Prediction of the waste stabilization pond performance using linear multiple regression and multi-layer perceptron neural network: a case study of Birjand, Iran

Maryam Khodadadi1, Alireza Mesdaghinia2, Simin Nasser11, Mohammad Taghi Ghaneian3, Mohammad Hassan Ehrampoush4, Mahdi Hadi5*

1Ph.D. Student of Environmental Health, International Campus, Shahid Sadoughi University of Medical Sciences, Yazd, Iran
2Professor, Center for Water Quality Research (CWQR), Institute for Environmental Research (IER), Tehran University of Medical Sciences, Tehran, Iran
3Associate professor, Department of Environmental Health, School of Public Health, Shahid Sadoughi University of Medical Sciences, Yazd, Iran
4Professor, Department of Environmental Health, School of Public Health, Shahid Sadoughi University of Medical Sciences, Yazd, Iran
5Ph.D. student, Center for Water Quality Research (CWQR), Institute for Environmental Research (IER), Tehran University of Medical Sciences, Tehran, Iran

Abstract

Background: Data mining (DM) is an approach used in extracting valuable information from environmental processes. This research depicts a DM approach used in extracting some information from influent and effluent wastewater characteristic data of a waste stabilization pond (WSP) in Birjand, a city in Eastern Iran.

Methods: Multiple regression (MR) and neural network (NN) models were examined using influent characteristics (pH, Biochemical oxygen demand [BOD], temperature, chemical oxygen demand [COD], total suspended solids [TSS], total dissolved solid [TDS], electrical conductivity [EC] and turbidity) as the regression input vectors. Models were adjusted to input attributes, effluent BOD (BODout) and COD (CODout). The models performances were estimated by 10-fold external cross-validation. An internal 5-fold cross-validation was also used for the training data set in NN model. The models were compared using regression error characteristic (REC) plot and other statistical measures such as relative absolute error (RAE).

Results: NN models (with RAE = 78.71 ± 1.16 for BODout and 83.67 ± 1.35 for CODout) and MR models (with RAE = 84.40% ± 1.07 for BODout and 88.07 ± 0.80 for CODout) indicate different performances and the former was better (P<0.05) for the prediction of both effluent BOD and COD parameters. For the prediction of CODout the NN model with hidden layer size (H) = 4 and decay factor = 0.75 ± 0.03 presented the best predictive results. For BODout the H and decay factor were found to be 4 and 0.73 ± 0.03, respectively. TDS was found as the most descriptive influent wastewater characteristics for the prediction of the WSP performance. The REC plots confirmed the NN model performance superiority for both BOD and COD effluent prediction.

Conclusion: Modeling the performance of WSP systems using NN models along with sensitivity analysis can offer better understanding on exploring the most significant parameters for the prediction of system performance. The findings of this study could build the foundation for prospective work on the characterization of WSP operations and optimization of their performances with a view to conducting statistical approaches.

Keywords: Data mining, Multiple regression, Neural network, Waste stabilization pond

Introduction

In recent decades, the application of computer modeling techniques has been introduced in many environmental issues (1). Given the significance of wastewater treatment and its role in reducing environmental pollution, the control and proper operation of a wastewater treatment system is the most important environmental issue. The waste stabilization ponds (WSPs) are commonly recommended wastewater treatment systems used in arid and semi-arid developing countries. WSPs are valuable treatment systems due to suitable climatic conditions and availability of land (2). A stabilization pond is the most simple, reliable and cost-effective process with low maintenance requirements that can be used as an appropriate alternative for wastewater treatment by reducing biological oxygen demand (BOD) (3). Inappropriate operation of a wastewater treatment systems may result in severe environmental and public health problems, as its effluent to a receiving...
water body can cause or spread various diseases to human beings (4). The performance of a WSP is often affected by various factors such as physical and biological factors (5). Moreover, the performance efficiency of a WSP is greatly controlled by raw wastewater characteristics because of the variations in raw wastewater characteristics, its strengths and flow rates which is attributed to the changing and complex nature of the treatment process (6). In a WSP treatment system, there are certain basic explanatory variables which can be used in explaining the plant performance. Among these variables, chemical oxygen demand (COD) and BOD, are the two important. Thus, predicting the COD and BOD in the effluent, as the WSP’s performance indices, depending upon the influent raw wastewater quality will aid the operator to discover the foremost effective factors on treatment efficiency and take necessary safety measures before the occurrence of any challenge. Furthermore, given the required time needed to measure BOD, (5 days) and COD (2 hours), and also their procedures which involve the use of several dangerous chemicals, the prediction of these parameters values instead of their measurement may be an environmentally and economically safe approach (7).

Attention has been diverted from manually treatment plant effluent to mathematically techniques due to these objectives (8). As a result, modeling of the effluent characteristics of treatment system is imperative for the prediction of plant performance and operation. Artificial neural network (ANN) technique is a non-parametric mathematical modeling technique which can be used for modeling such processes. It can be employed for better prediction of the process performance owing to their high accuracy, adequacy and quite promising applications in engineering (9-11). The ANN models have been used with the prediction objectives in several fields including air quality (12-14), water treatment (15), wastewater treatment (16), atmospheric sciences (17) among others.

The emphasis of this study is the analysis of a WSP treatment system data using data mining (DM) approaches. The data were collected from a WSP treatment system in Birjand, a city in the east of Iran. The purpose of this study is the prediction of the WSP performance in terms of COD and BOD as the main performance related parameters. Several analyses were performed by considering and comparing two DM techniques (i.e., multiple regression [MR] and neural network [NN]).

Methods

WSP description

The studied stabilization pond is located in Birjand, the capital city of Southern Khorasan province in east of Iran. It is located at a latitude of 32°86’ N and longitude of 59°21’ E and about 1490 m above sea level. Birjand city has a cold and dry climate. The average annual temperature is 16.35°C with the warmest time in June (average 27.5°C) and the coldest in February (average 3.2°C). The Birjand’s WSP has been constructed with a capacity of 10,500 cubic meter per day for a population of 64,000 people (18). The WSP configuration is divided into anaerobic, facultative and maturation ponds (Figure 1). According to Figure 1, this treatment system has a pretreatment unit which includes grit chamber and screens that is followed by the WSP systems. The wastewater samples were taken from the influent raw wastewater and maturation ponds effluent within a period of one year. The analysis of the influent raw wastewater composite samples characteristics (including pH, temperature, BOD, COD, total suspended solids [TSS], total dissolved solid [TDS], electrical conductivity [EC] and turbidity) were carried out according to the standard procedures (19). Only two main distinct parameters namely COD and BOD were analyzed for the effluent treated wastewater.

Data preparation

All performance parameters of Birjand’s WSP including pH, COD, BOD, TSS, TDS, EC and Turbidity were measured in Birjand wastewater treatment plant (water and wastewater laboratory). In this study, pH, EC and TDS were measured using portable devices. COD, BOD and TSS were determined according to standard methods (19). Data analyses were performed using statistical packages in R (20).

The descriptive statistics of the raw data of the influent and effluent wastewater characteristics are summarized in Table 1.

A dataset of the operation parameters of Birjand’s WSP was used where the objective was to estimate the effluent COD and BOD using influent wastewater characteristics (eight continuous attributes). The data comprised 96 raw instances, some of which are outlier values. Thus it is necessary to preprocess the data before fitting the DM models. This process includes operations such as choosing the data (e.g., attributes or examples) or dealing with missing values and outliers. Outlier data is defined as data with considerably distance from the normal distribution. However, in some instances, these outlier values may be correct because they are as a result of the natural product of the distribution of variables (21). All examples with missing values were deleted. The box-and-whisker plot was applied in order to detect outliers. Samples beyond the whiskers of the plot were considered as outliers (Figure 2).

Figure 2 summarizes each raw variable by four components as follows: a central line in each box is the sample median to specify the central tendency; a box (with edges of 25th and 75th percentiles) to indicate variability around the central tendency; whiskers around the box to show the range of the variable; and the observations beyond the

Figure 1. Schematic flow diagram for Birjand’s WSP. AP, anaerobic pond; PF, facultative pond; MP, maturation pond; WSP, waste stabilization pond (Adapted from reference 18).
whisker length are marked as outliers displayed with a circle sign that fall within the distance defined by quartile ±1.5× interquartile range (IQR). Where the IQR, or mid-spread or middle fifty, is a measure of statistical dispersion, being equal to the difference between the upper and lower quartiles, IQR = Q₃ − Q₁. In current study, using a normal distribution, the data falls out of three standard deviations (SDs) of the mean was considered as outliers. Thus, the data that were more than µ±3SD, were regarded as outliers. Figure 3 summarizes the Box-and-Whisker plot for each processed variable. The descriptive statistics of the pre-processed influent and effluent wastewater characteristics are summarized in Table 2.

The last step in the data preparation procedure was the data scaling. The objective here was to ensure that the statistical distribution of the values for each model input and output variable is approximately uniform. Therefore all attributes were standardized to a zero mean and one standard deviation (22).

### Table 1. Raw data descriptive statistics

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Unit</th>
<th>Mean</th>
<th>SD</th>
<th>SE</th>
<th>Max</th>
<th>Min</th>
<th>UB</th>
<th>LB</th>
</tr>
</thead>
<tbody>
<tr>
<td>pHin</td>
<td>-</td>
<td>7.91</td>
<td>0.13</td>
<td>0.03</td>
<td>8.20</td>
<td>7.50</td>
<td>7.94</td>
<td>7.89</td>
</tr>
<tr>
<td>Tin</td>
<td>°C</td>
<td>23.40</td>
<td>3.11</td>
<td>0.63</td>
<td>28.70</td>
<td>17.10</td>
<td>24.03</td>
<td>22.77</td>
</tr>
<tr>
<td>BODin</td>
<td>mg/L</td>
<td>510.64</td>
<td>101.63</td>
<td>20.59</td>
<td>726.00</td>
<td>210.00</td>
<td>531.23</td>
<td>490.04</td>
</tr>
<tr>
<td>CODin</td>
<td>mg/L</td>
<td>815.55</td>
<td>154.66</td>
<td>31.34</td>
<td>1167.00</td>
<td>516.00</td>
<td>846.89</td>
<td>784.22</td>
</tr>
<tr>
<td>TDSin</td>
<td>mg/L</td>
<td>1299.16</td>
<td>158.45</td>
<td>32.11</td>
<td>1680.00</td>
<td>962.00</td>
<td>1331.26</td>
<td>1267.05</td>
</tr>
<tr>
<td>TSSin</td>
<td>mg/L</td>
<td>238.73</td>
<td>210.52</td>
<td>42.66</td>
<td>2217.00</td>
<td>115.00</td>
<td>281.39</td>
<td>196.07</td>
</tr>
<tr>
<td>ECin</td>
<td>µ mohs/cm</td>
<td>3100.94</td>
<td>312.53</td>
<td>63.33</td>
<td>3810.00</td>
<td>2400.00</td>
<td>3164.26</td>
<td>3037.61</td>
</tr>
<tr>
<td>NTUin</td>
<td>NTU</td>
<td>211.22</td>
<td>113.00</td>
<td>22.90</td>
<td>1172.00</td>
<td>69.00</td>
<td>234.11</td>
<td>188.32</td>
</tr>
<tr>
<td>BODout</td>
<td>mg/L</td>
<td>116.85</td>
<td>20.80</td>
<td>4.22</td>
<td>182.00</td>
<td>67.00</td>
<td>121.07</td>
<td>112.64</td>
</tr>
<tr>
<td>CODout</td>
<td>mg/L</td>
<td>259.04</td>
<td>51.32</td>
<td>10.40</td>
<td>392.00</td>
<td>168.00</td>
<td>269.44</td>
<td>248.65</td>
</tr>
</tbody>
</table>

**Statistical modeling**

The modeling and evaluation of the model which is an iterative process are the key procedures in DM approach (23). This research addresses these steps with an emphasis on the use of NN and MR functional models to solve the regression goals for the prediction of the performance of a WSP treatment system. All experiments reported in this study were conducted using R statistical environment (20). R tool is an open access with a set of software packages that allows the manipulation of data, performing of calculations, drawing of graphics and conducting statistical analysis. Taking the advantage of open access, Cortez (24) developed the Rminer package that aids the use of DM techniques in classification and regression tasks. In the present work, the Rminer package was utilized in performing the statistical modeling. The regression dataset is made up of \( k \in \{1,...,N\} \) examples, each mapping an input vector \( (x^k_1, ..., x^k_n) \) to a given target \( y^k \). The error is given by \( e^k = y^k - \overline{y}_k \), where \( \overline{y}_k \) represents the average of the pre-

![Figure 2. Box-and-Whisker Plots for the raw data of WSP characteristics](image2)

![Figure 3. Box-and-Whisker Plots for the pre-processed data of WSP characteristics](image3)
Table 2. Processed data descriptive statistics

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Unit</th>
<th>Mean</th>
<th>SD</th>
<th>SE</th>
<th>Max</th>
<th>Min</th>
<th>UB</th>
<th>LB</th>
</tr>
</thead>
<tbody>
<tr>
<td>pHin</td>
<td>-</td>
<td>7.92</td>
<td>0.12</td>
<td>0.03</td>
<td>8.20</td>
<td>7.60</td>
<td>7.95</td>
<td>7.90</td>
</tr>
<tr>
<td>Tin</td>
<td>°C</td>
<td>23.36</td>
<td>3.11</td>
<td>0.65</td>
<td>28.70</td>
<td>17.10</td>
<td>24.00</td>
<td>22.71</td>
</tr>
<tr>
<td>BODin</td>
<td>mg/L</td>
<td>506.12</td>
<td>99.73</td>
<td>20.65</td>
<td>726.00</td>
<td>210.00</td>
<td>526.77</td>
<td>485.47</td>
</tr>
<tr>
<td>CODin</td>
<td>mg/L</td>
<td>807.25</td>
<td>147.92</td>
<td>30.63</td>
<td>1141.00</td>
<td>516.00</td>
<td>837.88</td>
<td>776.62</td>
</tr>
<tr>
<td>TDSin</td>
<td>mg/L</td>
<td>1297.57</td>
<td>158.83</td>
<td>32.89</td>
<td>1680.00</td>
<td>962.00</td>
<td>1330.46</td>
<td>1264.67</td>
</tr>
<tr>
<td>TSSin</td>
<td>mg/L</td>
<td>215.45</td>
<td>50.45</td>
<td>10.45</td>
<td>360.00</td>
<td>115.00</td>
<td>225.89</td>
<td>205.00</td>
</tr>
<tr>
<td>ECin</td>
<td>µmohs/cm</td>
<td>3100.98</td>
<td>315.17</td>
<td>65.27</td>
<td>3810.00</td>
<td>2400.00</td>
<td>3166.25</td>
<td>3035.71</td>
</tr>
<tr>
<td>NTUin</td>
<td>NTU</td>
<td>200.04</td>
<td>54.84</td>
<td>11.36</td>
<td>327.00</td>
<td>69.00</td>
<td>211.40</td>
<td>188.69</td>
</tr>
<tr>
<td>BODout</td>
<td>mg/L</td>
<td>115.51</td>
<td>18.88</td>
<td>3.91</td>
<td>176.00</td>
<td>67.00</td>
<td>119.42</td>
<td>111.60</td>
</tr>
<tr>
<td>CODout</td>
<td>mg/L</td>
<td>257.27</td>
<td>50.09</td>
<td>10.37</td>
<td>392.00</td>
<td>168.00</td>
<td>267.64</td>
<td>246.90</td>
</tr>
</tbody>
</table>

Abbreviations: SD, standard deviation; SE, standard error; UB, upper bound of 95% confidence interval for the mean; LB, lower bound of 95% confidence interval for the mean.
where $N$ denoted the number of considered cases, $y_i$ the observed value, $\hat{y}_i$ the predicted value and $\bar{y}$ the average of observed value. In all three metrics, lower values result in better predictive models. Nevertheless, the RMSE showing the overall accuracy of the model is more sensitive to high errors. Another approach used in comparing regression models is the regression error characteristic (REC) curve (30), which plots the error tolerance given in terms of the absolute deviation, versus the percentage of predicted points. An independent variable is expected to present low predictive error and high REC area. In this study, the comparison of NN and MR models was examined with a statistical test. The Welch’s two-sample $t$ test that handles inequality in variance by adjusting degrees of freedom was used to compare RAE of the models.

**Results**

Figure 4 shows the typical three-layered feed-forward ANN. Eight input nodes corresponding to eight independent attributes, four hidden layer nodes and one output node are estimating BODout (Figure 4a) and CODout (Figure 4b) concentrations. Connections between nodes are presented by solid lines, which are associated with synaptic weights adjusted during the training procedure. The bias nodes were also shown, with 1 as their output value. The NN model ($H = 4$, $decay = 0.75 \pm 0.03$) achieved the best predictive results for the prediction of CODout. Similar results were found for BODout in which an NN model ($H = 4$, $decay = 0.73 \pm 0.03$) predicts this parameter with lower error in comparison with MR regression. Table 3 shows the MAE, RAE and RMSE for the NN and MR models. The Welch’s two sample $t$ test was also carried out in order to compare the performance of NN and MR models for the prediction of BODout and CODout in terms of influent wastewater characteristics. The results are summarized in Table 4. The REC curves of NN models for CODout and BODout were shown in Figures 5a and 5b, respectively. The results of the sensitivity analysis (Figure 6a and 6b) procedure are useful for knowledge discovery for NN models. In this way, it is possible to quantify the contribution of a given attribute for the model. The diagnostic plots of fitted NN models for BODout and CODout are shown in Figures 7 and 8, respectively. For indicating the models performance, the actual vs. predicted values was visualized in Figure 7a and Figure 8a. It was observed that the targets were well tracked by the output. Figure 7b and Figure 8b show the residual histograms for the BODout and CODout NN models, respectively. The fitted normal distribution curve on the data set indicated that the residual of the models followed a normal distribution.
distribution. Figure 7c shows the residuals normal Q-Q plot of the BO-
Dout model. This plot helps to confirm the rationality behind the conclusion that the residuals are fairly normally distributed. Figure 8c shows the residuals normal Q-Q plot of the CODout model.

The relationship between standardized residuals and values predicted by the models were shown in Figure 7d and Figure 8d.

Discussion

According to Figure 6a and 6b which present the relative input importance of the NN model for CODout and BO-
Dout, respectively, the EC and TDS which are the measures of dissolved materials in wastewater are seen as the most important inputs. Conversely, the influent TSS has a minimal impact on the predicted variations of BODout and CODout. In other words, less than 10% of the BODout or CODout originated from the suspended solids entering the WSP and most of the influent suspended materials have a biodegradable nature and probably were converted or deposited through the treatment process. Thus the most fractions of the suspended biodegradable materials entering the WSP may be removed by treatment process, while some part of biodegradable materials, which are mainly dissolved materials, were leaved the system without being remarkably treated. This may be as a result of low considerable effect of the treatment process on the elimination of soluble materials entering the WSP. The lower metric values for NN models revealed the higher performance of NN models in comparison with MR models.

Welch's two sample t test was found to be statistically significant, indicating that the NN and MR have different performances and the former is better (P<0.05) for the prediction of both BODout and CODout parameters. The obtained average $RAE = 78.71 \pm 1.16$ for BODout and 83.67 ± 1.35 for CODout in the case of NN model are better when statistically compared with MR model results ($RAE = 84.40 \pm 1.07$ for BODout and 88.07 ± 0.80 for CODout). The REC curve plots in Figures 5a and 5b also confirm the NN model performance superiority. The whiskeys in all graphs represent the 95% t student CI. According to the findings of this study, both the REC curves and global metrics, present the best performance to be the NN model followed by MR model. In fact, the NN model for both BODout and CODout responses have lower errors and higher area under the REC curve.

The degradability of WSP’s influent wastewater as the ratio of BOD$_5$ to COD is called Biodegradability Index (BI). Generally, the BI Index ranges between 0.4 to 0.8 for domestic wastewater. If BOD$_5$/COD is >0.6 then the waste is biodegradable fairly and can be effectively treatable using biological treatment methods. If BOD$_5$/COD is >0.6 then the waste is fairly biodegradable and can be excellently treated using biological treatment. If BOD$_5$/COD ratio is between 0.3 and 0.6, then there is need to seed the wastewater in order to treat it biologically. In cases where the BOD$_5$/COD is <0.3, then the wastewater is not biologically treatable (13,31). From data in Table 2, the BI index was obtained to be 0.62 and 0.44 for influent and effluent of WSP, respectively. These values indicate that the influent wastewater can be classified as degradable and the ef-
fluent has no further biodegradability potential. This is consistent with Figure 6a where the BODin has a minimal impact on the predicted variations of CODout. In other words, most fractions of materials in the effluent are dissolved and non-biodegradable materials.

The coefficient of determination ($R^2$) between observed and predicted BODout and observed and predicted CODout were 0.56 and 0.54, respectively. Despite the slightly weak correlation values, the points are almost well aligned on the acceptable prediction diagonal of coordinates 1:1. These values which indicate that 56% and 54% of the variance of the BODout and CODout variables, respectively, can be explained by using the network input attributes. The remaining 44% and 46% of their variances can be attributed to unknown, lurking variables, or inherent variability (11). The Histogram of the residuals can be used to determine if the variance is normally distributed. A symmetric bell-shaped histogram which is evenly distributed around zero shows that the normality assumption of the residuals is likely to be valid (32,33).

According Figure 7b and Figure 8b the residuals are close to a normal distribution around a mean value of $0.003 \pm 2.9$ (SD=14.00) and $0.071 \pm 7.15$ (SD = 34.54) for BODout and CODout NN models, respectively. A Q-Q (Quantile-Quantile) plot is another graphic method for testing if the residuals of the models follow the normal distribution. The residuals are said to follow a normal distribution if all the scatter points are close to the reference line (34). The point pattern in the middle of Q-Q plot is fairly linear. Chambers (35) and Fowlkes (36) discuss the interpretations of commonly encountered departures from linearity. When the left end of pattern is below the line and right end of pattern is above the line, then there are long tails emanated from a normal distribution. Consequently, 95% of the standardized residuals will also fall between -2 and +2 in the residual plot (33). Figure 7d and Figure 8d show a random scatter around zero with only a few points outside the ±2 limits. Therefore, the points are properly distributed on both sides of the horizontal line of zero ordinate representing the standardized average of the residuals.

In recent years, modeling of wastewater treatment plant (WWTP) or constructed wetlands performances through NN for the prediction of wastewater characteristic parameters has gained enormous interest. The ANN application of these articles was summarized in Table 5. As can be seen in Table 5 the ANN models performance based on coefficient of determination for the prediction of effluent BOD$_5$ varies between 31% and 84 %. For effluent COD the models performances varies between 39 and 98%. The values of 0.56 for BOD$_5$ and 0.54 for COD obtained in this study are within the range of performance predictions reported in literature.

According to Table 5, it can be established that DM approaches were most commonly used for the performance prediction of the conventional WWTP systems. Indeed, DM approach were rarely utilized for the performance prediction of natural treatment system such as WSPs and constructed wetlands (CWs) which have been proven to be effective substitutes for treating wastewater (45). However, results of this study could be the basis for better understanding of DM approach using ANN models for the prediction of WSPs performance.

**Conclusion**

The real-world DM application case for the prediction of a WSP performance presented in this study indicates the possibility of analyzing data by using the R environment and `rmriner` package. It was shown that the influent WSP indices could be applied to the prediction of effluent quality. Two DM techniques were explored: MR and NN. Overall, Welch’s two-sample $t$ test on global metrics and

<table>
<thead>
<tr>
<th>Predicted parameters</th>
<th>Location</th>
<th>$R^2$</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>BOD$_5$, COD, TSS</td>
<td>WWTP, El-Agamy</td>
<td>Reaching up to 0.9</td>
<td>(39)</td>
</tr>
<tr>
<td>COD, TSS</td>
<td>Paper mill WWTP, China</td>
<td>0.98 for COD; 0.96 for SS</td>
<td>(40)</td>
</tr>
<tr>
<td>BOD$_5$, COD, TSS</td>
<td>WWTP, Doha</td>
<td>0.39-0.84 for COD; 0.31-0.84 for BOD$_5$; 0.54-0.96 for TSS</td>
<td>(10)</td>
</tr>
<tr>
<td>COD$_5$, pH, TSS, COD, TN</td>
<td>WWTP, Taiwan industrial park</td>
<td>0.92 for SS; 0.86 for COD; 0.90 for pH</td>
<td>(41)</td>
</tr>
<tr>
<td>NH$_4$, BOD$_5$, TSS, COD, TN</td>
<td>WWTP, Denmark</td>
<td>More than 0.95</td>
<td>(42)</td>
</tr>
<tr>
<td>pH, TDS, turbidity, TN, TP</td>
<td>Industrial WWTP, Iran</td>
<td>0.85-0.92 for pH; 0.23-0.53 for Turbidity; 0.07-0.45 for TP; 0.22-0.51 for TN; 0.67-0.82 for TDS</td>
<td>(43)</td>
</tr>
<tr>
<td>BOD$_5$, COD</td>
<td>Industrial WWTP, Govindpura</td>
<td>0.64-0.87 for BOD$_5$, 0.52-0.68 for BOD$_5$</td>
<td>(8)</td>
</tr>
<tr>
<td>BOD$_5$</td>
<td>Constructed wetlands (CWs), Greece</td>
<td>0.52-0.68 for BOD$_5$</td>
<td>(44)</td>
</tr>
<tr>
<td>BOD$_5$, COD</td>
<td>WSPs, Birjand, Iran</td>
<td>0.56 for BOD$_5$; 0.54 for COD</td>
<td>This study</td>
</tr>
</tbody>
</table>
REC curve analysis confirmed NN model as the best predictive model for the prediction of both BOD and COD parameters. The sensitivity analysis of the WSP influent characteristics for the selected NN model can offer guidance which will be very important for the prediction of WSP performance based on the effluent BOD and COD concentrations. EC and TDS were found to be the most descriptive influent wastewater characteristics for the prediction of the WSP performance.

Acknowledgments
The authors are thankful to the Center for Water Quality Research (CWQR) at the Institute for Environmental Research (IER) of Tehran University of Medical Sciences for providing facilities and supports for this research.

Ethical issues
The authors hereby certify that all data collected during the study are as stated in this manuscript and no data from the study has been or will be published elsewhere separately.

Competing interests
The authors declare that they have no competing interests.

Authors’ contributions
MK, MTG and MHE participated in raw data provision. AM and SN contributed to drafting and editing the manuscript. MH contributed in the design of study, data analysis and drafting the manuscript.

References