Research Article www.ijcps.com

International Journal of Chemical and Pharmaceutical Sciences
2014, Mar., Vol. 5 (1)

ISSN: 0976-9390
IJCPS

Back propagation neural network in predicting DO and BOD

[1] Sarala Thambavani D[*] and [2] Uma Mageswari TSR.

[1] Sri Meenakshi Government Arts College for Women (Autonomous), Madurai, Tamilnadu
Research and Development Centre, Bharathiar University, Coimbatore.
[2] PSNA College of Engineering & Technology, Dindigul, Tamilnadu
[*] **Corresponding Author:** E-Mail: sarala_dr@yahoo.co.in

ABSTRACT

This study examined the potential of Multilayer Perceptron Neural Network (MLP-NN) in predicting dissolved oxygen (DO) and Biochemical oxygen Demand (BOD) at Batlagundu, Dindigul District, Tamilnadu. The choice of input parameters based on statistical correlation analysis is the most popular analytical technique for selecting input. Based on the existing measured values and statistical analyses, the following water quality parameters are selected for the Artificial neural network (ANN) modeling in this study, namely pH, total dissolved solids (TDS), electrical conductivity (EC), total hardness (TH), calcium (Ca) ions, magnesium (Mg) ions, and total alkalinity (TA). The water quality parameters are monitored regularly during different seasons of summer, rainy and winter during the period of two years 2012 and 2013. To evaluate the performance of the proposed model, two statistical indexes were used; namely, mean Square Error (MSE) and Correlation Coefficient (CC). A relatively high correlation and low mean square error are obtained between the observed and predicted values in the testing data set.

Keywords: Dissolved Oxygen, Biochemical oxygen Demand, Multilayer Perceptron, Correlation Coefficient. Mean square error.

1. INTRODUCTION

Among the most important issues facing civilization in the 21st century is the growing scarcity of fresh and clean water. Maintaining and improving the quality and quantity of freshwater have long-term economic, health and ecological implications. Deterioration of water quality has waste disposal directly or indirectly has initiated serious management efforts in many countries. Most acceptable ecological and water related decisions are difficult to make without careful modeling, prediction and analysis of water quality for typical development population staying in the basins.

In general, the organic pollution in an aquatic system is measured and expressed in terms of the biochemical oxygen demand (BOD) and have increased the need for modeling techniques that can decline dissolved oxygen (DO) level. The DO level is the measure of the health of the aquatic system and certain level is essentially required for the aquatic life to survive. Measures of DO refer to the amount of oxygen contained in water and defined the living conditions for oxygen-requiring (aerobic) aquatic organisms. DO concentrations reflect equilibrium between oxygen producing processes (e.g. photosynthesis) and oxygen consuming processes (e.g. aerobic respiration, nitrification and chemical oxidation) and the rate of atmosphere exchange. BOD is an approximate measure of the amount of biochemical degradable organic matter present in a water sample. It is defined by the amount of oxygen required for the aerobic microorganisms present in the sample to oxidize the organic matter to a stable organic form. Thus, it's important to develop predictive DO and BOD model for the management of water quality. At present, there [4] are many mathematical models for assessing DO and BOD. For example, Ma *et.al*[1] presented an application study on connection model between BOD-DO and temperature. However, these traditional methods fail to solve the complicated nonlinear relationship between physico chemical factors of water and DO and BOD.

Artificial neural networks (ANNs) are self organizing, self teaching and nonlinear and can deal with systems that are difficult to be described with traditional mathematical models. 1) An ANN model does not require a prior knowledge of the

system and therefore, can be applied to solve the problems not clearly defined 2) The model has more tolerance to noise and incomplete data and thus, requires less data for model development and 3) The results are the outcome of the collective behavior of data, and thereby, the effect of outlier is minimized. In ANN, the gradient descent search optimization embedded with back propagation algorithm is quite popular in ANN for exploring diverse areas such as bio-medical, engineering, image processing, water resources, and others. ANNs have been widely applied to evaluate DO and BOD in water. Yang *et. al* [2] set up the prediction model of dissolved oxygen during the process of sewage disposal. But every model is difficult to assess different DO and BOD concentrations in water, which is affected by the different factors and mathematical simulation in the process, may be difficult to determine the expression of water pollution control. Wan *et.al* [3] applied artificial neural network and GIS in evaluating water quality. Kuo *et.al* [4] applied the ANN model to assess the variation of groundwater quality in an area with Blackfoot disease in Taiwan, China. These models abandoned the human factor, only based on water quality standard features to learn, thereby were brought to link weight matrix and threshold matrix and their outputs are more real objective.

This study demonstrated the application of Artificial Neural Network to forecast dissolved oxygen (DO) and Biochemical oxygen Demand (BOD) having the dynamic processes hidden in the measured data itself. The use of Multi-Layer Perceptron Neural Network (MLP-NN) model in water quality prediction in the study area could be complementary in capturing patterns of historical data set and improving the prediction accuracy.

2. MATERIALS AND METHODS

2.1. Study area and data analysis

The study area Batlagundu is bounded by Longitude 77 ^{0} 45' 33.84" E and Latitude 10 ^{0} 9' 55.80"N with an average elevation of 320 meters (1049 feet). The main occupation of this study area is agriculture. The sources of water supply in the area are hand pumps, bore holes and dug wells. The precipitation which is the sole source of ground water recharges in the study area is very low. The area is very humid (86%) and warm with an average temperature 22 ^{0}C. In order to achieve the research objective, samples were collected from the study area as shown in Fig.1. The water quality parameters were monitored regularly during different seasons of summer, rainy and winter over a period of two years ie., 2012 and 2013.

2.2. Identification of study objects and affecting factors

The selection of appropriate input parameters is a very important aspect for the neural network modeling. In order to use the MLP-NN structures effectively, the input parameters must be selected with great care. This highly depends on better understanding of the problem. The choice of input parameters based on statistical correlation analysis is the most popular analytical technique for selecting input. Based on the existing measured values and statistical analyses, the following water quality parameters were selected for the ANN modeling in this study, namely pH, total dissolved solids (TDS), electrical conductivity (EC), total hardness (TH), calcium (Ca) ions, magnesium (Mg) ions, and total alkalinity (TA), each of which affects the Dissolved Oxygen (DO) and Biochemical Oxygen Demand (BOD) to a certain degree. In addition, today's Dissolved Oxygen and Biochemical Oxygen Demand will have some impact on that of the next day, so in the end these factors are determined to be affecting factors for the forecast, i.e., the input variables of the model and DO and BOD being the study object i.e output variable. The model intends to achieve a forecast of the next day's Dissolved Oxygen and Biochemical Oxygen Demand value from today's water quality variables. Creating and testing the model is done via MATLAB, mathematical software introduced by Mathworks of USA in 1982 which has high-level numerical computation and data visualization capacity[5]. MATLAB is mathematical software with high-level numerical computation and data visualization capacity. It provides users with neural network design and simulation and enables them to work on design and simulation at greater convenience.

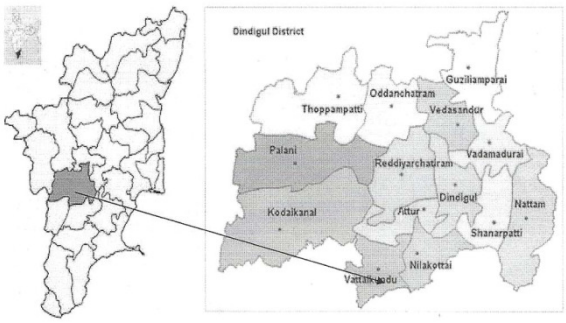

Figure - 1: Map of the study area

2.3. Artificial neural networks (ANN)

ANNs constitute an information-processing paradigm that is inspired by biological nervous systems [6]. The key element of this paradigm is the novel structure of the information processing system. It consists of a large number of highly interconnected processing elements

(neurons) working in unison to solve specific problems. An ANN is commonly divided into three or more layers: an input layer, hidden layer(s) and an output layer. The input layer contains the input nodes (neurons), i.e. the input variables for the network. The output layer contains the desired output of the system and the hidden layer usually contains a series of nodes associated with transfer functions. Each layer of the ANN is linked by weights that have to be determined through a learning algorithm. In this study, a three-layer feed-forward neural network with back propagation learning is employed (Fig.2). Back propagation is a commonly used learning algorithm in ANN applications. It uses the gradient descent algorithm to determine the weights in the network.

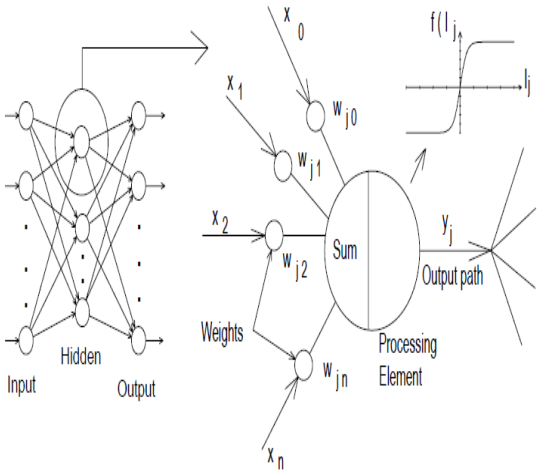

Figure - 2: Typical structure and operation of an ANN model.

Generally, forecasting models can be divided into statistical and physical based approaches. Statistical approaches determine the relationships between historical data sets, whereas physical based approaches model the underlying processes directly. MLP networks are closely related to statistical models and are the type of ANN most suited to forecasting applications [7]. When using ANNs for forecasting, the modeling philosophy employed is similar to that used in traditional statistical approaches. In both cases, the unknown model parameters (i.e. the connection weights in the case of ANNs) are adjusted to obtain the best match between the historical set of model inputs and the corresponding outputs.

The forecasting of DO and BOD to be governed by highly nonlinear processes and to introduce nonlinearity into the system, the commonly used sigmoid function

$$f(x) = \frac{1}{1+e^{-ax}} \qquad (1)$$

where 'a' is the slope parameter of the sigmoid function, adopted as the activation function that transfers the summed inputs to the output layer. This function also has the property that it squashes the independent variable, x, which may have a range from -∞ to +∞ to the range 0-1. At the output layer where the network output has to be compared with the target output, the target data needs normalization to the range (0, 1). The network is trained by adjusting the weights. The training process is done with a large number of training sets and training cycles (epochs). The main goal of the learning procedure is to find the optimal set of weights, which can ideally produce the correct output for the relative input. The output of the network is compared with the desired response to determine the error. The main steps involved in the development of an ANN, as suggested by Maier and Dandy [8] are illustrated in Figure 3.

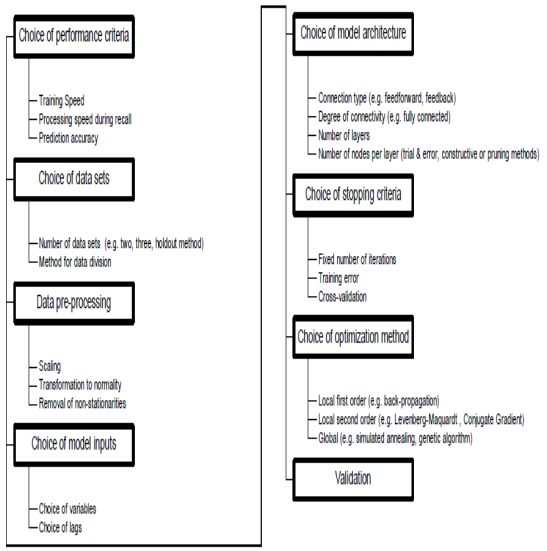

Figure - 3: The main steps involved in ANN model development

2.4. Selection of back propagation training algorithm

The back propagation (BP) learning algorithm [7] is a method conventionally used to perform the training of Artificial Neural Networks for adjusting weighed connections. A neural network can be used to represent a nonlinear mapping between input and output vectors. Neural networks are among the popular signal-processing technologies. In engineering, neural networks serve two important functions: as pattern classifiers and as nonlinear adaptive filters [9]. A general network consists of a layered architecture, an input layer, one or more hidden layers and an output layer [10]. The Multilayer perceptron (MLP) is an example of an artificial

neural network that is used extensively to solve a number of different problems, including pattern recognition and interpolation[11,12]. Each layer is composed of neurons, which are interconnected with each other by weights. In each neuron, a specific mathematical function called the activation function accepts input from previous layers and generates output for the next layer.

Standard back propagation is a gradient descent algorithm in which the network weights are moved along the negative of the gradient of the performance function. Although traditional BP uses a gradient descent algorithm to determine the weights in the network, it computes rather slowly due to linear convergence. The MLP is trained using the Levenberg–Marquardt technique as this technique is more powerful than the conventional gradient descent techniques. The Levenberg-Marquardt algorithm (LMA) is the most widely used optimization algorithm. It outperforms simple gradient descent and other conjugate gradient methods in a wide variety of problems. The LMA is a very simple but robust method, which provides a numerical solution to the problem of minimizing a function over a space of parameters for the function. Principally, it involves in solving the following equation:

$$\delta = (\boldsymbol{J}^t \boldsymbol{J} + \mu \boldsymbol{I})^{-1} \boldsymbol{J}^t \boldsymbol{E} \qquad (2)$$

where I is the identity matrix, J is the Jacobian matrix for the system, μ is the Levenberg's non-negative damping factor, δ is the weight update vector that we want to find and E is the error vector containing the output errors for each input vector used in training the network. The δ tells us by how much we should modify our network weights to reach a better solution. μ is adjusted at each iteration. If the reduction of E is rapid, a smaller value can be used, bringing the algorithm closer to the Gauss-Newton algorithm, whereas if iteration gives insufficient reduction in the residual, μ can be increased, giving a step closer to the gradient descent direction. In that way, LMA is considered as a hybrid between the classical Newton and steepest descent algorithms [13]. The Jacobian matrix can be created by taking the partial derivatives of each output in respect to each weight and has the following form:

$$\boldsymbol{J} = \begin{bmatrix} \frac{\partial \boldsymbol{F}(\boldsymbol{x_1},\boldsymbol{\omega})}{\partial \boldsymbol{\omega_1}} & \cdots & \frac{\partial \boldsymbol{F}(\boldsymbol{x_1},\boldsymbol{\omega})}{\partial \boldsymbol{\omega_W}} \\ \vdots & \ddots & \vdots \\ \frac{\partial \boldsymbol{F}(\boldsymbol{x_N},\boldsymbol{\omega})}{\partial \boldsymbol{\omega_1}} & \cdots & \frac{\partial \boldsymbol{F}(\boldsymbol{x_N},\boldsymbol{\omega})}{\partial \boldsymbol{\omega_W}} \end{bmatrix} \qquad (3)$$

2.5. Criteria and model performance evaluation

Two different criteria are used in order to evaluate the effectiveness of each network and its ability to make precise prediction. The Mean Square Error (MSE) can be used to determine how well the network output fits the desired output. The smaller values of MSE ensure better performance. It is defined as follows:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(DO_m - DO_p)^2 \qquad (4)$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(BOD_m - BOD_p)^2 \qquad (5)$$

The correlation coefficient (CC) is often used to evaluate the linear relationship between the predicted and measured values. It is defined as follows:

$$CC = \frac{\sum_{i=1}^{n}(DO_m - \overline{DO}_m)\,(DO_p - \overline{DO}_p)}{\sqrt{\sum_{i=1}^{n}(DO_m - \overline{DO}_m)^2 \sum_{i=1}^{n}(DO_p - \overline{DO}_p)^2}} \qquad (6)$$

$$CC = \frac{\sum_{i=1}^{n}(BOD_m - \overline{BOD}_m)\,(BOD_p - \overline{BOD}_p)}{\sqrt{\sum_{i=1}^{n}(BOD_m - \overline{BOD}_m)^2 \sum_{i=1}^{n}(BOD_p - \overline{BOD}_p)^2}} \qquad (7)$$

where n is the number of observations, DO_p , BOD_p and DO_m, BOD_m are the predicted and measured dissolved oxygen and Biochemical oxygen Demand respectively, and $\overline{DO}_m$, $\overline{BOD}_m$ and $\overline{DO}_p$, $\overline{BOD}_p$ is the average of measured and average of predicted dissolved oxygen and Biochemical oxygen Demand concentrations.

3. RESULTS AND DISCUSSION

It is common practice to divide the available data into two subsets; a training set, to construct the neural network model and an independent validation set to estimate the model performance in a deployed environment [8]. Usually, two-thirds of the data are suggested for model training and one-third for validation [14]. A modification of the above data division method is cross-validation in which the data are divided into three sets: training, testing and validation. The training set is used to adjust the connection weights, whereas the testing set is used to check the performance of the model at various stages of training and to determine when to stop training to avoid over-fitting. The validation set is used to estimate the performance of the trained network in the deployed environment. Shahin *et. al.*[15] found that there is no clear relationship between the proportion of data for training, testing and validation and model performance; however, they found that the best result was obtained when 20% of the data were used for validation and the remaining data were divided into 70% for training and 30% for testing. In this study there are 500 data record, 350 records taken for training and 150 records taken for testing.

The sine plot (Figure 4) demonstrated the actual DO values compared with the predicted values from the neural network model using Levenberg-Marquardt algorithm. The figure illustrates that predicted values from the network is close to the DO actual value. The correlation coefficient, which measures the strength and direction of the linear relation between two variables (actual and predicted values) is R=0.97. The mean square error (MSE) of the model is equal to 0.032.

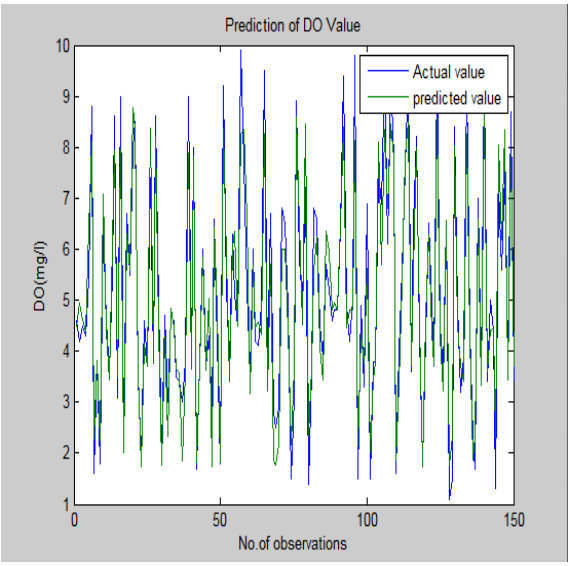

Figure - 4: Comparison of actual DO value and predicted DO value from neural network

The performances of the neural network models are assessed by comparing the measured and predicted values with one another. The model analyzed were functions of the readily available water quality parameters. Fig.4 shows the plot comparing the observed (actual) values of DO with the predicted DO from the network. Several forecast values in the figure are deviated from actual measured values due to the fact that forecast values can be affected by many factors during the study. In addition to the identified affecting factors, many other factors, such as weather condition or environmental pollution, can affect the forecast values every moment. They can be so unpredictable and thus make the study work more difficult. Nevertheless, the mean square error indicates that the overall forecast results are fairly good with error rate controlled within an acceptable range, proving the viability of the forecast model.

Figure 5 shows the forecasted concentrations of the BOD for the Levenberg-Marquardt algorithm. The back-propagation neural network models were applied to simulate and predict the concentration of BOD. The BOD is a major water quality parameter and is used as an indicator for the purity of the water. However, the determination of this parameter is resource-intensive, and exchange with an inexpensive surrogate parameter such as dissolved oxygen would be an advantage. The MSE measures the residual variances that indicate global goodness of fit between the predicted and observed variables. The correlation coefficients were relatively high during the training and testing period for Levenberg-Marquardt algorithm as shown in Table 1.

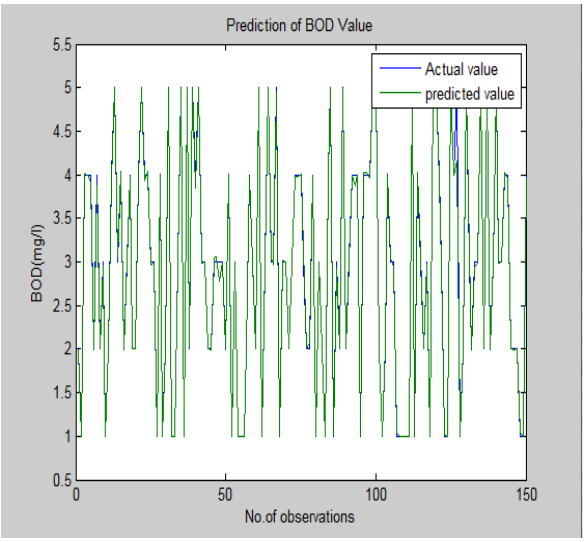

Figure - 5: Comparison of actual BOD value and predicted BOD value from neural network

Table - 1: Performance parameters of the ANN model for computation of the DO and BOD

Parameters	Data set	MSE	Correlation Coefficient
DO	Training	0.036	0.97701
	Testing	0.033	0.95815
	Validation	0.031	0.96915
	All	0.032	0.97147
BOD	Training	0.073	0.9986
	Testing	0.078	0.99858
	Validation	0.075	0.99786
	All	0.071	0.99848

4. CONCLUSION

In this paper, the MLP neural network using the Levenberg-Marquardt algorithm is applied to predict the DO and BOD in ground water at Batlagundu, Dindigul District, Tamilnadu. The networks were designed by putting weights between neurons, by using the hyperbolic tangent function of training. The results for the training and the test data sets were satisfactory. The

results showed that the neural network model provided high correlation coefficients and low mean square error. Dissolved oxygen and Biochemical oxygen Demand are important parameters for usage conditions of water. This result may be applied to automate DO and BOD estimations which are utilized in water management and treatment systems corresponding to the government's policy in water management. Compared to statistical methods, neural networks provide a general framework for determining relationships between water quality data and do not require the specification of any functional form.

5. REFERENCES

1. Ma Rui-jie, Li Xin, Liu Xiao-duan and Xu Qing. An application study of connection model between reservoir organic pollutant BOD-DO and temperature**. Rock and Mineral Analysis.** 2005; 24(2): 105−108.
2. Yang Xiao-ming, Li Ming-huan, Yang Pu and Jia Ming-xing. Dissolved oxygen prediction model and its error revision. **Control Engineering of China.** 2004; 11(2): 127−129.
3. Wan You-chuan, Xie Hong-yu, Wu Zheng-bing and Shen Xiao-li.Application of artificial neural network and GIS to water quality evaluation. **Journal of Wuhan University Engineering.** 2003; 36(3): 7−12.
4. Kuo Yi-ming, Liu Chen-ying and Lin Kao-hong. Evaluation of the ability of an artificial neural network model to assess the variation of groundwater quality in an area of Blackfoot disease in Taiwan, **Water Research.** 2004; 36(1): 148−158.
5. Zhang YH. Mastering MATLAB5. **Tsinghua University Press,** Beijing, 1999.p.1-2.
6. Haykin S. Neural Networks. **Prentice-Hall, Englewood Cliffs NJ,** 1994.
7. Rumelhart DE, Hinton G E and Williams R J. Learning representations by backpropagating errors, **Nature.** 1986; 323; 533−536.
8. Maier HR, Dandy GC. Neural networks for the prediction and forecasting of water resources variables: A review of modelling issues and applications. **Environmental Modelling & Software,** 2000; 15: 101-124.
9. Fausett L. Fundamentals of Neural Networks Architecture. **Algorithms and Applications, Pearson Prentice Hall,** USA, 1994.
10. Diamantopoulou MJ, Antonopoulos VZ and Papamichail DM. The Use of a Neural network Technique for the Prediction of Water Quality Parameters of Axios River in Northern Greece. **Journal 0f Operational Research, Springer,** 2005; 115-125.
11. Anguita D, Ridella S and Rivieccio F. K-folds Generalization Capability Assessment for Support Vector Classifiers, **Proceeding of International Joint Conference on Neural Network,** Canada, 2005; 855-858.
12. Li-hua Chen, and Xiao-yun Zhang. Application of Artificial Neural Network to Classify Water Quality of the Yellow River. **Journal 0f Fuzzy Information and Engineering,** 2009; 15- 23.
13. Souza CR. **Neural Network Learning by the Levenberg-Marquardt Algorithm with Bayesian Regularization,** 2009.
14. Hammertstrom D. Working with neural networks, **IEEE spectrum,** 1993; 30(7): 46-53.
15. Shahin MA, Maier HR and Jaksa MB. Data division for developing neural networks applied to geotechnical engineering. **Journal of Computing in Civil Engineering. ASCE**, 2004b; 18(2): 105-114.