Forecasting Models for Some Water Quality Parameters of Shatt Al-Hilla River, Iraq

Rafa H. Al-Suhili
College of Engineering, University of Baghdad/ A Visiting Prof. to the City College, NY, USA
rafeaalsuhili@yahoo.com
Nesrin J. Al-Mansori
College of Engineering, University of Babylon
nassrin20052001@yahoo.com

Abstract

This paper provides Artificial Neural Networks model versions for forecasting the monthly averages of some chemical water quality parameters of Shatt Al-Hilla River, which is located at Hilla City, south of Iraq. The water quality parameters investigated were Sulphate, Magnesium, Calcium, Alkalinity, and Total Hardness. Results indicate that for Sulphate and Calcium high correlation coefficients models were observed to be (0.9 and 0.88), while for Magnesium, Alkalinity and Hardness low correlation coefficients model were observed to be (0.48,0.58, and 0.51) respectively. Serial correlation behavior of these variables indicate at that high lag time correlations sequences are observed for the first two variables and low ones for the last three water quality parameters. A serial correlation coefficient analysis was done and indicates that as the variable exhibited weak lag correlation structure, then a successful ANN forecasting model could not be obtained even if many trials were done to enhance it's performance, such as increasing the number of nodes, the lagged input variables, and/or changing the learning rate and the momentum term values, or the use of different types of activation functions. On the other hand, those variables that have a strong lag correlation structure can easily fit successful ANN forecasting models.

Keywords: Artificial Neural Networks ; water quality parameters; forecasting models ; Prediction efficiency; Turbidity; Alkalinity.

الخلاصة

يتناول هذا البحث التنبؤ بالمعدلات الشهرية لبعض المحددات الكيميائية لمياه شط الحلة الواقع في محافظة بابل بأستخدام الشبكات العصبية الصناعية، المحددات المستخدمة في الموديل الكبريت، المغنيسيوم، الكالسيوم، القاعدية ، العسرة الكلية. بينت النتائج بأن الكبريت والكالسيوم يمتاز بمعامل ارتباط عالي عند تطبيق الموديل (0.8,0.9) ... بينما المغنيسيوم والقاعدية والعسرة كان معامل الارتباط (0.48, 0.58,0.51) على التوالي. من الملاحظ انه قيمة الارتباط العالي للمتغيرين الأوليين ولبقية العناصر بقيم واطئة ..كان بسبب محددات الوقت . نجاح تحليل الموديل بتنبؤ بالقيم بأستخدام ANN لا يمكن اجراءه حتى بكثرة عدد المحاولات ..وانما بزيادة عدد النقاط او تغيير معدل التعلم او تغيير قيمة الزخم او استخدام نوع اخر من المعادلات الفعالة .من جهة أخرى بناء هذا الموديل بأستخدام ANN كان مطابق للتخمين للقيم للمحددات الكيميائية للمياة.

الكلمات المفتاحية: الشبكات العصبية الصناعية ، محددات نوعية المياه، موديل التخمين، كفاءة التنبؤ، العكورة ، القاعدية.

Introduction

Water quality analysis of any source of water explains the suitability of this water to sustain various uses, such as drinking, irrigation and other uses. Most of studies, designs, and planning concerning the chemical properties of raw water requirement forecasts of future values . Hence, providing models for estimating expected future values of these chemical properties may enhance the decision-making about the system investigated. These values are essential, for example, for an efficient treatment plant planning, design and operation.

Artificial neural networks (ANNs) models successfully applied in many engineering fields including water quality analysis. Relative recent researches have reported that ANN forecasting models might offer a promising alternative of Auto - regressive forecasting models.

AL-Suhaili and Karim (2014)used different versions of artificial neural networks models for forecasting the daily inflows to Dokan Reservoir, located in the north of Iraq. The models developed proved the ability of such models to produce reliable forecasts, and to reproduce an excellent high and low persistency in the

forecasted series as compared to the persistency existed in the observed series, for different threshold values.

Al-Suhaili and Gafour (2013) developed an artificial neural networks model to predict the sodium adsorption ratio (SAR) for Tigris River near the city of Ammarah, located south of Iraq. These models capable of producing excellent predictions.

Gazzaz *et.al.*,2012, described the design and application of feed-forward, fully-connected, three-layer perceptron neural network model for computing the water quality index (WQI) for Kinta River (Malaysia). The analysis of the results indicated that the optimal network architecture was 23-34-1 as input, hidden and output layers nodes and that the best training algorithm was the quick propagation (QP); with a learning rate of 0.06; and a QP coefficient of 1.75. The correlation coefficient between the measured and predicted values was ($r = 0.977$, $p < 0.01$), this P-value indicate that the null hypothesis was accepted, i.e, there was no significant differences between the observed and predicted WQI values.

Palani *et.al.*,2008, used artificial neural networks (ANNs) to predict and forecast quantitative characteristics for Singapore coastal waters. The advantages of the ANN modeling process are due to its ability to (1) represent both linear and non-linear relationships and (2) obtain these relationships by the training process and the model parameters corrections using an optimization algorithm utilizing the input-output sets of the data being model. The water quality parameters included in the model were salinity, temperature, dissolved oxygen, and chlorophyll. A time lag ANN model up to two time lags degree found to be sufficient to yield reliable predicted results. The validation of the performance of the trained ANN model achieved by applying it to a data set from a station in the region that not used in the model building process. The results showed that the ANN's great potential to simulate water quality variables, with a determination coefficient of 0.9.

The importance of the ANN forecasting models was due to its capability to forecast future values of any time evolution variable using the historical record of this variable only. The need for other variables is usually essential for the development of other types of forecasting models. As in most cases in Iraq, other data are usually not available and/or available for short periods and at limited locations. Hence, the use of the artificial neural networks models for forecasting was more justified. However, the main disadvantage of the ANN forecasting models that the used of successive time lagged values of the variable forecasted as the input for forecasting its expected value for time steps a head is that it required a high lag correlation behavior. This implies that the degree of randomness included in this variable should be low.

ANN Modeling Technique

ANN modeling techniques are well known now and proved its capability to model different engineering problems. This modelling technique can represent the non-liner relationship among the input and output variables of any system,(Mustafa *et.al*.,2012). It consists of a grouped neurons or nodes in layers. The input layer neuron represented the input variables, while the output layer nodes represent the output variables. For a typical three layers ANN were existed between these input and output layers, a hidden layer with a certain number of hidden nodes. The nodes between the layers interconnected by weights. The input layer nodes received the input values and transmitted its liner weighted combination with an added bias term to the hidden nodes, where it processed with a suitable activation function to produce an output, from each node in the hidden layer. These outputs combined using weights between the hidden layer and the output layer, which received by the output layer nodes, and processed by an activation function to produce outputs. The process explained above called feed-foreword process.

In ANN, modelling the set of data required was the input variables with their corresponding output variables. In order to find weights of the model, the network trained using a partial set of the data. Hence the original data set to be divided to training, testing and hold out sub-sets. The training process performed by using the training sub-set and assuming initial weights, which were corrected during this training process. The input data was subject to a feed-foreword process to produce output data using the assumed weights. These outputs compared with the corresponding real ones and errors estimated. These errors used to adjust weights using certain algorithm such as back propagation (BP). Fig. (1) shows a typical three-layer ANN structure.

The artificial neural networks model matrices equations as presented by (Al-Suhaili and Ghfour ,2013) are as follows:

$$Zin\{p,1\} = Vo\{p,1\} + V\{n,p\}^T X\{n,1\} \quad (1)$$

where X is the standardized input vector, n is the number of nodes in the input layer, Vo is bias matrix of the input layer, V is the weight matrix between the input and hidden layer, p is the number of nodes in the hidden layer, and Zin is the input vector to the nodes in the hidden layer, and

$$Zout\{p,1\} = F1(Zin\{p,1\}) \quad (2)$$

where F_1 is the activation function of the hidden layer, and Zout is the output vector from the nodes of the hidden layer, and

$$Yin\{m,1\} = Wo\{m,1\} + W\{p,m\}^T Zout\{p,1\} \quad (3)$$

$$Yout\{m,1\} = F2(Yin\{m,1\}) \quad (4)$$

Where Yin is the vector input to the nodes of the output layer, Wo is the bias vector of the output layer, W is the weight matrix between the hidden and output layer, m is the number of nodes in the output layer, Yout is the standardized output vector, and F2 is the activation function of the output layer.

The statistical procedures for social sciences (SPSS, Version 20) software used herein to find the forecasting ANN models. This software initially assumed the model matrices parameters, Vo, V, Wo, ,W, and the number of nodes in the hidden layer p for a given activation functions F1, and F2, then estimated the errors of the forecasted outputs using the four matrix equation shown above, compared with the observed values and adjusted the assumed matrices using an optimization algorithm such as steepest gradient algorithm in a process called training until obtaining the best matrices and the best value of p. These matrices may be the model parameters that used in equations (1) to (4) for forecasting the output variable.

Case Study

In this research, ANN models developed for forecasting some of the monthly chemical parameters for raw water at Shatt AL-Hilla River. This river is the only surface water source in the area and which runs exactly in the middle of the Hilla City,100 km (62 mi) south of Baghdad City, between Longitude (44^0 26- 65" & 44^0 31-00") E and Latitude (32^0 25- 30" & $32^0$31-30") N.(Lafta and Naief, 1999).

Shatt Al-Hilla is the main channel that branches from the left side of the Euphrates River upstream of the Al-Hindiya Barrage. It irrigated an area of about 617500 km^2.The total length of Shatt Al-Hilla is about (101 km.) beginning from its head regulator(Al-Hindiya Barrage) throughout Babylon Governorate to the border of Al-Muthana Governorate. The maximum designed discharge of Shatt AL-Hilla is ($220m^3$/sec. with water level 31.30 m.a.s.l.) with hydraulic gradient of 8 cm/km (Babylon Water Resources Department/BWRD,2011) .Fig.(2), shows the location of this river relative to Iraq.

The data used for developing the models development were the measured monthly series for the period (2000-2013).The water quality parameters are the Sulfate (So_4), Magnesium (Mg), and Calcium (Ca), Alkalinity ($CaCO_3$), Total Hardness (HCO_3). Table (1) shows the descriptive statistics of the water quality parameters used in this study, for the monthly series data sample for the above-mentioned period.

Results and Discussions

The SPSS software used for obtaining the parameters of the ANN modelversions. Table (2) shows the number of nodes in the hidden layer, the activation functions and the correlation coefficients obtained for the model versions of Sulfate (So_4). These results obtained after many trials of data division to training, testing, and holding out sub-sets, different selections of the activation functions types, different selections of the learning rates and momentum terms. The results shown which gave the highest correlation coefficient between the predicted and the observed monthly input variables at time (t). It is clear that high correlation coefficients obtained for the first five model versions. Among the first five models versions the fifth one gave the highest correlation coefficient of 90% hence this version selected. Moreover, it is clear that the best activation function types for all of the five versions were the hyperbolic tangent and the identity functions for the hidden and output layer respectively. The largest correlation coefficient was for the fifth version with one node in the hidden layer. Introducing further lags increased the required number of nodes in the hidden layer with slight increase in the correlation coefficient; hence, the fifth version considered as the best one due to the highest correlation coefficients and relatively low number of hidden nodes required.

Table (3) shows the number of nodes in the hidden layer, the activation functions and the correlation coefficients obtained for the model versions of Magnesium (Mg). The results shown which gave the highest correlation coefficient between the predicted and the observed monthly input variables at time (t).

Among the seven models versions the sixth one gave the highest correlation coefficient of 48.0% which was relatively low, this may be due to the weak serial correlation structure of this variable .Moreover it is clear that the best activation function types for all of the six versions were the hyperbolic tangent and the sigmoid functions for the hidden and output layer ,respectively. The largest correlation coefficientis for the sixth version with four node in the hidden layer ; hence the sixth version was considered the best one due to the highest correlation coefficients in spite of increasing the number of hidden nodes required.

Table (4) shows the number of nodes in the hidden layer, the activation functions and the correlation coefficients obtained for the model versions of Calcium (Ca). The results shown which gave the highest correlation coefficient between the predicted and the observed monthly input variables at time (t). It is clear that high correlation coefficients obtained for the third model versions. Among the four model versions, the third one gave the highest correlation coefficient of 87.6% hence this version was selected. Moreover, it is clear that the best activation function types for all of the four versions are the hyperbolic tangent and the sigmoid functions for the hidden and output layer respectively. The largest correlation coefficients at third version with two node in the hidden layer ; hence the third version was considered the best one due to the highest correlation coefficients in spite of increasing the number of hidden nodes required.

Table (5) shows the number of nodes in the hidden layer, the activation functions and the correlation coefficients obtained for the model versions of Alkalinity ($CaCO_3$). The results shown which gave the highest correlation coefficient between the predicted and the observed monthly input variables at time (t). It is clear

that high correlation coefficients obtained for the fourth model versions. Among the fifth model versions, the fourth one gives the highest correlation coefficient of 52.9%, which were relatively low. Moreover, it is clear that the best activation function types for all of the five versions were the hyperbolic tangent and the identity functions for the hidden and output layer respectively. The largest correlation coefficients at four version with two node in the hidden layer ; hence the fourth version was considered the best one due to the highest correlation coefficients in spite of closing the number of hidden nodes required from other models.

Table (6) shows the number of nodes in the hidden layer, the activation functions and the correlation coefficients obtained for the model versions of Hardness (HCO_3). The results shown which gave the highest correlation coefficient between the predicted and the observed monthly input variables at time (t). It is clear that high correlation coefficients obtained for the fourth model versions. Among the fifth model versions, the fourth one gave the highest correlation coefficient of 57.9%, which was relatively low. Moreover, it is clear that the best activation function types for all of the five versions were the hyperbolic tangent and the identity functions for the hidden and output layer respectively. The largest correlation coefficientis atfour version with one node in the hidden layer ; hence the fourth version was considered the best one due to the highest correlation coefficients in spite of decreasing thenumber of hidden nodes required from other models.

The above results indicated that the ANN forecasting model was successful for Sulphate and Calcium only, and failed for Magnesium, Alkalinity and Hardness. To investigate the expected reason for this, as mentioned above the variables of the failed models may have weak serial correlation structure, a correlation study was done. Table (7) shows the serial correlation coefficients for the variables investigated herein up to lag 5. Figure (3), shows some values it where that Sulphate and Calcium exhibited strong lag correlation behavior, while the other three variables had a weak one.

Conclusions

Many version of forecasting artificial neural networks models tried herein for some water chemical parameters of Shatt Al-Hilla, located at south of Iraq. These models used the lag one, two, three, four, five, six and seven preceding monthly-measured values for each of these parameters. It found that as the variable exhibited a relatively high correlation, and then it more expected to get an ANN forecasting model with accepted performance, as observed for Sulphate and Calcium. Variables with weak serial correlation structure like Magnesium, Alkalinity, and hardness, failed to fit a forecasting model using ANNs, even if the number of hidden nodes increased, time lagging increased, and different activation functions and wide variations of the learning rate and momentum terms used.

References

Al-Suhaili R.H., Karim R. 2014: “Daily inflow forecasting for Dukan reservoir in Iraq using artificial neural networks", International Journal of water.

Al Suhaili, R.H. and Ghafour, Z.J. 2013, ‘Artificial neural netwok model to predict the sodium adsorption ratio for Tigris River at Ammara’, Journal of Engineering, College of Engineering, University of Baghdad, March, Vol. 19, No. 4.

Gazzaz N. M.,Yosoff M. K., Aris A Z., and Juahir H, 2012 "Artificial neural network modeling of the water quality index for Kinta River (Malaysia) using water quality variables as predictors ", Marine Pollution Bulletin, 08,1-12.

Palani S, Liong S. Y., and Tkalich P., 2008, "An ANN application for water quality forecasting ",Marine Pollution Bulletin 56 ,1586–1597.

Mustafa, M.R., Isa, M.H. and Rezaur, R.B.,2012 ,'Artificial neural networks modeling in water resources engineering: infrastructure and applications', World Academy of Science, Engineering and Technology, Vol. 6, No. 2, pp.317–325.

Lafta and Nayef,1999. Hydrochemistry of ground water in Hillaregion , Babylon University,Vol.4..

BWRD Babylon Water Resources Department, 2011.

Table:1 Descriptive Statistics for the Investigated Water Quality Parameters of Shatt Al-Hilla River (2000-2013).

	N	Range	Minimum	Maximum	Mean	Variance	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
SO_4	168	408.00	175.00	583.00	306.1071	8241.917	1.443	.187	1.800	.373
Mg	168	46.00	12.00	58.00	35.0893	78.609	-.053-	.187	.467	.373
Ca	168	100.00	40.00	140.00	95.4881	465.928	-.653-	.187	-.198-	.373
Alkalinity	168	94.00	80.00	174.00	127.2500	435.278	.191	.187	-.594-	.373
Total Hardness	168	562.00	178.00	740.00	383.5298	8209.700	-.044-	.187	2.233	.373

Table 2: Correlation Coefficients for the Developed ANN Model Versions for Sulfate (SO_4) model.

Model version Inputs	*Number of hidden nodes in the hidden layer*	*Activation function*		*Correlation coefficient*
		Hidden layer	*Output layer*	
It–1	*2*	Tansh	Identity	*0.836*
It–1, It–2 1	*1*	Tansh	Identity	*0.839*
It–1, It–2, It–3	*1*	Tansh	Identity	*0.841*
It–1, It–2, It–3, It–4	*2*	Tansh	Identity	*0.838*
It–1, It–2, It–3, It–4, It–5	*1*	Tanch	Identity	*0.901*

Table 3: Correlation Coefficients for the Developed ANN Model Versions for Magnesium (Mg) model.

Model version Inputs	*Number of hidden nodes in the hidden layer*	*Activation function*		*Correlation coefficient*
		Hidden layer	*Output layer*	
*It–1	1	Tansh	Sigmoid	0.395
*It–1, It–2 1	2	Tansh	Sigmoid	0.364
*It–1, It–2, It–3	2	Tansh	Sigmoid	0.392
*It–1, It–2, It–3, It–4	3	Tansh	Sigmoid	0.350
*It–1, It–2, It–3, It–4, It–5	3	Tanch	Sigmoid	0.469
*It–1, It–2, It–3, It–4,It–5,I-6	4	Tansh	Sigmoid	0.480
*It–1, It–2, It–3, It–4, It–5,I-6,I-7	4	Tanch	Sigmoid	0.441

Table 4: Correlation Coefficients for the Developed ANN Model Versions for Calcium (Ca) Model.

Model version Inputs	*Number of hidden nodes in the hidden layer*	*Activation function*		*Correlation coefficient*
		Hidden layer	*Output layer*	
*It–1	1	Tansh	Sigmoid	0.739
*It–1, It–2 1	2	Sigmoid	Sigmoid	0.759
*It–1, It–2, It–3	2	Tansh	Sigmoid	0.765
*It–1, It–2, It–3, It–4	3	Tansh	Sigmoid	0.876

Table 5: Correlation Coefficients for the Developed ANN Model Versions for Alkalinity ($CaCO_3$) Model.

Model version Inputs	*Number of hidden nodes in the hidden layer*	*Activation function*		*Correlation coefficient*
		Hidden layer layer	*Output*	
*It–1	2	Tansh	Identity	0.501
*It–1, It–2 1	1	Tansh	Identity	0.508
*It–1, It–2, It–3	3	Tansh	Identity	0.502
*It–1, It–2, It–3, It–4	2	Tansh	Identity	0.592
*It–1, It–2, It–3, It–4, It–5	4	Tanch	Identity	0.577

Table 6: Correlation Coefficients for the Developed ANN Model Versions for Total Hardness (HCO_3) Model.

Model version Inputs	*Number of hidden nodes in the hidden layer*	*Activation function*		*Correlation coefficient*
		Hidden layer layer	*Output*	
*It–1	2	Tansh	Identity	0.447
*It–1, It–2 1	4	Tansh	Identity	0.540
*It–1, It–2, It–3	3	Tansh	Identity	0.523
*It–1, It–2, It–3, It–4	1	Tansh	Identity	0.579
*It–1, It–2, It–3, It–4, It–5	2	Tanch	Identity	0.510

Table 7: Serial Correlation Coefficients of the Water Quality Parameters.

	lag1	lag2	lag3	lag4	lag5
So_4	0.816	0.744	0.686	0.583	0.492
Mg	0.403	0.169	0.059	0.08	0.067
Ca	0.712	0.645	0.513	0.487	0.449
Alkalinity	0.491	0.273	0.225	0.007	0.024
Total Ha.	0.35	0.209	0.212	0.0431	0.0345

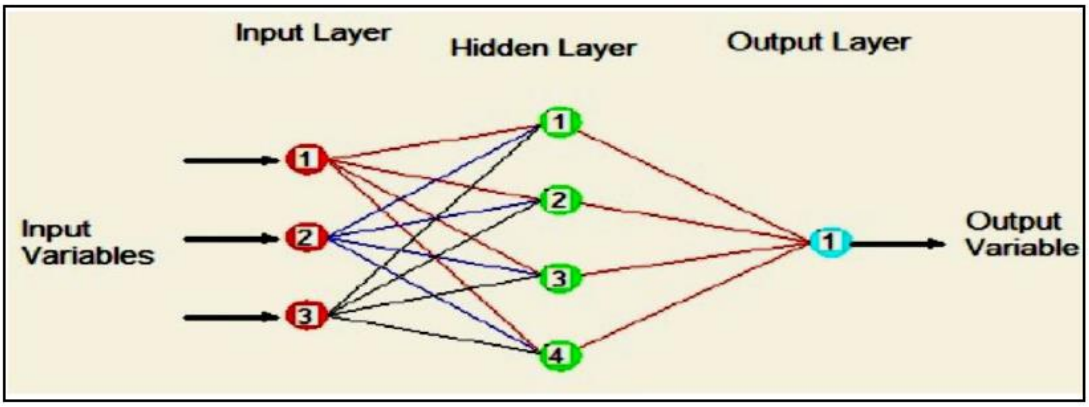

Figure 1 Three Layers Architecture of an Artificial Neural Network.

Figure.2 : Shatt Al-Hilla, Iraq.

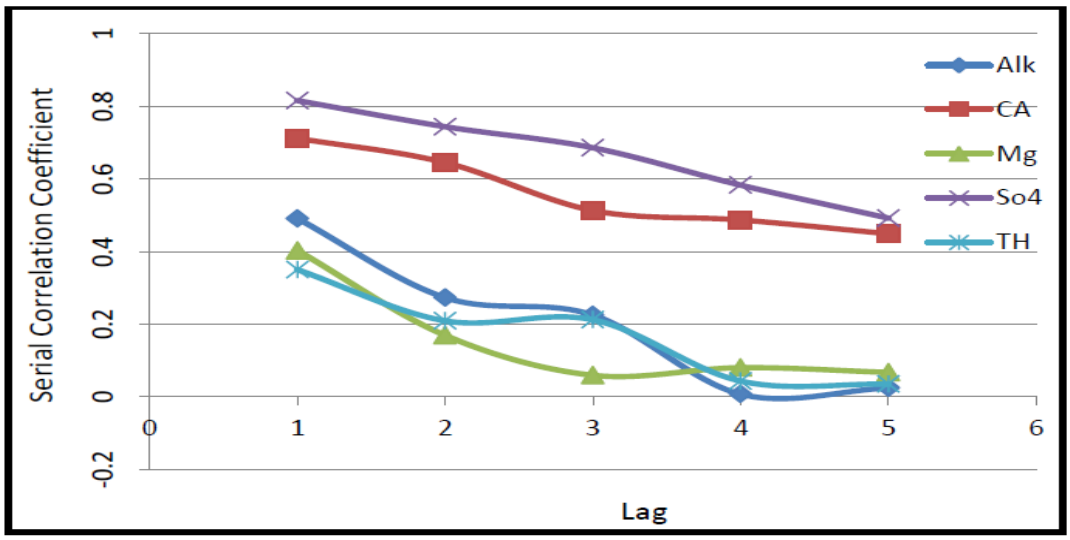

Figure. (3) Serial Correlation Coefficients of the Water Quality Parameters Investigated.