Avicenna J Environ Health Eng. 2018 June;5(1):15-20 http://ajehe.umsha.ac.ir

Original Article

doi 10.15171/ajehe.2018.03

Using Multilayer Perceptron Artificial Neural Network for Predicting and Modeling the Chemical Oxygen Demand of the Gamasiab River

Mohamad Parsimehr[1], Kamran Shayesteh[1*], Kazem Godini[2], Maryam Bayat Varkeshi[3]

[1]Department of Environmental Science, Faculty of Natural Resources and Environment, Malayer University, Malayer, Hamedan, Iran
[2]Department of Environmental Health, Health Sciences Research Center, Hamedan University of Medical Sciences and Health Services, Hamedan, Iran
[3]Department of Water Engineering, Faculty of Agriculture, Malayer University, Hamedan, Iran

***Correspondence to**
Kamran Shayesteh, Department of Environmental Science, Faculty of Natural Resources and Environment, Malayer University, Malayer, Hamedan, Iran.
Tel: +989123784864;
Email: K.shayesteh@malayeru.ac.ir

Published online June 23, 2018

Abstract
Concerns about water quality have widely increased in the last three decades; thus, water quality is now as important as its quantity. To study and model the quality of the Gamasiab River, its data, including chemical oxygen demand (COD), biological oxygen demand (BOD), dissolved oxygen (DO), total dissolved solids (TDS), total suspended solids in water, acidity, temperature, turbidity, and cations and anions were measured at four stations. Then, the correlations between these parameters and COD were measured using Pearson's correlation coefficient and modeled by multilayer perceptron artificial neural network. In order to minimize the cost of the experiments performed and to provide the input parameters to the artificial neural network based on the correlations between the data and COD, the number of input parameters was reduced and finally, model No.3, with the Momentum training function and the TanhAxon activation function with the validation correlation coefficient of 0.97, mean absolute error of 2.88, and normalized root mean square error of 0.11 was identified as the most accurate model with the lowest cost. The results of the present study showed that the multilayer perceptron neural network has high ability in modeling the COD of the river, and those data correlated with each other have the greatest effect on the model. Moreover, the number of input parameters can be reduced in order to lower the cost of experiments while the performance of the model is not undermined.

Keywords: Artificial Intelligence, River Quality, Environmental Assessment

Received April 5, 2018; **Revised** June 10, 2018; **Accepted** June 13, 2018

1. Introduction

The importance of water, as a vital element for the survival of civilizations, has been proven throughout human history. Concerns about water quality have widely increased in the last 3 decades; thus, water quality is now as important as its quantity. Water pollution not only affects its quality, but also threatens human health, economic development, and social welfare (1).

Rivers are the most important sources of water for drinking, industry, and agriculture. Therefore, the study of quality parameters and prediction of their changes are the objectives of environmental managers and planners (2). Chemical oxygen demand (COD) is widely used as an important index for determining the relative presence of organic pollutants in water to monitor the environment and assess the environmental impacts. Nutrients in industrial wastewaters and household sewages discharged into the rivers in result of human activities can lead to eutrophication phenomena, environmental issues, loss of aquatic ecosystem performance, and so on (3). In this regard, COD is considered as one of the best indicators of the amount of water pollution (4).

Gamasiab River passes near the city of Nahavand and its spring is also located there. In many cases, urban and industrial effluents of this city and its neighbor villages are released into the river. It should be pointed out that the quality of this river has been influenced by these kinds of polluters. Unfortunately, the demand for water intake and discharge of pollutants into the river has also increased; therefore, simulation and prediction of pollution for this river is of great importance.

In order to implement water quality standards (i.e., to ensure that the maximum permissible concentration of a substance in water is not exceeded), river water quality models are commonly utilized in research as well as planning. However, most models are based on linear

functions. In the past, it was tried to predict dissolved oxygen (DO) in the river under various scenarios using certain models, but, in practice, the statistical accuracy of these models was generally low, because natural systems were and still are very complicated for deterministic models. Artificial neural network (ANN) is a fast and flexible tool used to make a model for the estimation of water quality. In recent years, ANN has shown exceptional performance as a regression tool, especially when used for model identification and performance estimation (5). However, the Streeter–Phelps equation is used to investigate water pollution through explaining a decrease in DO in a river or stream along a certain distance by degradation of biological oxygen demand (BOD) (6). On the other hand, neural networks, dislike experimental methods, have the ability to establish a meaningful relationship between the data about the water quality (7).

The neural network approach has several advantages compared to conventional methods or semi-experimental models. Most of these methods, for example, require a lot of input parameters, while neural networks predict changes in desired parameters with acceptable accuracy by means of the least number of measured parameters (8).

Talib and Amat examined the quality of the Dondang River using the ANN for modeling the COD. The input variables of their model were: DO, BOD, suspended solids (SS), pH, ammonia (NH_3), temperature, nitrate (NO_3), total solids (TS), and phosphate (PO_4). In their study, the COD prediction included training, testing, and validation of the model. Their results showed that BOD was the most important variable determining COD, followed by phosphate, DO, solid contents, and temperature (9).

To estimate the COD concentration, Ay and Kisi developed a model by combining the k-means clustering method and the perceptron ANN. The performance of this model was compared with that of multivariable regression, multilayer perceptron, radial neural network, generalized regression neural network, as well as two different fuzzy and neural fuzzy comparative inference methods. The results indicated that the proposed model had higher accuracy and less error. In addition, k-means clustering with multilayer perceptron could be used as a tool for modeling daily COD concentration (10). Further, Ruben et al described how ANN is applied to predict COD levels. In their study, it was described how the ANN was used to predict COD of the river in Chisinau City, China. In their study, the simulation was performed using 10 neurons in a hidden layer and seven input variables (temperature, DO, total nitrogen (TN), total phosphorus (TP), suspended sediment, turbidity, and NH3-N), and a multilayer perceptron (MLP). The correlation coefficients of the modeled results were 0.96, 0.94, and 0.89 in the validation stage, respectively (11).

In the present study, the multilayer perceptron artificial neural network was used to model the COD levels of the Gamasiab River. Its optimal structure was also examined to reduce the cost of experiments performed to obtain input data.

2. Materials and Methods

Case under this study: The Gamasiab River is a calcareous spring located 61 km southeast of the city of Nahavand, Hamedan, Iran. It originates from the northern slopes of Garrin Mountain, named Sarab Gamasiab. The average rainfall in this area is about 350-400 mm per year. In the vicinity of this river, there are wide agricultural fields where the most important products are wheat, vegetables, beetroot, tobacco, and so on. Fig. 1 shows the location of the case under this study. From the geological perspective, the catchment area of the Gamasiab River is the High Zagros Region. The studied area is severely crushed and faulty as a thin and narrow strip parallel to the High Zagros and located in the northeast of it. The highlands of the area are mainly made up of sedimentary and metamorphic rocks. The sedimentary rocks of the region are lime in karst regime and conglomerate.

2.1. Used Data

The data used in the present study included: COD, BOD, DO, total dissolved solids (TDS), total suspended solids (TSS), acidity (pH), temperature (T), turbidity (NTU), cations and anions. They were obtained from Nahavand Department of Environment. These data were collected monthly from April 2005 to March 2007 at four stations. The location of the stations is shown in Fig. 1. Maximum, minimum, mean values and standard deviations were calculated using SPSS software version 25.0. The summary of the results is presented in Table 1.

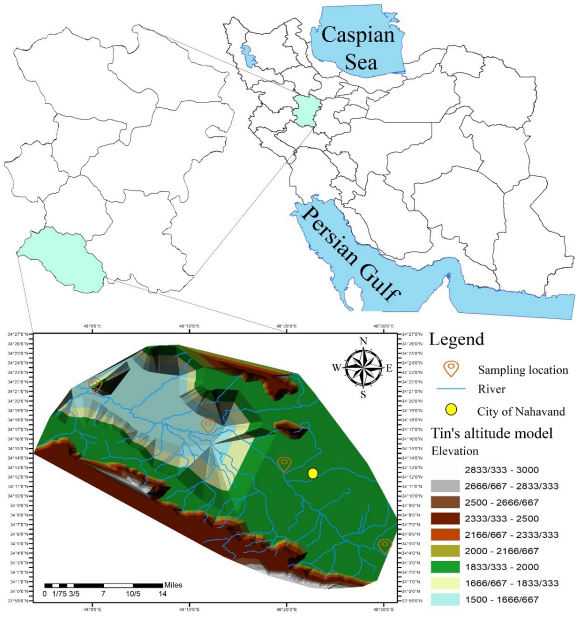

Fig. 1. The Location of the Gamasiab River.

Table 1. Dispersion Indicators of Used Data

	Mean	SD	Min	Max
COD	29.65	14.44	8	78
BOD	14.79	7.64	3	42
TDS	265.25	74	104	486
TSS	39.98	25.77	7	186
pH	7.75	0.315	7.1	8.42
T	11.01	4.81	4.1	24.8
DO	5.68	1.75	2.1	11.3
EC	440.93	123.89	173	810
NTU	13.93	13.93	1.93	63.1
CL^{-}	14.74	9.80	3	38
MG^{++}	20.86	9.14	3.36	46.5
CA^{++}	42.26	19.26	3.2	109
HW	195.25	48.28	106	306
Alk	198.62	44.64	82	354

Table 2. Results of the Pearson Correlation Coefficient Between the Parameters

Parameter	COD		
EC	0.249	0.055	60
COD	1		60
TDS	0.25	0.054	60
BOD	0.968**	0	60
TSS	0.558**	0	60
pH	-0.031	0.814	60
TEMP	0.18	0.168	60
DO	-0.288*	0.026	60
NTU	0.055	0.676	60
CL-	0.273*	0.035	60
MG++	0.117	0.375	60
CA++	0.226	0.082	60
HW	0.314*	0.015	60
Alk	0.066	0.616	60

2.2. Statistical Analysis

Pearson correlation coefficient was used to assess the correlation between the measured parameters and the impact of COD on the water quality compared to other parameters, using SPSS software version 25.0 (Table 2). Then, the optimal ANN models were studied to predict COD.

2.3. Artificial Neural Network

ANNs include an information processing pattern and a powerful tool for simulation, inspired from biological neurons. With this approach, ANN, like the human brain's biological structure, is able to solve complicated problems with linear or nonlinear nature by combining the features such as learning power, generalization, parallel processing, and decision-making (12).

The general structure of the ANN consists of 3 layers with separate tasks: the input layer with the role of distributing data in the network, the intermediate (hidden) layer with the task of information processing, and the output layer which, in addition to the processing per the input vector, shows its outputs. Neuron is considered to be the smallest processing unit of a network. As shown in Equation 1, the ANN operates in such a way that the net input of the neuron is obtained by summing up the product of multiplying of the input matrix P with the elements Pi (i = 1,2, ..., r) by the weight matrix W with the elements Wi (i = 1,2, ..., r), and a constant value with weight b.

$$n = \sum_{i=1}^{k} pi.wli + b = W.P + b \qquad \text{Eq. (1)}$$

where, k is the number of input parameters and b is a bias. Then, the network is obtained as Equation 2 by applying the output convention function f.

$$a = f(P.W + b) \qquad \text{Eq. (2)}$$

In the architectural step of the network, or the step of choosing network structure, the number of layers and how they are placed, as well as the synaptic weight of the network are determined by the designer. Fig. 2 shows the structure of the ANN. In the present study, a multilayer perceptron network with an error backpropagation algorithm was used. About 90% of ANN used in water-specific problems are of error backpropagation algorithm (13).

The network was designed based on the combination of parameters affecting the river water quality in previous times in the form of different structures of information in the input layer. In each structure, the post-processing input information is transmitted from the output of the neurons of first layer to the neurons of the subsequent layers, and eventually, to the network output if it is acceptable. Otherwise, computations are repeated again by propagating computational error to the previous layers. This process continues until an acceptable result is obtained. In order to increase the speed of information processing and not to stop the network at the local minimums, normalized data were used as the input. In the first structure, these structures were run in the software under Neuro Solution windows operating system with the ability of normalizing the data. Another advantage of this software is the existence of various functions with various

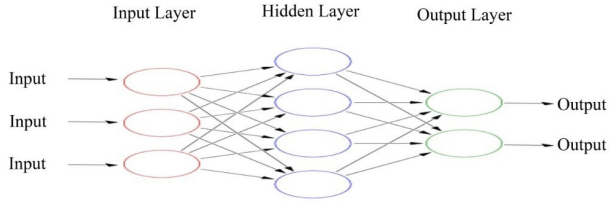

Fig. 2. The Overall Structure of the Artificial Neural Network.

algorithms in the software bank (14).

In order to model the ANN, first, a percentage of data is used for training the network. In the next step, the remaining percentage of data is used for validation and network testing. In the present study, 60% of the data were used for training and 40% for the validation and testing the network. In the training phase, the ANN learned the relationship between the variables by analyzing the input data. Then, in the testing and validation phases, the network was able to predict the COD indicator for 40% of the data by using 60% of the input data used to train the relationship between the variables. Next, the predicted data were compared with the actual data and the error value was calculated.

The error values should be the least one. For this purpose, the desired network should be designed and the practice of training and testing should be repeated with different modes to minimize the error. In order to evaluate and compare the results of the mean absolute error (MAE), the correlation coefficient (R), and normalized root mean square error (NRMSE), equations 3, 4, and 5 were used.

$$MAE: \frac{1}{n}\sum_{i}^{N}=1\left|x_i - y_i\right| \qquad \text{Eq. (3)}$$

where, x_i is the actual data, y_i is the estimated data, and n is the total number of data.

$$R = \frac{\sum_{i=1}^{n}(Y\,act - \hat{Y}\,act)(Y\,est - \hat{Y}\,est)}{\sqrt{\sum_{i=1}^{n}(Y\,act - \hat{Y}\,act)^2 \sum_{i=1}^{n}(Y\,est - \hat{Y}\,est)}} \qquad \text{Eq. (4)}$$

where, Y act is the actual value, $\hat{Y}$ act is the mean of actual values, Y *est* is the estimated value, and *est* is the mean of the estimated values.

$$NRMSE = \frac{RMSE}{COD\,average} \qquad \text{Eq. (5)}$$

where, the RMSE is the error rate and the COD average is the mean COD obtained from the tests.

Correlation coefficient (R) represents the correlation between the predicted values and the actual data. It is obvious that in this equation, the closer the R is to 1, the more acceptable the results are, and the closer the NMSAE and MAE values are to 0, the smaller the errors are.

The order in which the parameters in each model is used, accords with the correlations presented in Table 2, and they are used in the tests in accordance with greater convenience and lower cost, respectively. In Table 3, the numbers of parameters are reduced (15).

3. Discussion

The standard limit of COD for effluents to be released into the rivers is 60 mg/L. Sometimes the COD level discharged into this river is by 78 mg/L; this is indicative of the critical condition in which pollution exceeds the acceptable limit. Therefore, assessment and performance of protective and management programs are essential for this river. In the current study, the multilayer perceptron artificial neural network was used to model and predict the COD content of the Gamasiab River.

Table 3. The Parameters Studied in 3 Models

Input Parameters	Model
All 13 existing parameters	1
Five parameters have the possibility of correlation with COD	4
Two parameters have the highest probability of correlation with COD	9

Based on the results obtained from Pearson correlation coefficient, there were correlations between the COD parameter and the BOD, TSS, DO, Cl^- and hardness parameters, and the highest correlation was found to be between BOD and TSS parameters.

According to Table 3, a model with Momentum training function, and a model, with 13 inputs and 10 neurons in a hidden layer, with TanhAxon activation function showed the best results . The findings are presented in Fig. 3. TanhAxon activation function for a model with 5 inputs and 4 neurons in a hidden layer showed the best results (Fig. 4). By reducing the input data to 2 parameters, the accuracy of the model increases. Finally, a model with a Momentum training function and a model with 2 inputs and 4 neurons in a hidden layer with a linear TanhAxon activation function indicated the highest correlation with the least error rate between the results and the experimental data. The results are presented in Fig. 5. By reducing the number of input parameters and choosing the parameters with the highest probability of correlation, the error rate was decreased and the accuracy of the model was increased. As a result, Model No. 3 had the highest accuracy with the lowest input and cost of the experiments.

In order to confirm what parameters have the highest effect on the model, a sensitivity analysis was performed. The sensitivity of each variable was tested by removing it from the selected networks with the best performance. When a variable has a marginal effect on the output amount, the model, after the removal of the variable, will still show a high correlation between the results and the actual data and its error amount will be low. Five variables were selected as follows: BOD, TSS, DO, Cl, and turbidity. The results of the sensitivity analysis had the highest effect on the output of the model.

The results of the present study indicated that the variables, which had the least impact on the model's performance, could be found and eliminated from the input data, so that the performance of the model was not significantly undermined or even more accurate results

Table 4. Results of Different Models of Artificial Neural Network Along With Their Structure

MAE	NRMSE	R	Processing Elements	Hidden Layers	Transfer	Learning Rule	Input PEs	Model
4.77	0.11	0.95	10	1	TanhAxon	Momentum	13	1
2.95	0.09	0.95	4	1	TanhAxon	Momentum	5	2
2.88	0.11	0.97	4	1	TanhAxon	Momentum	2	3

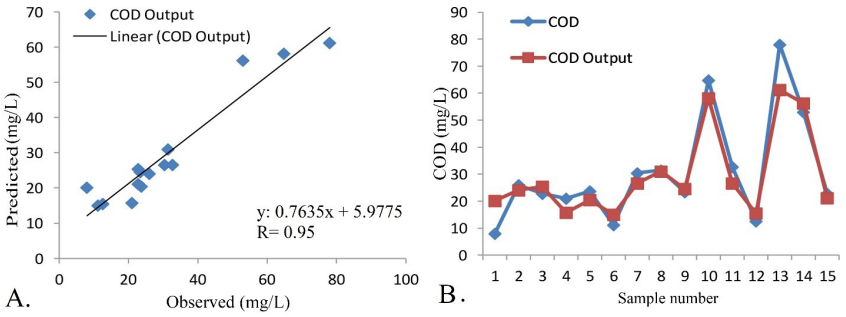

Fig. 3. Overlaping Graphs and Scatter Plots of the Results With the Actual Results in a Model With 13 Inputs, a Momentum Training Function and a TanhAxon Actuator Function.

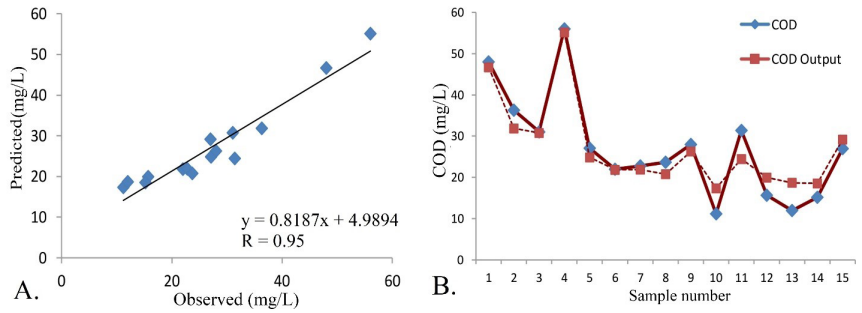

Fig. 4. Overlaping Graphs and Scatter Plots of the Results With the Actual Results in a Model With 5 Inputs, a Momentum Training Function and a TanhAxon Actuator Function.

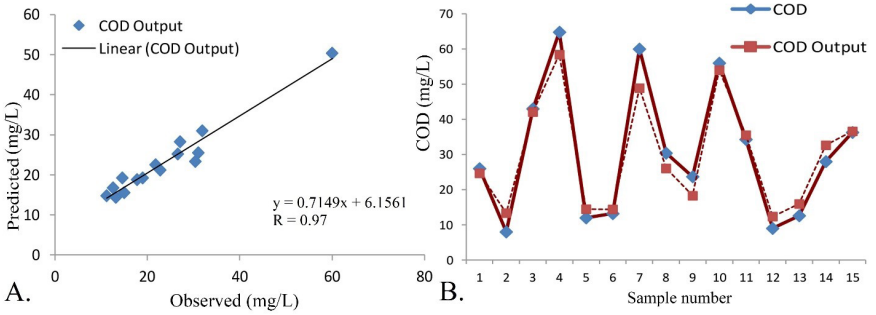

Fig. 5. Overlaping Graphs and Scatter Plots of the Results With the Actual Results in a Model With 2 Inputs, a Momentum Training Function and a TanhAxon Actuator Function.

were achieved. These results corroborated the results of a study by Ruben et al (11). Moreover, the results of the present study agreed with those of a study by Talib and Amat (9) in that BOD was one of the most effective parameters in COD modeling. Similar to the studies by Ay and Kisi (10), Talib and Amat (9), and Ruben et al (11), this study showed the high ability of the ANN to predict and model the COD of the river.

4. Conclusion

The findings illustrated that the Gamasiab River receives industrial effluents more than the standard limit:

sometimes the levels of COD are extremely higher than the acceptable limit. Thus, it is imperative to establish management and protective programs for this river. In this regard, in order to maintain the river water quality, the application of modeling methods for all variables affecting the water quality is inevitable. According to the results of various tests and assessment of the correlation between the results of models and the actual data and evaluation of various errors for choosing the structure of the neural network in each model, it was found that the ANN shows the best results based on the data, affirming the presence of correlations between them. Furthermore, the main upside of ANN in this study was the low number of input variables to the model. This model could present acceptable results with only two input parameters. In this study, by using the model, the number of the parameters was reduced from 13 to 2, which kept the accuracy of the predicted values still high.

Conflict of Interest Disclosures

The authors declare that they have no conflict of interests.

References

1. Milovanovic M. Water quality assessment and determination of pollution sources along the Axios/Vardar River, Southeastern Europe. Desalination 2005 2007; 213(1):159–73. Doi: 10.1016/j.desal.2006.06.022.
2. Singh KP, Basant A, Malik A, Jain G. Artificial neural network modeling of the river water quality—A case study. Ecolo Modell. 2009;220(6):888–95. Doi:10.1016/j.ecolmodel.2009.01.004.
3. Zhao X, Huang X, Liu Y. Spatial autocorrelation analysis of Chinese inter-provincial industrial chemical oxygen demand discharge. Int J Environ Res Public Health. 2012;9(6):2031–44. Doi:10.3390/ijerph9062031
4. Jia J, Jian H, Xie D, Gu Z, Chen C. Multi-perspectives' comparisons and mitigating implications for the COD and NH3-N discharges into the wastewater from the industrial sector of China. Water. 2017; 9(3):201.
5. Sarkar A, Pandey P. River Water Quality Modelling Using Artificial Neural Network Technique. In: International Conference on Water Resources, Coastal and Ocean Engineering (ICWRCOE'15) 2015; 4:1070–7.
6. Frost WH, Streeter HW. A Study of the Pollution and Natural Purification of the Ohio River. II. Report on Surveys and Laboratory Studies. U.S. Government Printing Office; 1924.
7. Yu L, Salvador NN. Modeling water quality in rivers. Am J Appl Sci. 2005;2(4):881–6. Doi:10.15666/aeer/1401_383395.
8. Bowers JA, Shedrow CB. Predicting stream water quality using artificial neural networks. Miscellaneous series. Westinghouse Savannah River Co; 2000.
9. Talib A, Amat MI. Prediction of chemical oxygen demand in Dondang River using artificial neural network. International Journal of Information and Education Technology. 2012; 2(3):259. Doi:10.7763/IJIET.2012.V2.124
10. Ay M, Kisi O. Modelling of chemical oxygen demand by using ANNs, ANFIS and k-means clustering techniques. J Hydrol. 2014;511:279–89. Doi:10.1016/j.jhydrol.2014.01.054.
11. Ruben GB, Zhang K, Bao H, Ma X. Application and sensitivity analysis of artificial neural network for prediction of chemical oxygen demand. Water Resources Management 2018; 32(1):273–83.
12. Stergiou C, Siganos D. Neural networks 2008. Available from: https://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol4/cs11/report.html.
13. Braddock RD, Kremmer ML, Sanzogni L. Feed-forward artificial neural network model for forecasting rainfall run-off. Environmetrics 1998; 9(4):419–32.
14. Asadpour G, Nasrabadi T. Municipal and medical solid waste management in different districts of Tehran, Iran. Fresenius Environmental Bulletin 2011;20(12):3241–5.
15. Zare AH, Bayat VM, Maroufi S, Amiri CR. Evaluation of artificial neural network and adaptive neuro fuzzy inference system in decreasing of reference evapotranspiration parameters. Journal of Water and Soil 2010;24(2):297–305.