Environmental Health Engineering and Management Journal 2016, 3(x), x–x

http ehemj.com

Environmental Health Engineering and Management Journal

Original Article

Environmental Health Engineering and Management

Open Access Publish Free

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

**Maryam Khodadadi[1], Alireza Mesdaghinia[2], Simin Nasseri[2], Mohammad Taghi Ghaneian[3], Mohammad Hassan Ehrampoush[4], Mahdi Hadi[5*]**

[1]Ph.D. Student of Environmental Health, International Campus, Shahid Sadoughi University of Medical Sciences, Yazd, Iran

[2]Professor, Center for Water Quality Research (CWQR), Institute for Environmental Research (IER), Tehran University of Medical Sciences, Tehran, Iran

[3]Associate professor, Department of Environmental Health, School of Public Health, Shahid Sadoughi University of Medical Sciences, Yazd, Iran

[4]Professor, Department of Environmental Health, School of Public Health, Shahid Sadoughi University of Medical Sciences, Yazd, Iran

[5]Ph.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 [$BOD_5$], 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 $BOD_5$ (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). Sensitivity analysis was also applied to extract useful knowledge from NN model.

**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 $BOD_5$ 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

**Citation:** Khodadadi M, Mesdaghinia A, Nasseri S, Ghaneian MT, Ehrampoush MH, Hadi M. Prediction of the waste stabilization pond performance using linear multiple regression and multi-layer perceptron neural network: a case study of Birjand, Iran. Environmental Health Engineering and Management Journal 2016; 3(x): x–x.

**Article History:**
Received: 3 March 2016
Accepted: 30 May 2016
ePublished: 10 June 2016

***Correspondence to:**
Mahdi Hadi
Email: hadi_rfm@yahoo.com

## 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 ($BOD_5$) (3). Inappropriate operation of a waste water 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 $BOD_5$ are the two important. Thus, predicting the COD and $BOD_5$ 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 $BOD_5$ (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 $BOD_5$ 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, $BOD_5$, 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, $BOD_5$, 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

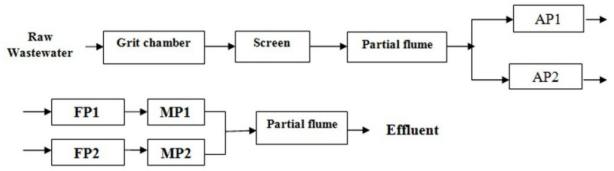


**Figure 1.** Schematic flow diagram for Birjand's WSP. AP, anaerobic pond; PF, facultative pond; MP, maturation pond; WSP, waste stabilization pond (Adapted from reference 18).

**Table 1.** Raw data descriptive statistics

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

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 = $Q_3 - Q_1$. In current study, using a normal distribution, the date 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).

## 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 ( $x_1^k,...,x_n^k$ ) to a given target $y_k$. The error is given by $e_k = y_k - \overline{y}_{pk}$, where $\overline{y}_{pk}$ represents the average of the pre-

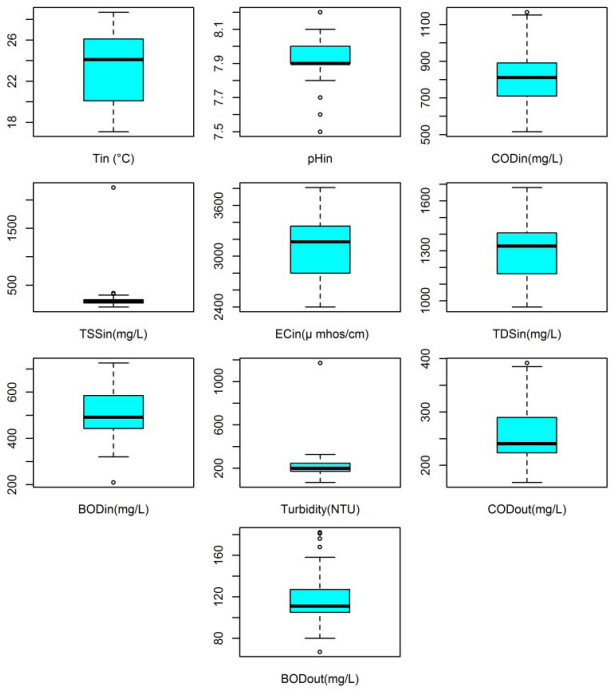


**Figure 2.** Box-and-Whisker Plots for the raw data of WSP characteristics

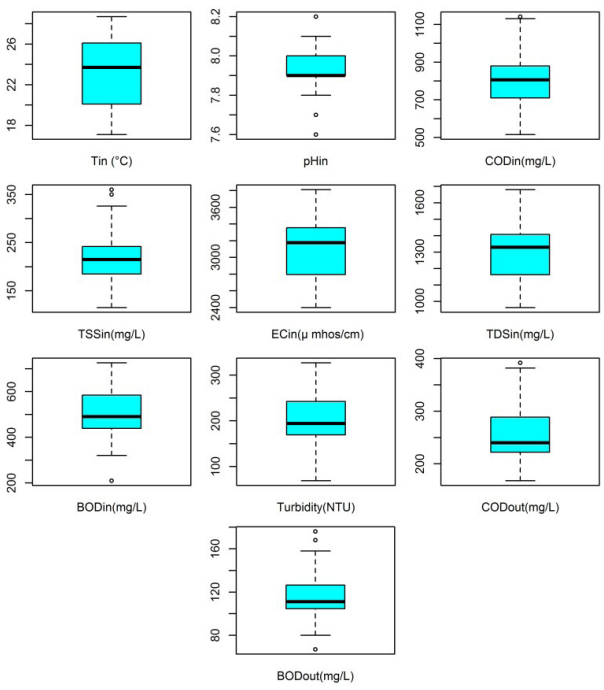


**Figure 3.** Box-and-Whisker Plots for the pre-processed data of WSP characteristics.

**Table 2.** Processed data descriptive statistics

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

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.

dicted value for the *k* input pattern.

Here, we selected eight wastewater treatment plant's influent characteristics (pH, BODin, Tin, CODin, TSSin, TDSin, ECin and NTUin) as the regression dataset input vectors. Both models were assessed with a supervised learning, where each model was adjusted to a dataset with examples that map input attributes into a given target (BODout and CODout).

The generalization performance of the models is often estimated by the holdout validation (i.e. train and test split). However, in some conditions the *k*-fold cross-validation is also used as the more robust approach (22). The latter approach is a more powerful method which requires *k* times more computation since *k* models are fitted in this approach. In this study, 10-fold cross-validation approach was used for the model's generalization performance estimation. The multilayer perceptron network was utilized for NN model. This network includes one hidden layer of H neurons with logistic functions. The overall model is given in the form:

$$y_i = f_i(w_{i,0} + \sum_{j=I+1}^{I+H} f_j(\sum_{n=1}^{I} x_n w_{m,n} + w_{m,0}) w_{i,n}) \quad (1)$$

where $y_i$ is the output of the network for node i, $w_{i,j}$ is the weight of the connection from node *j* to *i* and $f_j$ is the activation function for node *j*. In regression method, the output neuron is usually a linear function. The *Rminer* use the *nnet* package for modeling the NN. In *nnet* the optimization is done through the BFGS (Broyden, Fletcher, Goldfarb and Shanno )method of *optim* package. The BFGS is a quasi-Newton method, specifically published simultaneously in 1970 by Broyden, Fletcher, Goldfarb and Shanno (25). This uses function values and gradients to build up a structure of the surface to be optimized. For regression tasks, the algorithm minimizes the squared error (22). To solve this issue, the solution adopted is to train *Nr* different networks and then select the NN with the lowest error (22). In *Rminer*, this option is set using *model = 'mlp.'* The performance of NN model depends significantly on the number of nodes in hidden layer. NN in simplest form has *H = 0*, while some more complex NN may use a high values for *H* (24). Optimizing the network structure is a crucial step in the design of NNs. The NN structure must be optimized to minimize computer processing and obtain a good performance to avoid overfitting (10). There is no way to determine the best number of hidden units without training several networks and estimating the generalization error of each (26). If the hidden units are few, then high training error and high generalization error due to under-fitting may result. Conversely, if many hidden units are used, low training error can be achieved at the expense of network generalization which degrades overfitting (27). Since the NN network was not too complex in this study, the number of hidden nodes (*H*) was estimated by using the formula, *H=I/2*, where *I* is the number of nodes in input layer. However, the weight decay factor hyper-parameter was optimized with a 10-range grid-search between zero and one. To avoid overfitting for the training data set, an internal 5-fold cross-validation was used. In this study all input and output attributes were standardized with zero mean and standard deviation of unity using setting *scale* argument in *fit* function on *all* value.

After selecting the best parameters, the model was retrained with all training data. To quantify the importance of input variables in the model, the sensitivity analysis was applied after the training phase in order to analyze the model responses when a given input is changed. The weight of each input variable was measured by varying its value through its full range while the other input variables remained with their mean values (28). If the analyzed input variable is very important, its variance in the model output will be high. Therefore, the most important variable in the model output is the input variable that has a higher variance (24). The overall performance of each model was computed by the global metrics, namely the mean absolute deviation (MAD), root mean squared error (RMSE) and relative absolute error (RAE), which can be computed as (29):

$$MAD = \frac{\sum_{i=1}^{N} |y_i - \hat{y}_i|}{N} \quad (1)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{N}} \quad (2)$$

$$RAE = \frac{\sum_{i=1}^{N}|y_i - \hat{y}_i|}{\sum_{i=1}^{N}|y_i - \bar{y}_i|} \times 100\% \quad (3)$$

where $N$ denoted the number of considered cases, $y_i$ the observed value, $\hat{y}_i$ the predicted value and $\bar{y}_i$ 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 presents 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 ± 0.03) achieved the best predictive results for the prediction of effluent COD (CODout). Similar results was found for BODout in which an NN model ($H$ = 4, $decay$ = 0.73 ± 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 RMSE is the root of mean squared difference between outputs and targets.

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.

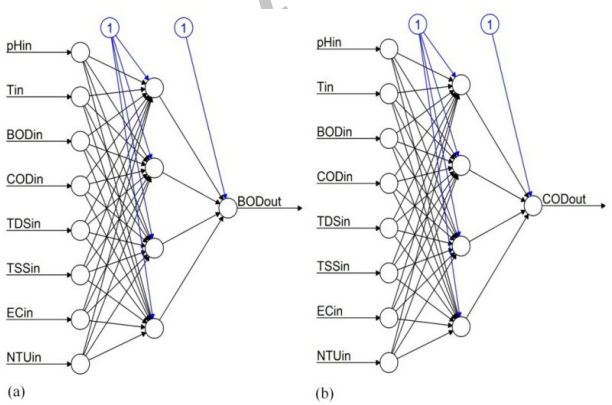


**Figure 4.** Structure of the constructed three-layer feed-forward ANN to predict BODout (a) and CODout (b) .

**Table 3.** Global metrics for the NN and MR models

| Mertic | BODout | | CODout | |
|---|---|---|---|---|
| | NN | MR | NN | MR |
| MAE | 11.41±0.16 | 12.24±0.15 | 33.72±0.54 | 35.50±0.32 |
| RAE | 78.71±1.16 | 84.40±1.07 | 83.67±1.35 | 88.07±0.80 |
| RMSE | 16.37±0.21 | 16.81±0.22 | 44.75±0.74 | 45.55±0.36 |

**Table 4.** Welch's two sample *t* test results for comparing NN and MR models

| Parameter | Model | RAE | t statistic | df | *P* value* |
|---|---|---|---|---|---|
| BODout | NN | 78.71±1.16 | -8.1337 | 17.89 | 0.0000 |
| | MR | 84.40±1.07 | | | |
| CODout | NN | 83.67±1.35 | -6.3215 | 14.64 | 0.0000 |
| | MR | 88.07±0.80 | | | |

*Ho = no difference between means; H1 = true difference in means is not equal to 0; 95% CI.

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

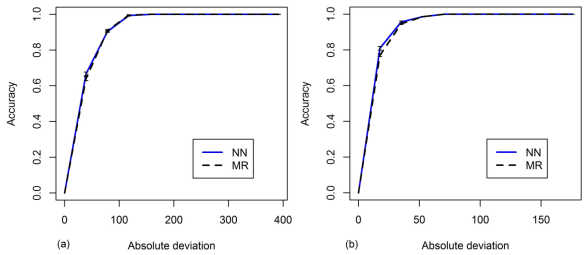


**Figure 5.** The REC curves confirm the NN performance superiority for the prediction of CODout (a) BODout (b).

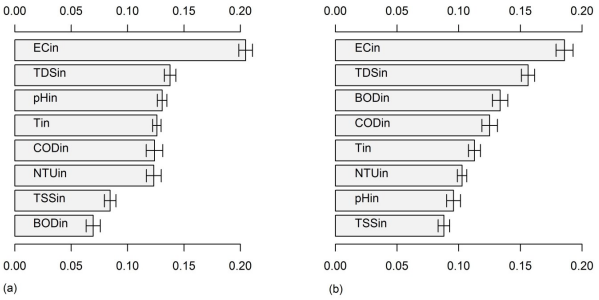


**Figure 6.** The relative input importance of the NN model for CODout (a) and BODout (b) in the order of importance shows the ECin and TDSin as the most relevant inputs.

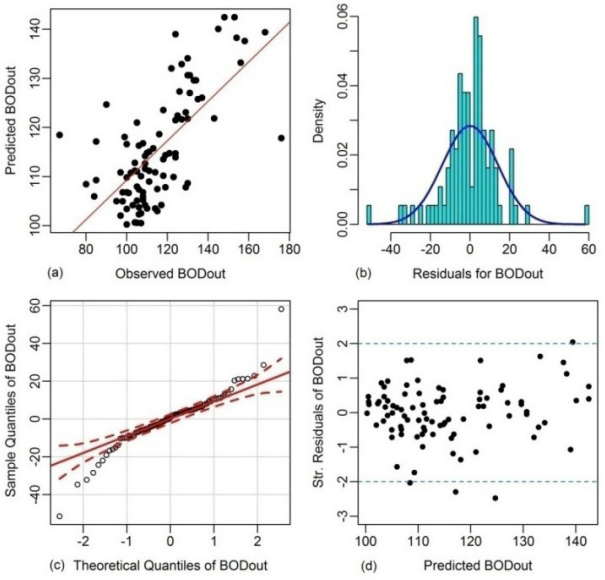


**Figure 7.** Result of NN model to predict BODout with a 10-fold external and a 5-fold internal cross-validation approach. Scatter plot of estimated values vs. observed values ($R^2$ = 0.56) (a). Histogram of residuals with normal adjustment curve (mean = 0.003 ± 2.9) in the data set (b). Q-Q plot of the agreement between the residual quantiles and normal quantiles (c). Relationship between standardized residuals and values predicted by the model (d).

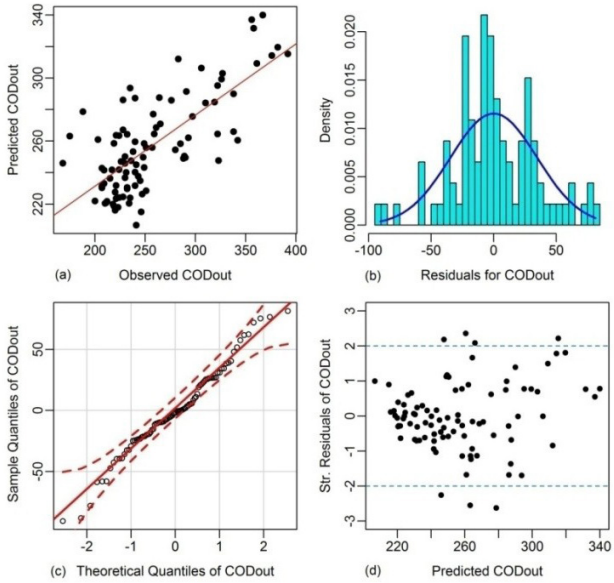


**Figure 8.** Result of NN model to predict CODout with a 10-fold external and a 5-fold internal cross-validation approach. Scatter plot of estimated values vs. observed values ($R^2$ = 0.54) (a). Histogram of residuals with normal adjustment curve (mean = 0.071 ± 7.15) in the data set (b). Q-Q plot of the agreement between the residual quantiles and normal quantiles (c). Relationship between standardized residuals and values predicted by the model (d).

distribution.

Figure 7c shows the residuals normal Q-Q plot of the BODout 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 BODout, 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 ± 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% ± 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 whiskers 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 at both ends of the data distribution. No curvature was seen in the middle in Figure 7c thus showing that there is no skewness in the residuals distribution. It is noted that at the left and right ends of the plot, the circles are somewhat farther away from the line than elsewhere in the plot. But it is common for points at either end of the plot to be farther from the line than elsewhere, even when the data are normal (37). When the distribution of the residuals is skewed and their variance is found not to be constant, a transformation on the response variable may be quite advantageous (38). For CODout model (Figure 8c) the circles or points all lie quite close to the line and within the 95% CI zone; as such is safe to say that these residuals emanated from a normal distribution. Consequently, the normality assumption of residuals for both BODout and CODout models may be appreciated. In the standardized residuals rescale residual values by the regression standard error, if the residuals are found to be distributed normally, about 95% should fall within 2σ around the fitted curve. 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 *rminer* 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 5.** Summary ANN application of researcher's studies in modeling of WWTPs and WTPs

| Predicted parameters | Location | $R^2$ | References |
|---|---|---|---|
| $BOD_5$, COD, TSS | WWTP, El-Agamy | Reaching up to 0.9 | (39) |
| COD, TSS | Paper mill WWTP, China | 0.98 for COD ; 0.96 for SS | (40) |
| $BOD_5$, COD, TSS | WWTP, Doha | 0.39-0.84 for COD; 0.31-0.84 for $BOD_5$; 0.54-0.96 for TSS | (10) |
| $COD_5$, TSS, pH | WWTP, Taiwan industrial park | 0.92 for SS; 0.86 for COD; 0.90 for pH | (41) |
| $NH_3$, $BOD_5$, TSS, COD, TN | WWTP, Denmark | More than 0.95 | (42) |
| pH, TDS, turbidity, TN, TP | Industrial WWTP, Iran | 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 | (43) |
| $BOD_5$ | Industrial WWTP, Govindpura | 0.64-0.87 for $BOD_5$ | (8) |
| $BOD_5$ | Constructed wetlands (CWs), Greece | 0.52-0.68 for $BOD_5$ | (44) |
| $BOD_5$, COD | WSPs, Birjand, Iran | 0.56 for $BOD_5$; 0.54 for COD | This study |

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

1. Maier HR, Dandy GC. Neural network based modelling of environmental variables: a systematic approach. Math Comput Model 2001; 33(6): 669-82.
2. Mozaheb S, Ghaneian M, Ghanizadeh G, Fallahzadeh M. Evaluation of the stabilization ponds performance for municipal wastewater treatment in Yazd, Iran. Middle-East Journal of Scientific Research 2010; 6(1): 76-82.
3. Wiley PE, Brenneman KJ, Jacobson AE. Improved algal harvesting using suspended air flotation. Water Environ Res 2009; 81(7): 702-8.
4. Peterson JD, Murphy RR, Jin Y, Wang L, Nessl MB, Ikehata K. Health effects associated with wastewater treatment, reuse, and disposal. Water Environ Res 2011; 83(10): 1853-75.
5. Shammas NK, Wang LK, Wu Z. Waste stabilization ponds and lagoons. In: Wang LK, Pereira NC, Yung-Tse Hung YT, eds. Biological Treatment Processes. Springer; 2009. p. 315-70.
6. Saqqar M, Pescod M. Modelling the performance of anaerobic wastewater stabilization ponds. Water Sci Technol 1995; 31(12): 171-83.
7. Al-Asheh S, Mjalli FS, Alfadala HE. Forecasting influent-effluent wastewater treatment plant using time series analysis and artificial neural network techniques. Chemical Product and Process Modeling 2007; 2: 3.
8. Vyas M, Modhera B, Vyas V, Sharma A. Performance forecasting of common effluent treatment plant parameters by artificial neural network. J Eng Appl Sci 2011; 6(1): 38-42.
9. Hanbay D, Turkoglu I, Demir Y. Prediction of wastewater treatment plant performance based on wavelet packet decomposition and neural networks. Expert Syst Appl 2008; 34(2): 1038-43.
10. Mjalli FS, Al-Asheh S, Alfadala H. Use of artificial neural network black-box modeling for the prediction of wastewater treatment plants performance. J Environ Manage 2007; 83(3): 329-38.
11. Solaimany Aminabad M, Maleki A, Hadi M. Application of artificial neural network (ANN) for the prediction of water treatment plant influent characteristics. J Adv Environ Health Res 2014; 1(2): 89-100.
12. Rostami Fasih Z, Mesdaghinia A, Nadafi K, Nabizadeh Nodehi R, Mahvi AH, Hadi M. Forecasting the air quality index based on meteorological variables and autocorrelation terms using artificial neural network. Razi Journal of Medical Sciences 2015; 22(137): 31-43. [In Persian].
13. Motesaddi S, Nowrouz P, Alizadeh B, Khalili F, Nemati R. Sulfur dioxide AQI modeling by artificial neural network in Tehran between 2007 and 2013. Environmental Health Engineering and Management Journal 2015; 2(4): 173-8.
14. Shakerkhatibi M, Mohammadi N, Zoroufchi Benis K, Behrooz Sarand A, Fatehifar E, Asl Hashemi A. Using ANN and EPR models to predict carbon monoxide concentrations in urban area of Tabriz. Environmental Health Engineering and Management Journal 2015; 2(3): 117-22.
15. Khataee AR, Kasiri MB. Artificial neural networks modeling of contaminated water treatment processes by homogeneous and heterogeneous nanocatalysis. Journal of Molecular Catalysis A: Chemical 2010; 331(1): 86-100.
16. Hamed MM, Khalafallah MG, Hassanien EA. Prediction of wastewater treatment plant performance using artificial neural networks. Environ Model Softw 2004; 19(10): 919-28.
17. Gardner MW, Dorling S. Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. Atmos Environ (1994) 1998; 32(14): 2627-36.
18. Hayati H, Doosti M, Sayadi M. Performance evaluation of waste stabilization pond in Birjand, Iran for the treatment of municipal sewage. Proceedings of the International Academy of Ecology and Environmental Sciences 2013; 3(1): 52-8.
19. Rice EW. Standard methods for the examination of water and wastewater. Washington, DC: American Public Health Association; 2012.
20. R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2013. Available from: http://www.R-project.org.
21. Masters T. Practical neural network recipes in C++. California: Morgan Kaufmann; 1993.
22. Hastie T, Tibshirani R, Friedman J, Franklin J. The elements of statistical learning: data mining, inference and prediction. The Mathematical Intelligencer 2005; 27(2): 83-5.
23. Chapman P, Clinton J, Kerber R, Khabaza T, Reinartz T, Shearer C, et al. CRISP-DM 1.0 Step-by-step data mining guide. 2000. Available from: https://www.the-modeling-agency.com/crisp-dm.pdf.
24. Cortez P. Data mining with neural networks and support vector machines using the R/rminer tool. In: Perner p, ed. Advances in Data Mining Applications and Theoretical Aspects. Springer; 2010. p. 572-83.

25. Fletcher R. A new approach to variable metric algorithms. Comput J 1970; 13(3): 317-22.
26. Nguyen N, Cripps A. Predicting housing value: A comparison of multiple regression analysis and artificial neural networks. Journal of Real Estate Research 2001; 22(3): 313-36.
27. Geman S, Bienenstock E, Doursat R. Neural networks and the bias/variance dilemma. Neural Comput 1992; 4(1): 1-58.
28. Kewley RH, Embrechts MJ, Breneman C. Data strip mining for the virtual design of pharmaceuticals with neural networks. IEEE Trans Neural Netw 2000; 11(3): 668-79.
29. Witten IH, Frank E. Data Mining: Practical Machine Learning Tools and Techniques. Amsterdam: Morgan Kaufmann; 2005.
30. BI J, Bennett P. Regression error characteristic curves. Twentieth International Conference on Machine Learning (ICML-2003); Washington, DC; 2003.
31. Rim-Rukeh A, Agbozu L. Impact of partially treated sewage effluent on the water quality of recipient Epie Creek Niger Delta, Nigeria using Malaysian Water Quality Index (WQI). J Appl Sci Environ Manag. 2013; 17(1): 5-12.
32. Montgomery DC, Runger GC, Hubele NF. Engineering Statistics. New York: John Wiley & Sons; 2009.
33. Graphic Residual Analysis. Available from: http://www.originlab.com/doc/Origin-Help/Residual-Plot-Analysis.
34. Thode HC. Testing for normality. Boca Raton: CRC Press; 2002.
35. Chambers JM. Graphical Methods for Data Analysis. New York: Champman & Hall; 1983.
36. Fowlkes EB. A folio of distributions: A collection of theoretical quantile-quantile plots. New York: Marcel Dekker; 1987.
37. Brown DE. Bro. Brown's statistics reference pages: examples of interpreting Q-Q plots 2014. Available from: http://emp.byui.edu/BrownD/Stats-intro/dscrptv/graphs/qq-plot_egs.htm.
38. Neter J, Kutner MH, Nachtsheim CJ, Wasserman W. Applied Linear Statistical Models. Chicago: Irwin; 1996.
39. Nasr MS, Moustafa MA, Seif HA, El Kobrosy G. Application of artificial neural network (ANN) for the prediction of EL-AGAMY wastewater treatment plant performance-EGYPT. Alexandria Engineering Journal 2012; 51(1): 37-43.
40. Wan J, Huang M, Ma Y, Guo W, Wang Y, Zhang H, et al. Prediction of effluent quality of a paper mill wastewater treatment using an adaptive network-based fuzzy inference system. Appl Soft Comput 2011; 11(3): 3238-46.
41. Pai T, Yang P, Wang S, Lo M, Chiang C, Kuo J, et al. Predicting effluent from the wastewater treatment plant of industrial park based on fuzzy network and influent quality. Appl Math Model 2011; 35(8): 3674-84.
42. Raduly B, Gernaey KV, Capodaglio A, Mikkelsen PS, Henze M. Artificial neural networks for rapid WWTP performance evaluation: methodology and case study. Environ Model Softw 2007; 22(8): 1208-16.
43. Naser M, Hamed H, Mohammad TJ, Hamidreza H, Hamid A. Simulation of low TDS and biological units of Fajr industrial wastewater Treatment plant using artificial neural network and principal component analysis hybrid method. Journal of Water Resource and Protection 2012; 6(4): 370-6.
44. Akratos CS, Papaspyros JN, Tsihrintzis VA. An artificial neural network model and design equations for BOD and COD removal prediction in horizontal subsurface flow constructed wetlands. Chem Eng J 2008; 143(1): 96-110.
45. Kayombo S, Mbwette T, Katima J, Ladegaard N, Jørgensen S. Waste stabilization ponds and constructed wetlands -design manual. UNEP-IETC with the Danish International Development Agency (Danida); 2005.