstrate that the predicted values of COD, TSS, TN and TP based on Q(t-i) do not differ more than 5% from the results produced with the models in which independent variables were wastewater quality indicators. A similar relation was found for BOD_5 values. Those findings can prove useful in case of the measurement system failure.

The value of the coefficient of correlation for BOD_5 prediction model, based on Q(t-i), developed using ANN+SOM method (Table 4) is lower than that obtained by Abyaneh [3] (R = 0.83) with ANN. To predict BOD_5, Abyaneh [3] relied on temperature, pH of influent wastewater, and TSS. Dogan et al. [15] received a better fit (R = 0.92) of BOD_5 simulation results to measurements than Abyaneh [3]. Additionally, the values of MAPE computed for COD prediction model are greater than those (MAPE = 7.35%) obtained by Minsoo et al [5] with the k-NN method. The values of R computed for TSS prediction model turned out to be lower than the results received by Verma and Kusiak [4] (R = 0.93), who used the ANN method and relied on $CBOD_5$ and flow measurements.

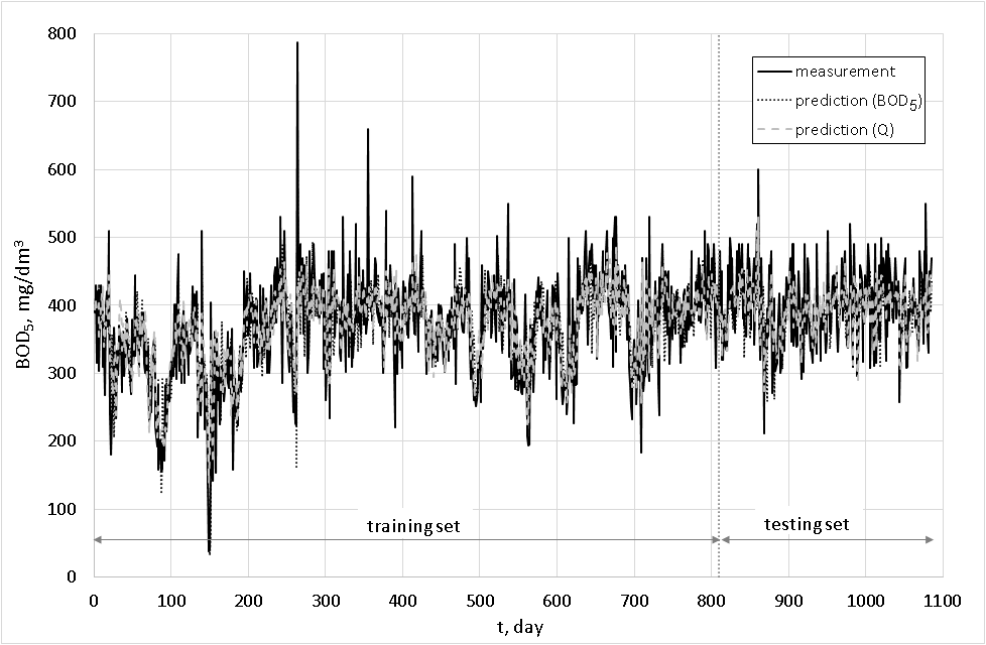

Fig. 1. Comparison of BOD_5 computations with hybrid models and measurements in the period of concern

Values of MAPE in models for the prediction of TN and TP with the ANN+SOM method (Table 4) are greater than those produced with the k-NN method (MAPE = 5.79% and MAPE = 8.87%), obtained by Minsoo et al [5]. The analyses presented above indicate that the parameter which considerably affects the variation in quality indicators is wastewater inflow to WWTP. That is illustrated by the parameters of fit of computational results to measurements obtained in this study, and also by other researchers. At the same time the analyses conducted for this study confirm that it is possible to predict wastewater quality indicators solely on the basis of flow data. The accuracy of predictions is satisfactory, which is illustrated by the values of errors obtained (Tables 2-4). Although the parameters of computations fit to measurements (R, MAE, MAPE) received in this study were worse than those reported by Minsoo et al [5], Dogana et al [15], Verma and Kusiak [4] the models designed in this study can find practical applications. When the number of independent variables in the statistical models is limited to only wastewater inflow data, the models could be employed under failure conditions of the measurement system, or during the system overhaul. Such a scenario was not considered in other studies. As remaining researchers [3,4,5,7] took into account other independent variables in addition wastewater inflow, it is necessary to conduct further research aimed at finding the impact of those variables on the predictive abilities of the models.

5 Conclusions

The study discusses the possibility of modelling the values of selected indicators (BOD_5, COD, TSS, TN, and TP) of wastewater quality at the inflow to the treatment plant on the basis of the data from the last measurements, and also solely on the basis of flow rate values. The computations demonstrated that lower values of errors in the prediction of biochemical and chemical oxygen demand, concentration of nitrogen, phosphorus, and total suspended solids were generated in the statistical model developed using cascade neural networks than with the MARS method. In the majority of cases considered, the best predictive abilities as regards wastewater quality indicators were shown by hybrid models that combine Kohonen neural networks and multilayer perceptron.

Slightly better results of simulations of wastewater quality indicators were obtained when independent variables were the values of wastewater quality indicators rather than the flow values. That refers to models based both on MARS and CNN. For hybrid models, the prediction error values were lower than for MARS and CNN-based models. It was only for TP computations with the use of Q(t-i) that MAE and MAPE values were lower when CNN-based model, rather than ANN+SOM–based one, was used.

It was shown that for the wastewater treatment plant of concern, it is possible to ensure continuity in predictions of the values of selected pollutant indicators using the two independent systems developed in the study. That could ensure reliable operation of the plant facilities when the wastewater quality monitoring system fails or is being upgraded.

References

1. K. V. Gernaey, M.C.M. van Loosdrecht, M. Henze, M. Lind, S.B. Jørgensen, Environ Modell Softw **19**, 9, 763 (2004)
2. A. Dellana, D. West, Environ Modell Softw **24**, 1, 96 (2009)
3. H. Z. Abyaneh, J Environ Health Sci **12**, 40, 2 (2014)
4. A.Verma, X.Wei, A. Kusiak, Eng Appl Artif Intel **26**, 1366 (2013)
5. K.Minsoo, K. Yejin, K. Hyosoo, P. Wenhua, K. Changwon, Front Env Sci Eng **10**, 2, 299 (2016)
6. S. Grieu, F. Thiery, A. Traoré, T. P. Nguyen, M. Barreau, M. Polit, Chemical Engineering Journal **1**, 116, 1 (2006)
7. B. Szeląg, L. Bartkiewicz, J. Studzinski, Ochrona Srodowiska **38**, 4, 39 (2016)
8. W . Zhang, A. T. C. Goh, Geoscience Frontiers **7**, 1, 45 (2016)
9. D Güçlü., Ş. Dursun, Bioproc and Biosyst Eng **33**, 9, 1051 (2010)
10. B. Szeląg, J. Gawdzik, Pol J Environ Stud **25**, 4, 1709 (2016)
11. R. Rustum, A. J. Adeloye, M. Scholz, Water Environ Res **80**, 1, 32 (2008)
12. G.Capizzi, G. L. Sciutto, P. Monforte, C. Napoli, International Journal of Electronics and Telecommunications **61**, 327 (2015)
13. S. B.Al – Batah, M.S Alkhasawneh., L.T Tay., U.K. Ngah, H. H. Lateh, M.T.A Isa, Math Probl Eng **2015**, 2015, 1 (2015)
14. K. Hirschen, M. Schäfer, Computer Methods in Applied Mechanics and Engineering **195**, 7-8, 481 (2006)
15. E. Dogan, A.Ates, E.C. Yilmaz, B. ErenEnvironmental Progress **27**, 4, 439 (2008)