International Journal of ChemTech Research
CODEN (USA): IJCRGG ISSN: 0974-4290
Vol.8, No.3, pp 1282-1289, 2015

Application of Multivariate Linear Regression and Neural Network in the Assessment of Groundwater Quality

Sarala Thambavani D[1], Uma Mageswari T.S.R [2]*

[1]Sri Meenakshi Government Arts College for Women (Autonomous), Madurai, Tamilnadu, Research and Development Centre, Bharathiar University, Coimbatore, India
[2]PSNA College of Engineering & Technology, Dindigul, Tamilnadu, India

Abstract: This paper examined the efficiency of multivariate linear regression (MLR) and artificial neural network (ANN) models in prediction of three major water quality parameters such as electrical conductivity, total alkalinity and nitrate concentrations. Results showed that using measured parameters is convenient to model these three parameters with acceptable and appropriate accuracy. ANN and MLR methods are able to predict electrical conductivity, total alkalinity and nitrate concentration at the desirable level of accuracy. Comparison of ANN analysis with MLR model results showed that ANN requires fewer parameters with more accuracy in comparison to MLR models. Performance of the ANN models was evaluated using coefficient of correlation (r) and root mean square error (RMSE). The computed values of the parameters by model, ANN method and regression analysis were in close agreement with their respective measured values. Results showed that the ANN performance model was better than the MLR model.
Key words : Electrical conductivity, Alkalinity, Nitrate, Multivariate Linear Regression, Artificial Neural Network.

Introduction

Groundwater is the major source of water supply in different cities around the world and therefore several studies have highlighted different aspects of groundwater such as, storage potential, hydrogeology, water quality, vulnerability and sustainability and so on[1,2,3].A variety of factors contribute to variations in groundwater quality. Their inherent uncertainty carries weight, when more than one variable affect quality of water. The in homogeneity of the medium has thrown the quality prediction and the approaches adopted by researches into complexity

The electrical conductivity, alkalinity and nitrate are the important water quality parameters for drinking and agricultural purposes. Kney and Brandes[4] hypothesized that alkalinity values can be used as an index of bedrock geology and that it can be expected that, under natural conditions, a particular range of electrical conductivity values will correspond to a particular range of alkalinity. Ionic pollutants from anthropogenic sources contribute to electrical conductivity, however, and it is this portion of the EC that should be of primary interest in monitoring and assessment. Kney and Brandes [4] suggest that anthropogenic impacts will result in a deviation in the relationship between electrical conductivity and alkalinity and it should therefore be possible to use concurrent alkalinity and electrical conductivity measurements to indicate anthropogenic impacts. Nitrates, being extremely soluble in water, move easily through the soil and into the ground water. Ingestion of excessive amounts of nitrates causes ill health effects in infants less than six months old and susceptible to adults. It causes “blue baby syndrome” or Methemoglobinemia in infants, which can lead to brain damage and sometimes death. Also, the Maximum Contaminant Level (MCL) for nitrates in public drinking

water established by the federal government is 10 mg l^{-1}. Agricultural activities and operations that disturb and aerate the soil enhance the soil nitrogen oxidation, which, along with the fertilizer nitrate components will be leached to the groundwater.

Artificial neural networks (ANNs) are able to approximate accurately complicated non-linear input-output relationships. The ANN is used as an approximation tool rather than a complex mathematical calculation, which results in a ten percent deviation of predicted value from observed data [5].There are a number of studies in which neural networks are applied to water quality problems. In the recent years many works have been reported in electrical conductivity modeling with ANN due to its ability in modeling complex non-linear problems. For instance, Tutmez[6] attempted Neuro-fuzzy modeling of electrical conductivity variation as a function of concentration of dissolved solids (such as sodium, potassium, calcium and magnesium) whose variation in space can be considered as non-linear. Similarly, Najah[7] employed a complex 2 hidden layer neural network architecture for electrical conductivity modeling in the study. Very little work has been done so far in building stochastic models to predict alkalinity in groundwater using regression and neural networks.

The ANNs have been widely used in various studies on surface water pollution control for predicting stream nitrogen concentration [8], forecasting raw water quality parameters [9], prediction of water quality parameters, water quality management [10] and identification of non-point sources of microbial contamination [11,12]. Due to increased agricultural activity which is necessary for enhanced food production and also due to industrial activity, there is an increasing evidence of nitrate pollution of groundwater [13].

This study is to assess the effect of parameters in ANN technique on the prediction of water quality of Virudhunagar district, Tamil Nadu. The objective of this analysis is to enhance the effectiveness in parameter training and validation and to expand the application of ANN. These results can then be used as a baseline for monitoring the impacts on water quality that may occur in the future.

Materials and Methods

Study Area and Data Analysis

The present study was carried out in Virudhunagar District of Tamil Nadu State. It lies between Latitude $9^0$12'N to $9^0$45'N and Longitude $77^0$24'E to $78^0$18'E. The district is bounded by Sivagangai district and Madurai district on the North, Tirunelveli district and Tuticorin district to the South and Ramanathapuram district on East and Kerala state to the West and Theni district to the North West. The location of the study area is shown in Fig.1. Total area of Virudhunagar district is 4243.23 sq. km and the district is divided into 8 taluks. It has an average elevation of 102 m (335 ft) above sea level and is largely flat with no major geological formations. The district receives the rainfall under the influence of both southwest and northeast monsoons. The Northeast monsoon chiefly contributes to the rainfall in the district. The town has a humid climate and receives 780 mm rainfall annually. The relative humidity is on an average between 65 and 85%. Virudhunagar district is characterized by relatively high level of groundwater development in both hard rock and sedimentary aquifers. Occurrence, movement and storage of groundwater are influenced by lithology, thickness and structure of the rock formation. The presence of black clayey soil has resulted in reduced natural recharge to groundwater system. It has also resulted in water quality problem. The ground water samples collected from 8 taluks namely Aruppukkottai (S_1), Kariapatti (S_2), Rajapalayam (S_3),Sattur (S_4), Sivakasi (S_5), Srivilluputtur (S_6) and Tiruchuli (S_7) and Virudhunagar (S_8).

Fig.1 Map of the study area

Multivariate Linear Regression (MLR)

Statistical methods, such as regression models, are the best tools for investigating any relationship between dependent and independent variables of small sample size . The MLR is a method used to model the linear relationship between a dependent variable and one or more independent variables. MLR is based on least squares. In the best model, sum of square error between observed and predicted parameters should be minimum value. Electrical conductivity, Alkalinity and nitrate estimation also can be performed using linear models which explain linear relationship between parameters. Furthermore, the same input variables for MLR models are considered for linear models Eq.1

Y= a pH+ b TDS+ c DO+ d TH+ e TA+ f Ca+ g Mg+ h----------(1)
Where,
Y= Electrical conductivity, Alkalinity and nitrate,
a, b, c, d, e, f, g and h are constant coefficients of linear regression model,
pH, TDS, EC, TH, TA, Ca, Mg, DO are input parameters.

Artificial Neural Network (ANN)

The ANN models are increasingly being used for forecasting or simulating water resources variables because they are often capable to model complex systems with unknown or difficult behavioral rules or underlying physical processes. The ANN is a non-linear modeling tool capable of handling a large number of inputs (independent variables), to determine one or more outputs (dependent variables) . There are many types of neural networks for various applications available in researches. The multi layer perceptron (MLP) is a widely used ANN configuration and has been frequently applied in the field of hydrological modeling .Steps involved in ANN is given in Fig.2

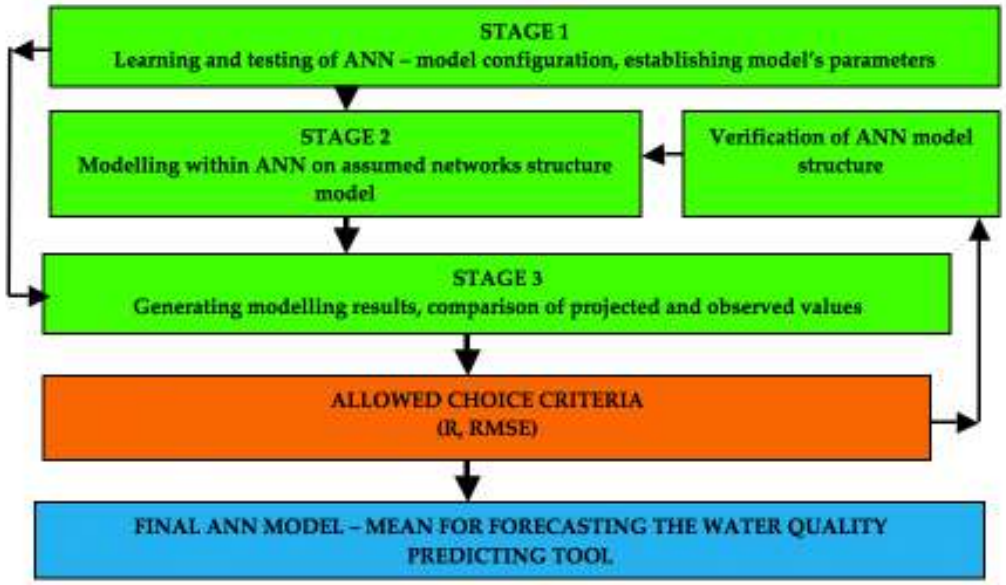

Fig.2 Steps involved in Artificial Neural Network

Fig. 3 provides an overview of the structure of this network. The MLP consists of three layers of neurons: (1) an input layer; (2) an output layer, and (3) intermediate (hidden) layer or layers. Each neuron has a number of inputs (from outside the network or the previous layer) and a number of outputs (leading to the subsequent layer or out of the network). A neuron computes its output response based on the weighted sum of all its inputs according to an activation function.

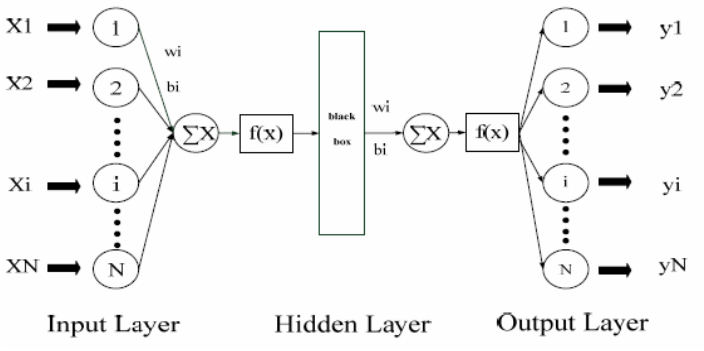

Fig.3 Typical Multilayer structure of ANN

The MLP is the simplest and therefore most commonly used neural network architectures. It is a network with seven input variables, a hidden layer with two to a maximum of ten processing neurons and three output variables (Electrical conductivity, Total alkalinity and nitrate). For a simple regression analysis the units in the input layer introduce normalized or filtered values of each input variable into the network, then these values are transferred to all units of the hidden layer multiplied by a " weight " factor that is, in general, different for every connection, and its magnitude characterizes the importance of some connection (Fig. 3).

The methodology of research is divided to four separate parts: description of data sets; ANN and MLR inputs; description of ANN characteristics of neurons, layers and in-out parameters; and method of sensitivity analysis. The choice of the type of network depends on the nature of the problem to be solved [14]. The number of input and output neurons is determined by the nature of the modeling problem, the input data representation and the form of the network output required. The number of hidden layers is related to the complexity of the system being modeled. Although some researchers suggest that one hidden layer is usually sufficient [15]. So in this study a three-layer ANN (input-output layers with a hidden layer) with Levenberg-Marquardt algorithm and a tan-sigmoid transfer function for the hidden layer and a linear transfer function for the output layer were used.

In order to achieve the research objective, samples were collected from the study area as shown in Fig.1. 150 groundwater samples were analyzed to model Electrical conductivity, alkalinity and nitrate contamination change. Seventy percent of the samples were used to train the ANN and develop MLR models, and the remaining 30% of data were used to evaluate the models. Samples were analyzed in the laboratory for the major ions using standard methods. The analyses were carried out within 48 hrs after sampling. Parameters of pH, electrical conductivity (EC), magnesium, chloride and calcium ionic concentrations, total dissolved solids, dissolved oxygen and total hardness were measured. Nitrates were measured using colorimetric method with an UV – visible spectrophotometer.

First, 7 parameters of water quality were used as a primary input of artificial neural network. For the selection of the most important artificial neural network input parameters the periodic remove method was used. Therefore, by eliminating any input parameter, the structure of optimized artificial neural network was run. The network sensitivity to any input parameter was calculated by comparing the neural network output and by eliminating any input parameter by the following equation:

$$PC = \frac{|X_1 X_2|}{X_1} X100 \quad \text{..... (2)}$$

Where *PC*– percent of change (%), X_1 – artificial neural network output with 7 input parameters, X_2 – output of artificial neural network with eliminating any input parameter. To select the most appropriate input parameters, different combinations of parameters were used to predict electrical conductivity, alkalinity and nitrate concentration. The combined parameters are given in Table 1. Two different forecast consistency measures of the root mean square error (R*MSE*) and the correlation coefficient (*r*) were used to evaluate models results and to compare ANN and MLR:

$$= \frac{\Sigma(Q_o - M_o)(Q_p - M_p)}{\sqrt{\Sigma(Q_o - M_o)^2 \Sigma(Q_p - M_p)^2}} \quad \text{---------- (3)}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}[(Q]_o - Q_p)^2}{n}} \text{----------(4)}$$

Where Q_o and Q_p are observed and predicted value. M_o and M_p are mean of observed and predicted values.

Table 1. Different combinations of input parameters in ANN and MLR models

S.No	Output Parameters	Combinations of input parameters
1	Electrical conductivity	pH, TDS, TH, TA, DO, Ca, Mg pH, TDS, TH, TA, Ca, Mg pH, TDS, TH, TA pH, TDS, TH
2	Alkalinity	pH, TDS, TH, EC, DO, Ca, Mg pH, TDS,TH,EC, Ca, Mg pH, TDS,TH,EC pH, TDS,EC
3	Nitrate	pH, TDS, EC,TH, TA, Ca, Mg pH, TDS, EC,TH, TA pH, TDS, EC,TH pH, TDS, EC

Results and Discussion

To predict electrical conductivity of ground water 7 parameters are selected. The parameters are pH, total dissolved solids, total hardness, dissolved oxygen, total alkalinity, calcium and magnesium ions. The best network structure is determined for forecasting electrical conductivity. To reduce the input parameters and to determine parameters with less influence on conductivity, sensitivity to each of the selected parameters was studied. Percentage of variation of calculated conductivity was determined for each index. As a result, four different combinations of input parameters were used to evaluate the accuracy measures of equations 2-4.The results are presented in Table 2. ANN1 structure with 7 parameters had a greater error than those of other structures, which means that increasing the number of input parameters is not always effective.ANN2 structure with 6 parameters was selected as the appropriate structure considering error and correlation coefficients. It is clearly noted in Table 2 that the proposed ANN 2 model has impressively well learned the nonlinear relationship between the input and the output variables with r=0.9963 and RMSE= 0.095. The same input parameters were used in MLR models as independent variables for electrical conductivity modeling. The results are presented in Table 2. The best regression model includes 7 independent variables (MLR1).This structure is able to estimate electrical conductivity concentration with an error of 0.6708 and with correlation coefficient of 0.852.Comparison of two selected models, ANN2 and MLR1 is less accurate. Fig.4 shows the results of ANN2 structure of electrical conductivity.

Table 2. Evaluation of the ANN structures and MLR models for Electrical conductivity

Structure	Input parameters	ANN		MLR	
		r	RMSE	r	RMSE
ANN1/MLR1	pH, TDS, TH, TA, DO, Ca, Mg	0.984	0.1581	**0.852**	**0.6708**
ANN2/MLR2	pH, TDS,TH,TA, Ca, Mg	**0.999**	**0.095**	0.727	0.7211
ANN3/MLR3	pH, TDS,TH,TA	0.979	0.1317	0.653	0.8426
ANN4/MLR4	pH, TDS,TH	0.966	0.1378	0.612	0.8544

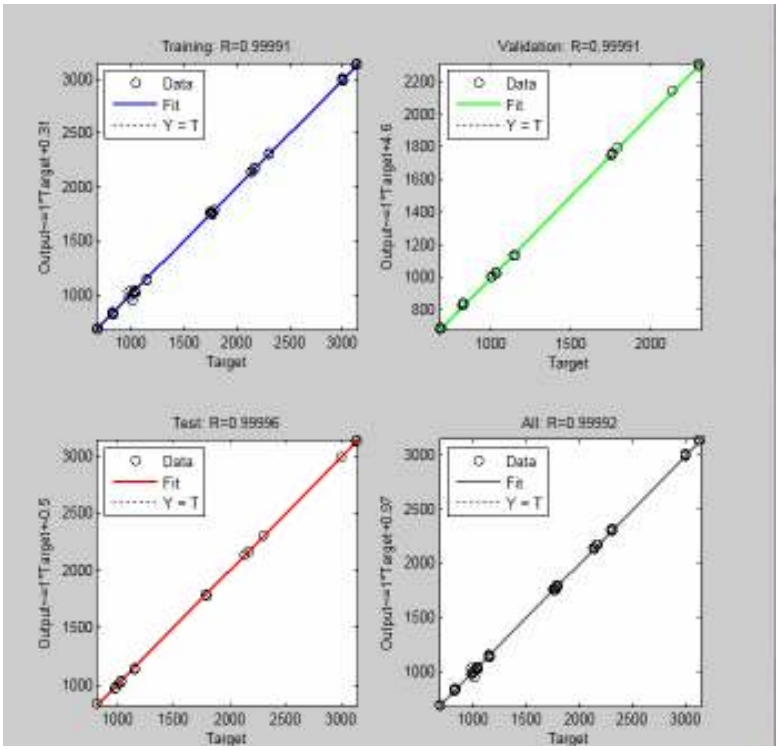

Fig.4 Evaluation of ANN 2 results of electrical conductivity

To predict the groundwater alkalinity concentration 7 parameters are used as input variables. The parameters are pH, total dissolved solids, total hardness, dissolved oxygen, electrical conductivity, calcium and magnesium ions. Table 3 clearly noted that ANN 2 and MLR1 structures are best model to predict alkalinity concentration with high correlation coefficient and low root mean square error. Fig.5 shows the results of ANN2 structure of alkalinity.

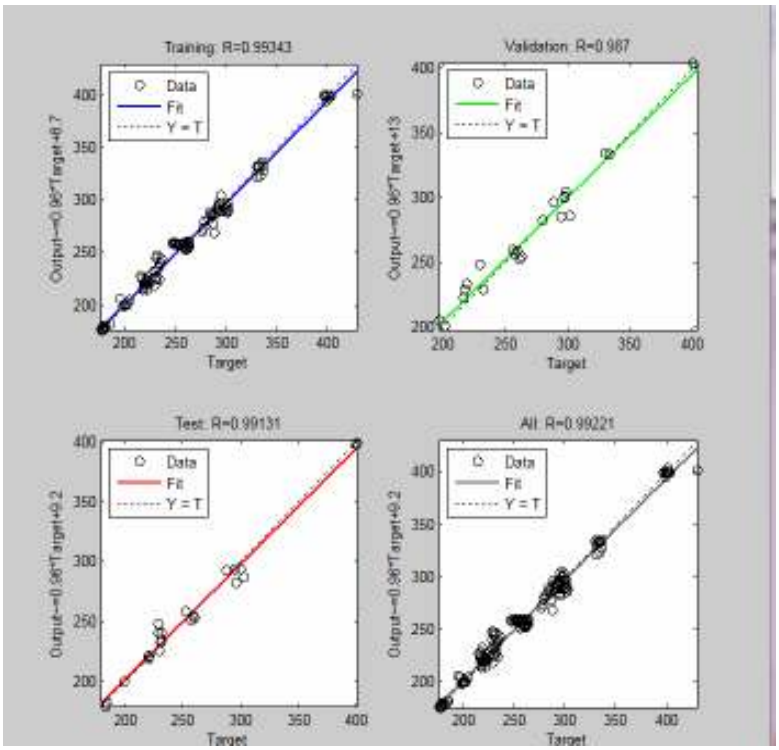

Fig.5 Evaluation of ANN 2 results of Alkalinity

Table 3. Evaluation of the ANN structures and MLR models for Alkalinity

Structure	Input parameters	ANN		MLR	
		r	RMSE	r	RMSE
ANN1/MLR1	pH, TDS, TH, EC, DO, Ca, Mg	0.987	0.2915	**0.872**	**0.8307**
ANN2/MLR2	pH, TDS,TH,EC, Ca, Mg	**0.992**	**0.270**	0.702	0.9219
ANN3/MLR3	pH, TDS,TH,EC	0.982	0.2983	0.700	0.9110
ANN4/MLR4	pH, TDS,EC	0.980	0.02811	0.681	0.9592

To predict the nitrate concentration 7 parameters are used as input variables. They are pH, total dissolved solids, total hardness, dissolved oxygen, total alkalinity, electrical conductivity, calcium and magnesium ions. Table 4 clearly noted that ANN 4 (r=0.83,RMSE=1.249) and MLR1 (r=0.55,RMSE=2.315) structures are best model to predict nitrate concentration. Fig.6 shows the results of ANN4 structure of nitrate.

Table 4. Evaluation of the ANN structures and MLR models for Nitrate

Structure	Input parameters	ANN		MLR	
		R	RMSE	r	RMSE
ANN1/MLR1	pH, TDS, EC,TH, TA, Ca, Mg	0.771	1.797	**0.558**	**2.315**
ANN2/MLR2	pH, TDS, EC,TH, TA	0.784	1.609	0.523	2.362
ANN3/MLR3	pH, TDS, EC,TH	0.793	1.425	0.510	2.551
ANN4/MLR4	pH, TDS, EC	**0.814**	**1.249**	0.497	2.634

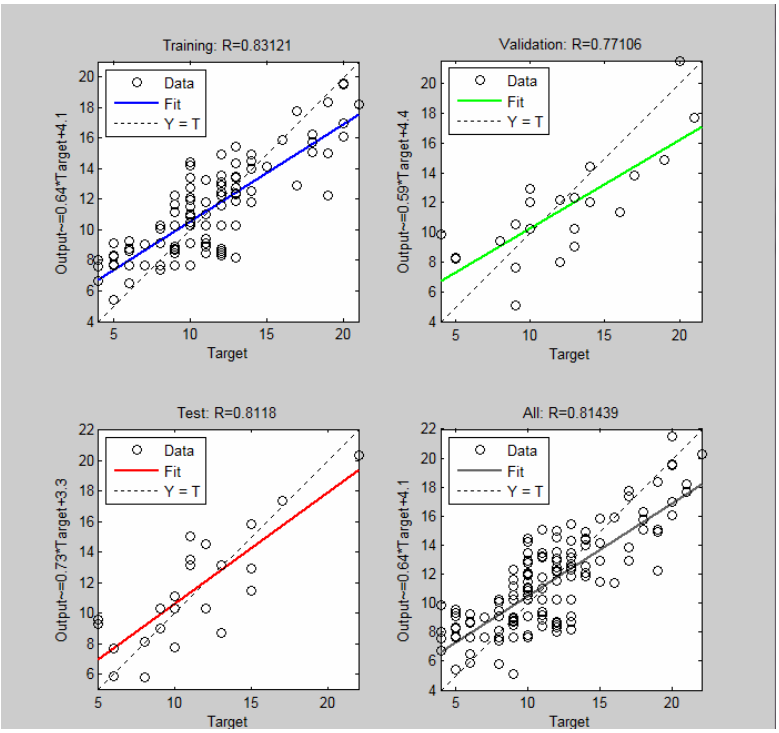

Fig.6 Evaluation of ANN 4 results of nitrate concentration

From the above results, it is understood that the reducing input variables of MLR models decreases model accuracy, where as for ANN it increases. However, the main limitations of statistical techniques are the rigid assumptions that are essential for justifying their applications, such as those of sample size, linearity and continuity. The main advantage of reduction in ANN input parameters are computational economy, decrease in

computation time and cost. This point is confirmed on estimation of free water evaporation and reference evapo transpiration (Wang *et al.*, 2008). In above mentioned studies using minimum parameters, the neural network to predict the unknown parameter (output) was investigated.

Conclusion

In this study, electrical conductivity, total alkalinity and nitrate concentration were simulated by ANN and MLR techniques using seven water quality variables as the input to the models. Systematic or hierarchical models of ANN and MLR were developed by trimming the most complicated network and using backward selection procedure in order to investigate the significant input variables and their contribution order. Results of the study indicated that the reducing input variables of MLR models decreases model accuracy, where as for ANN it increases. Thus artificial neural network (ANN) models need fewer parameters for prediction of parameters compared to MLR models. This result suggests that the use of more input parameters will not necessarily lead to improvements of predicted results, but type of input parameters is more important. Moreover, because of all of their advantages, ANNs are easy and practical to apply from site to site. Their fast execution should also be helpful for simulation of electrical conductivity, total alkalinity and nitrate concentration on a large scale.

References

1. Pandey, V.P.; Kazama. F.;Environmental Earth Sciences, 2011,62,. 1723 – 1732.
2. Pandey, V.P.; Shrestha, S.;Chapagain, S.K.;Kazama, F.;Environmental Science & Policy, 2011 14. 396 – 407.
3. Chapagain, S.K.; Pandey, V.P.; Shrestha, S.; Nakamura, T.;Kazama, F.;Water Air and Soil Pollution, 2010, 210, 277 – 288
4. Kney, A.D.;Brandes. D.;J. Environ. Manage. 2007,82,519–528
5. Lingireddy. S. and Ormsbee L.E., Neural Networks in Optimal Calibration of Water Distribution Systems, Artificial Neural Networks for Civil Engineering: Advanced Features and Applications (Eds I. Flood, R. Kartam). ASCE Press, New York, USA.(1998)
6. Tutmez, B.;Hatipoglu, Z.;Kaymak, U.;Computers and Geosciences.2006,32, 421 – 433
7. Najah, A.;Elshafie, A.; Karim, O.A.;Jaffar, O.;European Journal of Scientific Researach,2009,28, 422 – 435.
8. Lek, S.;Maritxu. G.;Giraudel, J.;Wat. Res., 1999, 33, 3469-78.
9. Zhang. Q;. Stephen, J.S.;Water Res.,1997, 31, 2340-2351.
10. Wen, C.G.; Lee, C.S.;Water Resour. Res., 1998, 34, 427-436.
11. Brion. G.M.; Lingireddy. S.; 2000. Academic Press, London, UK.2000
12. Zaheer. I.; Cui. G.;J. Environ. Hydrol.,2003,11, 1-8.
13. Prakasa Rao, E.V.S.; Puttanna, K.;Current Sci.,2000, 79, 1163-1168.
14. Goethals, P.;Dedecker, A.P.;Gabriels, W.;Lek, S.;Pauw, N.;Ecol.,2007, 41, 491-508.
15. El-Din, A.G.; Smith, D.W.;Water Res.,2002,36, 1115-1126.

*****