INTELLIGENT MULTIVARIATE MODEL FOR THE OPTICAL DETECTION OF TOTAL ORGANIC CARBON

TAHER AHMADZADEH KOKYA[1], NASER MEHRDADI[1], MOJTABA ARDESTANI[1], AKBAR BAGHVAND[1], ARASH KAZEMI[2], ARAM A. M. KALHORI[3]*

[1] *Department of Environmental Engineering, Faculty of Environment, University of Tehran, Tehran, Iran.*
[2] *Electronics Laboratory, Array Computers Co., Tehran, Iran.*
[3] *Global Change Research Group, San Diego State University, San Diego, CA, USA.*

ABSTRACT

UV inactivity and fluorescence irradiance of various organic substances are the major drawbacks for a wide applicability of UV based TOC assessment models, especially in drinking water utilities and environmental fields. The adoption of an intelligent model is the key factor to access a reliable and effective detection. The accurate training of the artificial neural network model and backward elimination of less significant parameters, conferred more predictive properties to TOC detection. This led to an efficient optimal TOC detection model based on turbidity, UV_{254} absorbance and true color. The validation of model performance was investigated through application of untrained scenarios. The outcome of the validation analysis showed a correlation coefficient of 0.87 and root mean square error of 0.48 while the training performance of the model showed 0.95 and 0.33 respectively. The results indicated that the trained ANN model was efficiently capable for TOC detection in water resources based on the main drivers.

Keywords: Total Organic Carbon; Modeling; Artificial Neural Network; UV_{254}; Color; Turbidity

1. INTRODUCTION

Surface water quality largely depends on the extent of industrial and agricultural activities as well as natural land uses in the area. The river systems are most adversely affected due to their dynamic nature and accessibility to waste disposal through drains and tributaries. In the last few decades increased anthropogenic effluents into the rivers and reduced water flow have caused many-fold increase in the organic pollution load of the surface water bodies [1].

In general, the organic pollution of an aquatic system is measured and expressed in terms of the Total Organic Carbon (TOC). The TOC measures an approximate amount of natural and anthropogenic organic matter in water body and serves as an indicator parameter for the extent of water pollution. Detection and monitoring of TOC as the main source of disinfection by-products (DBPs) is critical for the environmental regulators and drinking water suppliers as well as water resources protection plan managers [2].

TOC is highly related to the dissolved organic matter (DOM) in water, and the high values of the earlier indicate for a high level of the dissolved organic carbon (DOC). However, it is common to see the terms TOC and DOC used interchangeably [3].

Currently available methods for the determination of TOC concentration are tedious or prone to the measurement errors [4]. Laboratory techniques for TOC detection include two major classes: (1) Standard analytical methods that are critically time consuming, user experience dependent and evidently vulnerable to the interferences of other existing chemicals especially chloride ion; (2) Instrumental TOC analyzing methods which are costly and rather difficult to maintain and use. In a typical TOC analyzer, sample undergoes an inorganic carbon removal process, combustion process and subsequent CO_2 detection. These processes need to be taken out with care, for instance, inorganic carbon removal step is not entirely selective as it also affects the carbon in the remaining phase, possibly because of the presence of volatile organic substances in sample [5].

Online monitoring and field techniques for TOC detection are usually based on simple UV-TOC regression models. Although, detection of TOC surrogates such as UV_{254} is relatively fast and simple to maintain but suffers from inaccuracy issues that stem from nonlinear nature of UV-TOC correlation.

The present work tried to address the mentioned drawbacks by modification of an artificial neural network model for TOC assessment and demonstrate its application on limited water quality data to show how it can improve the interpretation of the results. The ANN model approach has several advantages over traditional semi-empirical models, since they require known input data set without any assumptions [6]. Moreover, the application of artificial neural network to spectrophotometric determination of challenging chemical substances is known to be very efficient [7]. The ANN model develops a mapping of the input and output variables, which can subsequently be used to predict as a function of suitable inputs making it very popular in handling various water quality problems [8-17].

The real-life environmental problems are very complex and highly dependent on several process configurations, different influent characteristics and various other conditions [18]. Successful application of ANN based models for environmental problems in past decades stands for its reliability, robustness and adjustability [16-18]. These properties stem from the ability of learning complex nonlinear relationships within multiple variables particularly in situations where the explicit form of the relations is unknown [19].

TOC surrogates nonlinear modeling shows more adoptability to extensive variations of organic carbon concentration in comparison to less reliable UV-TOC regression models. For the first time, artificial neural network modeling was chosen as a TOC assessment optimization tool in water resources. In this case, the possibility of a neural net model training has been investigated to predict TOC as the secondary attribute of primary water quality variables. The results should serve as a more reliable TOC detection and monitoring method for its better accuracy, low cost, and instantaneous nature.

2. Materials and Methods

2.1. Artificial Neural Network Modeling

ANN model is basically comprised of three distinctive layers; the input layer - where the data are introduced and the weighted sum of the input is computed; the hidden layer(s) - where the data are processed; and the output layer - where the results are produced. Each layer consists of one or more basic element(s) called a node. A node is a non-linear function, parameterized with boundary values [20]. The signal passing through the node is modified by weights and transfer functions. This process is repeated until the output layer is reached [21]. The number of the nodes in the input, hidden and output layers is application dependent and should be taken large enough to provide sufficient degree of freedom [22].

There are many types of neural network based on the architecture and training algorithm used. In current study, a three-layer feed-forward neural network with error back propagation learning was applied to predict the TOC on the bases of several input variables (Fig. 1).

The feed-forward neural network architecture is proved to be very responsive for the prediction of water resources variables [23,24]. The error back propagation training was provided by Broyden-Fletcher-Goldfarb-Shanno (BFGS) training algorithm which is well suited for the unconstrained nonlinear multi-dimensional problems [25]. This algorithm provided a fast and efficient back propagation neural network training by adaptively modifying the initial search direction to improve the training efficiency [26].

2.2. Data

The sampling station for water quality data is located near Muncie at the White River upstream, starting south of Winchester in Randolph County at 40° 04’ 46” N, 84° 55’ 58” W. The river which was called Wapahani by the local

e-mail: T.ahmadzadeh@ut.ac.ir; Taher.tak@gmail.com

Indians, is currently the main water resource of several communities located along the 502 km of the river path [27]. The average annual river flow was 5.7 m^3s^{-1} and average annual precipitation was approximately 890 mm [28,29].

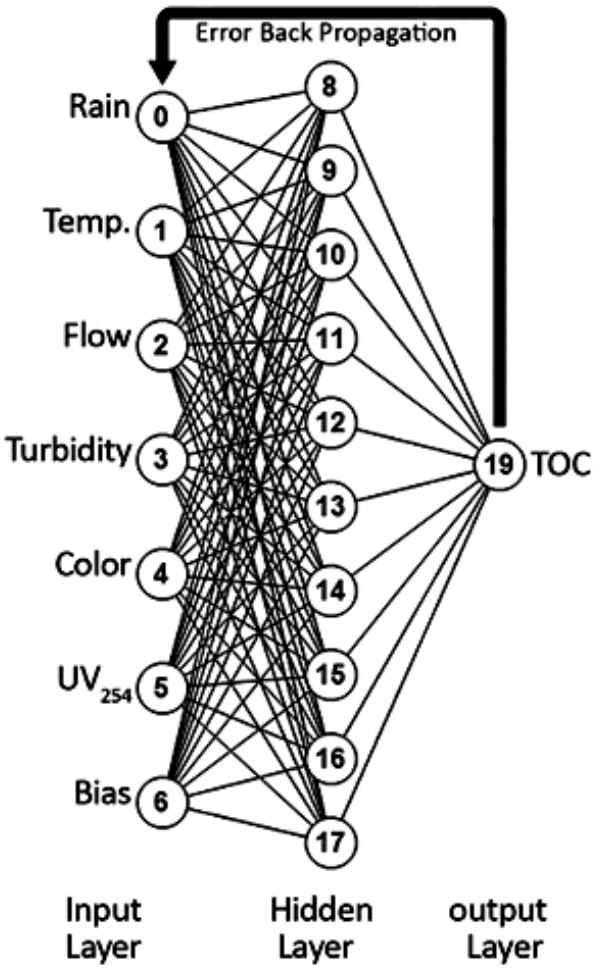

Fig. 1: Feed-forward artificial neural network with error back propagation learning.

A total of 657 daily data from August 1998 to June 2000 including rainfall, water temperature, river flow, turbidity, true color, UV_{254} and TOC concentration were studied as the input parameters for training and validation phases. Hydrological parameters such as rainfall, temperature and river flow are known to affect the water TOC content which in turn affects the gratitude of other studied parameters [30,31].

River water sample was filtered through preheated glass fiber filters (Whatman GF/F) for the removal of large particles before any analytical measurement. TOC concentrations was measured in triplicate, using a Sievers 800 TOC analyzer with an inorganic carbon removal module (Ionics-Sievers Instruments, Inc., Boulder, CO). UV absorbance at 254nm and true color at 455nm were determined with a DR/4000 spectrophotometer (Hach Company, Loveland, CO), and turbidity (NTU) was measured with a SS6/SE turbidimeter (Hach Company, Loveland, CO). Water temperature was measured daily with a calibrated thermometer and daily precipitation was measured with a calibrated rain collector. River discharge was recorded continuously at the U.S. Geological Survey gauging station 03347000 on the White River at Muncie [28].

The data set was divided into two parts for model training and validation phases in the order of 6:1. Therefore 563 samples were selected for the training phase and 94 samples for the validation phase. The summary of the collected data is shown in Table 1.

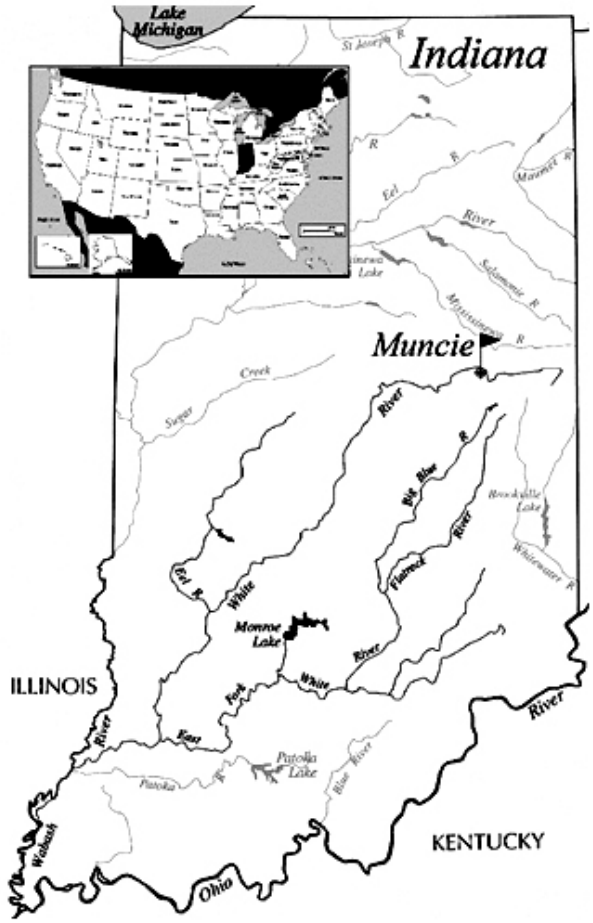

Fig. 2: Water sampling station was located at the White River upstream, Muncie, IN.

3. RESULTS AND DISCUSSION

3.1. Model Optimization

Model optimization is an important part of the cost effective modeling. This was performed to provide the best predictive model with the least number of input parameters. Prediction accuracy of an ANN model is highly dependent to the training algorithm as well as the number of nodes per layer and the corresponding transfer functions [32-34]. Most of these factors were chosen through multiple examination of model to provide a best fit model for the data sets. Further optimization was performed through backward elimination of less significant variables.

Backward elimination method was conducted for decreasing the number of input variables while keeping the best probable coefficient of determination (R^2) for the TOC prediction. However, there were no predefined criteria for identification of the significance of parameters. Therefore, the method was executed repeatedly with an omitted variable each time and comparing the model performance coefficients. This led to a minimum of three optically measurable input variables. Modified parameters of the full and optimal ANN model are summarized in Table 2.

Table 1 Summary of data used for model training and validation phases.

	Parameters	*Rainfall (mm)*	*Water temperature (°C)*	*River flow (m^3s^{-1})*	*Turbidity (NTU)*	*True color (Pt-Co)*	*UV_{254} (abs.)*	*TOC ($mg\ l^{-1}$)*
Training	Mean value	2.3	13.25	7.5	15.53	11.74	0.11	3.90
	Max. value	53.3	26.48	237.8	277.96	50.78	0.73	11.31
	Min. value	0.00	2.63	0.5	1.75	2.61	0.04	2.13
	RSD %	254.7	50.9	2.2	151.3	53.7	66.1	27.4
Validation	Mean value	1.00	13.36	7.5	15.92	11.26	0.11	3.86
	Max. value	15.34	25.44	167.5	105.29	32.61	0.52	7.57
	Min. value	0.00	2.74	0.5	1.75	3.47	0.04	2.18
	RSD %	247.38	51.02	2.1	131.43	50.99	71.86	25.25

Table 2 Summary of optimum ANN model parameters.

Parameter	*Full model values*	*Optimal model values*
Input variables	Water Temperature, River Flow, Rainfall, UV_{254}, Color, Turbidity	UV_{254}, Color, Turbidity
Number of training Iterations	50000	100000
Number of input nodes	7	4
Number of hidden nodes	15	10
Number of output nodes	1	1
Learning rate	0.3	0.5
Activation steepness	1	1
Hidden Layer Activation function	Elliot	Elliot
Output Layer Activation function	Sigmoid, Stepwise	Elliot Symmetric

This process minimized the model to 3 input variables while keeping the coefficient of determination as high as possible. The prediction capability of the optimized model is shown in Figure 3 as how it competes with the full model in a 1:1 diagram.

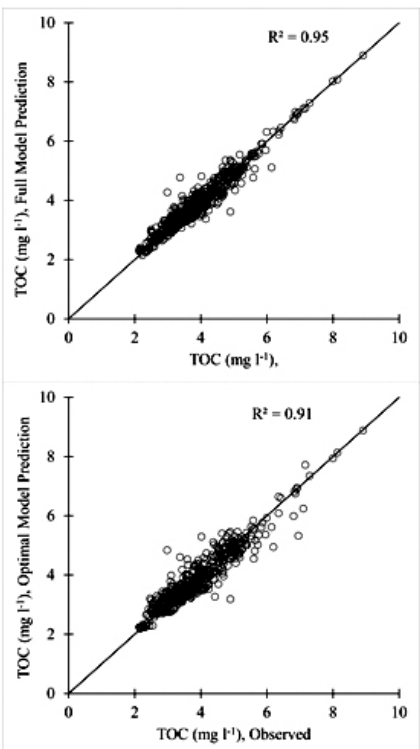

Fig. 3: Performance of optimized model in 1:1 diagram.

3.2. Model Performance

Model performance and prediction accuracy was investigated as the various correlation limiting factors. Statistical analysis was performed on both preliminary full model and the reduced size model. The correlation coefficients of predictions and relevant errors were repeatedly calculated during each training and validation phases to get the most accurate probable prediction. Though, the best results of the modeling performance indicators are summarized in Table 3.

According to Table 3 there is a quite competing results for the optimized model with only 3 input factors in comparison with the full model in training phase which was remarked by comparable R and R_s values. Relatively high values of the coefficient of determination (R^2) indicated reliable predictive characteristics which provide enough evidence for the significance of the input parameters for TOC assessment.

Table 3: Summary of ANN model performance in training and validation phases.

	Model performance coefficients	*TOC (Full model)*	*TOC (Optimal model)*
Training	Pearson Correlation coefficient (R)	0.97	0.95
	Coefficient of determination (R^2)	0.95	0.91
	Spearman correlation coefficient (R_s)	0.95	0.92
	Mean Absolute Percentage Error (MAPE) %	5.04	6.23
	Root Mean Square Error (RMSE)	0.25	0.33
Validation	Pearson Correlation coefficient (R)	0.87	0.87
	Coefficient of determination (R^2)	0.75	0.75
	Spearman correlation coefficient (R_s)	0.90	0.92
	Mean Absolute Percentage Error (MAPE) %	7.21	7.29
	Root Mean Square Error (RMSE)	0.50	0.48

The validation results showed some expected loss of accuracy. Knowing the fact that validation scenarios were completely new for the model to predict on the basis of training information, the predicted results are in a very good agreement with the actual observations. It is noteworthy that the optimized model actually did a better job in validation phase with the higher spearman coefficient (0.92 vs 0.9) and lower RMSE (0.48 vs 0.50). The performance of both models in the validation phase was shown in Figure 4 as the model prediction versus the actual observation in a 1:1 diagram.

Although the higher predictive capabilities of the ANN model usually stand for its higher accuracy, these capabilities seem to be a result of the model adjustability in the application of two more surrogate parameters – Color and Turbidity – in addition of UV_{254} in the prediction model.

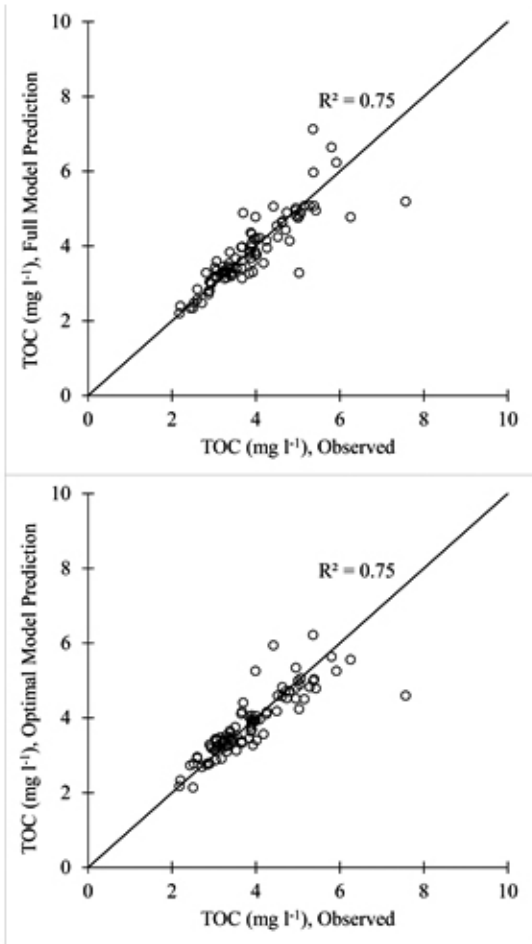

Fig. 4: Predicted vs. observed TOC concentrations in ANN model's validation phase.

Long term prediction of ANN model and the relevant prediction error is presented in Fig. 5. The performance of the ANN model for predicting TOC variations was reasonable. However, the model showed some difficulties in predicting peak values which is shown in peak values of TOC. This is apparently because the ANN model were trained with 563 days of data which include only 10 days with TOC concentration over 7 mg l^{-1}.

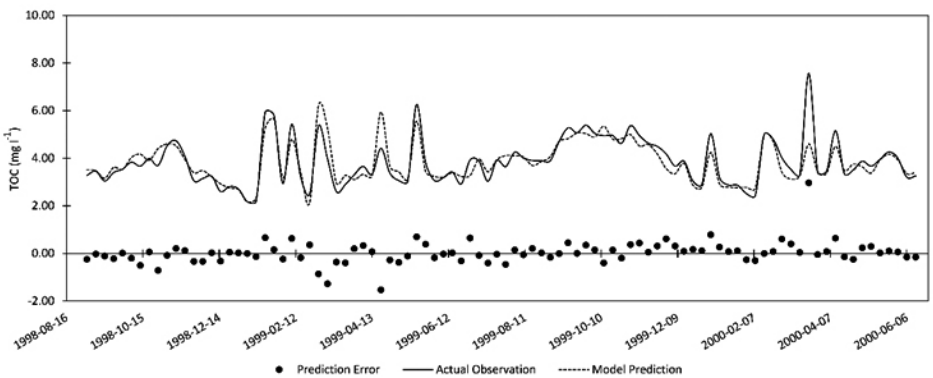

Fig. 5: Predicted vs. observed TOC concentrations during the studied time scale.

A descriptive comparison of the specified method in the present work with common TOC determination methods is shown in table 4. A detailed information on the applicability and characteristics of compared methods is well describe by Matilainen and Bisutti [5,35].

Table 4: Comparison of common TOC detection methods with the proposed method[+]

Detection Method	*Sample Treatment*	*Intended Use*	*TOC Modeling*	*Monitoring Compatibility*	*Cost Efficiency*	*Time Efficiency*	*Environment Friendly*	*User Friendly*	*Reliability*	*Accuracy*
Titration of Excess $Cr_2O_7^{2-}$	Wet Oxidization by $Cr_2O_7^{2-}$	Laboratory	Linear Regression	L	H	L	L	L	H	H
Colorimetric Detection of Excess $Cr_2O_7^{2-}$	Wet Oxidization by $Cr_2O_7^{2-}$	Laboratory	Linear Regression	L	M	L	L	L	H	H
IR Detection of CO_2	Wet Oxidization by Persulfate	Laboratory	Linear Regression	L	L	M	M	L	L	H
Thermal Conductivity Detection of CO_2	Wet Oxidization by Persulfate	Laboratory	Linear Regression	L	L	M	M	L	L	H
Detection of Fluorescence Irradiance	No Particular treatment	Laboratory	Linear Regression	M	M	M	H	M	M	L
Detection of UV_{254} Absorbance	No Particular treatment	Field & Laboratory	Linear Regression	H	H	H	H	H	M	L
This work; Detection of UV_{254}, Color, Turbidity	No Particular treatment	Field & Laboratory	ANN (Nonlinear)	H	H	H	H	H	H	M

[+] L: Low, H: High, M: Moderate

4. CONCLUSIONS

An intelligent TOC detection model for water resource monitoring was developed using artificial neural networks. The results indicated that the modified model converged rapidly during the training phase and the performance of the ANN model in variations of TOC was reasonable. The predicted TOC concentrations is in good agreement with the observed values with the correlation coefficient of 0.87 and root mean square error of 0.48. According to the results, the ANN model reasonably balanced the cost and the accuracy of TOC detection through measurement of three optically measurable surrogate parameters – UV_{254}, Color and Turbidity. As a matter of fact, ANN supported spectrophotometry showed to be a reliable alternative for TOC analysis. The results are critical for water monitoring systems and drinking water suppliers for fast, low cost, and maintenance-free TOC monitoring. The proposed method is a suitable support system for other TOC analytical methods, indicating outlier data and instrumental failure or can be further exploited to examine the effects of other challenging water quality parameters such as DBPs.

ACKNOWLEDGEMENTS

Authors gratefully acknowledge Dr. Christian Volk for providing water quality data used in model training and validation. Authors also acknowledge many beneficial conversations with Prof. Khalil Farhadi and Mr. Bahman Ahmadzadeh. This project was funded by the laboratory of KIMIA AB, Urmia.

REFERENCES

1.- K.P. Singh, A. Malik, D. Mohan and S. Sinha, J. Water Res. 38(18), 3980–3992 (2004).
2.- Y. Hou, W. Chu and M. Ma, J. Environ. Sci. 24(7), 1204–1209 (2012).
3.- R. Beckett and J. Ranville, in Interface Science in *Drinking Water Treatment: Theory and Applications*, Edited by G. Newcombe and D. Dixon (Elsevier Ltd, London, 2006).
4.- G. Visco, L. Campanella and V. Nobili, Microchem. J. 79, 185–191 (2005).
5.- Matilainen, E.T. Gjessing, T. Lahtinen, L. Hed, A. Bhatnagar and M. Sillanpää, Chemosphere 83, 1431–1442 (2011).
6.- M.W. Gardner and S.R. Dorling, J. Atmos. Environ. 32, 2627–2636 (1998).
7.- M.R. Moghadam, A.M.H. Shabani and S. Dadfarnia, J. Hazard. Mater. 197, 176–182 (2011).
8.- I.O. Bucak and B. Karlik, Ekoloji 20(78), 75–81 (2011).
9.- P. Kulkarni and S. Chellam, J. Sci. Tot. Environ. 408(19), 4202–4210 (2010).
10.- E. Dogan, A. Ates, E. C. Yilmaz and B. Eren, Environ. Prog. 27(4), 439–446 (2008).
11.- E. Dogan, B. Sengorur and R. Koklu, J. Environ. Manag. 90(2), 1229–1235 (2009).
12.- P.S. Kunwar, A. Basant, A. Malik and G. Jain, Ecol. Model. 220(6), 888–895 (2009).
13.- P.S. Kunwar and S. Gupta, Chemomet. Intell. Lab. Sys. 114, 122–131 (2012).
14.- Q. Cong, W. Yu and T. Chai, in *Advances in Computational Intelligence vol.61*, edited by W. Yu, E.N. Sanchez (Springer–Verlag, Berlin, 2009).
15.- Najah, A. El-Shafie, O.A. Karim, O. Jaafar and A.H. El-Shafie, Int. J. Phys. Sci. 22(6), 5298–5308 (2011).
16.- A.R. Khataee, M. Zarei and M. Pourhassan, Clean 38(1), 96–103 (2010).
17.- E.R. Rene and M.B. Saidutta, Int. J. Environ. Res. 2(2), 183–188 (2008).
18.- K. Yetilmezsoy and S. Demirel, J. Hazard. Mater. 153, 1288–1300 (2008).
19.- K. Yetilmezsoy, B. Ozkaya and M. Cakmakci, Neural Net. World 11(3), 193–218 (2011).
20.- G. Dreyfus, J.M. Martinez, M. Samuelides, M. B. Gordon, F. Badran and S. Thiria, L. Hérault, *Reseaux de Neurones - Méthodologie et applications*, (Eyrolles, Paris, 2002).
21.- R.S. Govindaraju, J. Hydrol. Eng. 5(2), 124–137 (2000).
22.- J.M. Ortiz-Rodríguez, M.R. Martínez-Blanco, J.M. Cervantes Viramontes and H.R. Vega-Carrillo, in *Artificial Neural Network-Architectures and Applications*, edited by K. Suzuki (InTech, Rijeka, 2013)
23.- R.C. Schweitzer and J.B. Morris, United States Army Research Laboratory Report No. ARL–TR–2155, 2000.
24.- S. Palani, S. Liong, P. Tkalich and J. Mar. Pollut. Bull. 56(9), 1586–

1597 (2008).
25.- J. Nocedal and S.J. Wright, *Numerical Optimization*, 2nd ed. (Springer–Verlag, New York, 2006).
26.- N.M. Nawi, M.R. Ransing and R.S. Ransing, Presented at the Sixth International Conference on Intelligent Systems Design and Applications IEEE, Jinan University, China, 2006 (unpublished).
27.- L. Sonneborn, *Chronology of American Indian History* (Infobase Publishing, New York, 2007).
28.- C. Volk, L. Wood, B. Johnson, J. Robinson, H. W. Zhuc and L. Kapland, J. Environ. Monit. 4, 43–47 (2002).
29.- C. Volk, L.A. Kaplan, J. Robinson, B. Johnson, L. Wood, H.W. Zhu and M. LeChevallier, Environ. Sci. Technol. 39(11), 4258–4264 (2005).
30.- I. Delpla, A.V. Jung, E. Baures, M. Clement, O. Thomas, Environ. Inter. 35, 1225–1233 (2009).
31.- J.K. Edzwald, G.S. Kaminski, J. NEWWA, 123(1), 15–31 (2009).
32.- T.Y. Lin and C.H. Tseng, J. Eng. App. Art. Intell. 13, 3–14 (2000).
33.- J.A. Frenie, A. Jiju, Work Study 50(4), 141–149 (2001).
34.- T.T. Soong, *Fundamentals of probability and statistics for engineers*, (John Wiley & Sons Inc., New York, 2004).
35.- I. Bisutti, I. Hilke, M. Raessler, Trends Anal. Chem. 23(1), 10–11 (2004).