Journal of Zhejiang University SCIENCE A
ISSN 1673-565X (Print); ISSN 1862-1775 (Online)
www.zju.edu.cn/jzus; www.springerlink.com
E-mail: jzus@zju.edu.cn

Water quality forecast through application of BP neural network at Yuqiao reservoir*

ZHAO Ying†, NAN Jun, CUI Fu-yi, GUO Liang
(*School of Municipal and Environmental Engineering, Harbin Institute of Technology, Harbin 150090, China*)
†E-mail: zhaoying@hit.edu.cn

Received Dec. 26, 2006; revision accepted Apr. 9, 2007

Abstract: This paper deals with the study of a water quality forecast model through application of BP neural network technique and GUI (Graphical User Interfaces) function of MATLAB at Yuqiao reservoir in Tianjin. To overcome the shortcomings of traditional BP algorithm as being slow to converge and easy to reach extreme minimum value, the model adopts LM (Levenberg-Marquardt) algorithm to achieve a higher speed and a lower error rate. When factors affecting the study object are identified, the reservoir's 2005 measured values are used as sample data to test the model. The number of neurons and the type of transfer functions in the hidden layer of the neural network are changed from time to time to achieve the best forecast results. Through simulation testing the model shows high efficiency in forecasting the water quality of the reservoir.

Key words: Water quality forecast, BP neural network, MATLAB, Graphical User Interfaces (GUI)
doi:10.1631/jzus.2007.A1482 **Document code:** A **CLC number:** TU991

INTRODUCTION

The Yuqiao reservoir covers a flow area of 2060 km^2 with a total storage capacity of 1.559×10^6 m^3. It is the only reservoir in Tianjin that incorporates various functions including flood control, urban water supply, irrigation, power generation and aquaculture. It is also a key regulating reservoir for the Diversion Project from Luanhe River to Tianjin and plays a major role in securing the economic growth of Tianjin and the life and property of the people living in the lower region. However in recent years, along with the development of farmland fertilization and fishery in the area, the content of nitrogen and phosphorus in the water keeps rising, resulting in massive reproduction of undesirable water bodies such as fungus and waterweeds and their over-nutrition. To ensure the safety of drinking water, it is important to initiate a comprehensive treatment of the reservoir's water quality in addition to the restriction and standardization of development in farming and fishery in the proximity of the reservoir. As a result, this paper studies water quality forecast at Yuqiao reservoir to gain better understanding of possible water quality changes in the future. The study will provide scientific evidences for water quality management and propose necessary action to foresee and prevent any future problems, thus playing a positive role in securing the safety of water supply in Tianjin.

In recent years, many researches have been conducted on water quality forecast model (Chen *et al.*, 2003; Kurunç *et al.*, 2005; Li, 2006). However, as water quality can be affected by so many factors, traditional data processing methods are no longer good enough for solving the problem (Wu *et al.*, 2000; Xiang *et al.*, 2006) as such factors show a complicated nonlinear relation to the variables of water quality forecast. On the other hand, the neural network can imitate such basic characteristics of the human brain as self-adaptability, self-organization and error tolerant and has been widely adopted for

* Project (No. 2006AA06Z305) supported by the Hi-Tech Research and Development Program (863) of China

mode identification, analysis and forecast, system recognition and design optimization (Niu *et al.*, 2006; Shu, 2006). 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. This paper focuses on forecasting the water quality COD (chemical oxygen demand) and DO (dissolved oxygen) values of Yuqiao reservoir. Based on BP neural network technique and MATLAB's GUI function, the study creates a water quality forecast model and adopts the 2005 measured values of Yuqiao reservoir as sample data to conduct testing and simulation in order to verify the viability of the model.

IDENTIFICATION OF STUDY OBJECTS AND AFFECTING FACTORS

COD and DO are two major values for evaluating water quality. This paper studies the forecast of COD and DO at Yuqiao reservoir. Based on existing measured values and correlative analysis, 8 factors are identified as affecting factors including water temperature, turbidity, pH, alkalinity, chloride, NH_4^+-N, NO_2^--N and hardness, each of which affects the water quality COD and DO to a certain degree. In addition, today's COD and DO value will have some impact on that of the next day, so in the end 10 factors are determined to be affecting factors for the forecast, i.e., the input variables of the model, and COD and DO, being the study objects, are output variables. The model intends to achieve a forecast of the next day's COD and DO value from today's 10 water quality variables, or affecting factors.

CREATING THE WATER QUALITY FORECAST MODEL

MATLAB

Creating and testing the model is done via MATLAB, a mathematical software introduced by Mathworks of USA in 1982 which has high-level numerical computation and data visualization capacity (Zhang, 1999). MATLAB Neural Network Toolbox 4.0 is an integral part of MATLAB6.x high-performance visualized numerical computation software. Aimed for analysis and design of neural network, Toolbox 4.0 offers many toolbox functions that can be called directly. GUI and Simulink, the simulation tool, and has become an ideal tool for analysis and design of neural network. It provides the GUI for design and simulation of neural network, making it more convenient for the users to design and conduct simulation testing. The model can also be modified subject to actual needs to forecast water quality under various conditions.

Basic theory of creating the model

In recent years, neural network technology has been widely adopted in water quality forecast, in which BP network is commonly used (Lee *et al*., 2003; Mo *et al*., 2004). The model created in this paper is a BP neural network model with a single hidden layer (Fig.1), with R as the input layer, S^1 the hidden layer, S^2 the output layer, $\boldsymbol{IW}^{1.1}$ the weight matrix of the input layer, $\boldsymbol{LW}^{2.1}$ the weight matrix from the hidden layer to the output layer, b^1 and b^2 threshold values of the hidden and output layer respectively and f^1 and f^2 the neuron transfer functions of the hidden and output layer respectively. As theoretically proven, such BP model as shown in Fig.1 can approach any nonlinear function with limited interruptions at any accuracy as long as the neurons in the hidden layer of the model are sufficient (Xu and Wu, 2002).

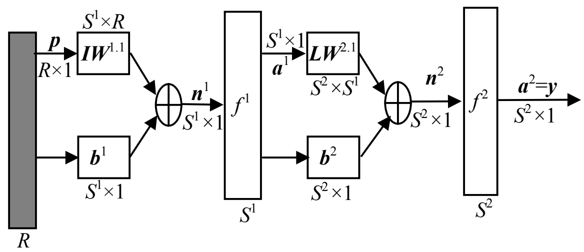

Fig.1 BP neural network with a single hidden layer

In Fig.1, the input and output variables are established for evaluation of water quality. The input variables are the factors affecting the output variables, which are also the study objects. It is assumed that the actual output of the neuron j in the output layer is $y_j(t)$ at time t and the expected output $d_j(t)$, so the network error function $E(t)$ at the time t will be defined as follows:

$$E(t)=\frac{1}{2}\sum_{j=1}^{q}(y_j(t)-d_j(t))^2,$$

q is the number of neurons in the output layer (S^2); ε is a pre-set error margin. The model that stops testing when $E(t)$ is less than ε is the desired model (Guo *et al*., 2001; Kuo *et al*., 2004).

Any input and output water quality variables that are correlated with each other can find a suitable network model to connect the input and output ends by adjusting the internal structure of the network and variables of the model. In the model, the number of neurons in the hidden layer, the transfer function of neurons in the hidden and output layers can be changed and the appropriate learning algorithm can be selected to meet the pre-determined standards on error rate.

Any type of model always relies on the use of sample data to train the network so as to find the best working model (Xu *et al*., 2007). Taking the existing water quality measured values as the input and output sample data, the model will be tested and put to a simulation testing. If the error rate is within an acceptable range, the model can then be applied for water quality forecast in real life (Chang and Chao, 2006).

Creating water quality forecast model with MATLAB

By keying in nntool in the command window of MATLAB, the user will enter the main page of the neural network GUI (Fig.2), the Network/Data Manager.

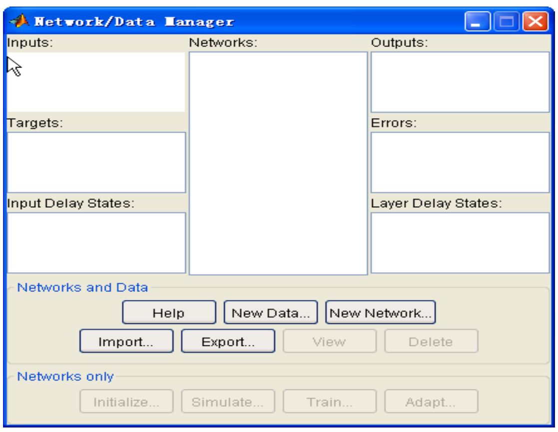

Fig.2 GUI main page

First, upload respectively under Inputs and Targets in GUI main page the input and output (target) data that have been previously written into Excel worksheet. The input variables are set at 10 and output variables 2. Next, click New Network to create a new network model as shown in Fig.3. Select Feed-forward backdrop as the Network Type. Based on experiences, the number of neurons in the hidden layer (i.e. Layer 1) can be chosen between 10 and 20, and *LOGSIG* or *TANSIG* as the neuron transfer function of the hidden layer. The output characteristics of the entire neural network will be decided by the characteristics of the last layer of the BP network. When *Sigmoid* functions are applied to the last layer, the output of the entire network will be limited to a smaller range; and if *Purelin* is applied to the last layer, the output could be an arbitrary value. As a result, *Purelin* is chosen as the transfer function for the neurons of the output layer.

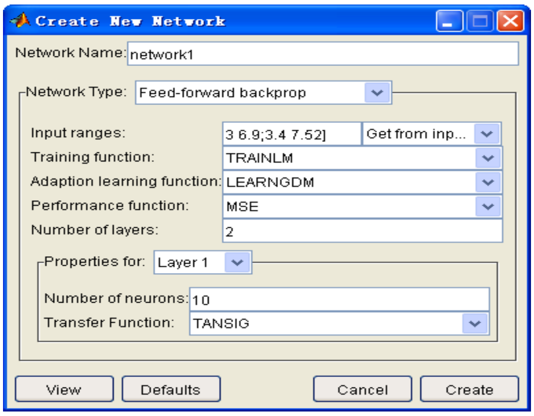

Fig.3 Create new network

While traditional BP algorithm is a gradient descent algorithm, which computes rather slowly due to linear convergence, LM (Levenberg-Marquardt) algorithm, improved from BP algorithm, is much faster since it adopts the method of approximate second derivative (Wang, 2004). Therefore, LM algorithm is used in the model, i.e., select *TRAINLM* for the Training Function in the figure above.

The input and output data come from the 3 months, a total of 90 days' measured values of 2005 at Yuqiao reservoir, Tianjin, therefore there are 90 groups of data in total (90×12). Since the number of neurons in the hidden layer can be selected among 11 options between 11 and 20 and the transfer functions from either *TANSIG* or *LOGSIG*, 22 models can be totally created subject to various numbers of neurons and transfer functions in the hidden layer. Fig.4 shows one example of the model. Each model will be trained and tested with sample data separately.

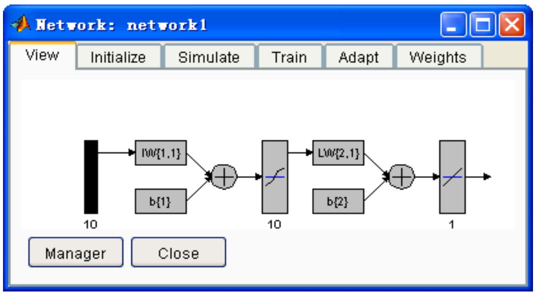

Fig.4 Network structure 1

RESULT ANALYSIS AND MODEL SELECTION

Each model gets tested three times. Since initial weights and thresholds are randomly generated, training results will turn out to be different every time. Select the lowest error rate as the minimum value that the model can achieve. When each model completes 2000 trainings and the errors of the 22 models are compared, the model with the lowest error rate will be the desired model. Table 1 shows the error rate of each model at each test.

The table shows that the lowest error rate is 0.0002190. Therefore, the corresponding network will be the desired forecast model, of which the number of neurons in the hidden layer is 18 and transfer function *TANSIG*. The error rate of the model is rather small. Then simulation testing will be done to validate the forecast results of the model.

SIMULATION TESTING

Once the best working model is obtained through network training, the COD and DO forecast values of November will be worked out and compared with actual measured values to evaluate the forecast results.

Figs.5 and 6 show the simulation results of COD. Fig.5 compares the COD forecast values with the actual measured values, and Fig.6 shows the COD forecast error rate. It can be seen in Fig.5 that the two curves almost overlap each other. Correlative analysis

Table 1 Comparison of the error rate of each model at each test

Transfer function of the hidden layer	Number of neurons in the hidden layer	Test 1	Test 2	Test 3	Optimal value
	10	0.0233099	0.0449321	0.0400328	0.0233099
	11	0.0315491	0.0918271	0.0151227	0.0151227
	12	0.0027208	0.0367409	0.0129202	0.0027208
	13	0.0382716	0.1375070	0.0089725	0.0089725
	14	0.2451930	0.0026240	0.0482054	0.0026240
TANSIG	15	0.0497664	0.0035787	0.0054661	0.0035787
	16	0.0017094	0.0023878	0.0079453	0.0017094
	17	0.0056553	0.0014324	0.0111583	0.0014324
	18	0.0002190	0.0306762	0.0006335	0.0002190*
	19	0.0015005	0.0012583	0.0015936	0.0012583
	20	0.0012252	0.0171138	0.0569807	0.0012252
	10	0.0168789	0.0227590	0.0686120	0.0168789
	11	0.0123424	0.1012560	0.0362708	0.0123424
	12	0.0805504	0.0048781	0.0338589	0.0048781
	13	0.0388126	0.0389666	0.0132847	0.0132847
	14	0.0123701	0.0149234	0.0110111	0.0110111
LOGSIG	15	0.0117938	0.0618509	0.0304595	0.0117938
	16	0.0012492	0.0472839	0.0021735	0.0012492
	17	0.0146762	0.0248416	0.0529890	0.0146762
	18	0.0181797	0.0380558	0.0665121	0.0181797
	19	0.0013349	0.0346547	0.0241650	0.0013349
	20	0.0460249	0.0122586	0.0088362	0.0088362

*0.0002190 is the lowest error rate

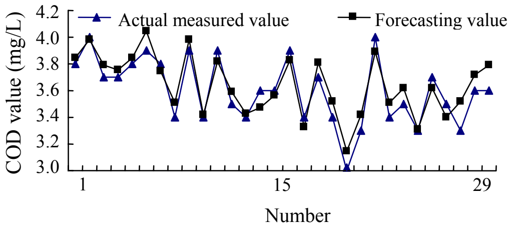

Fig.5 COD forecast value and actual measured value

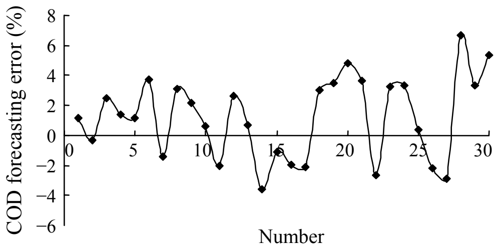

Fig.6 COD forecast error rate

between the two groups of data indicates that the COD correlative coefficient is 0.8537, and analysis of the forecast error rate shows that the average forecast error is 2.56%, with maximum at 6.70% and minimum at 0.36%.

Figs.7 and 8 show the simulation results of DO. Fig.7 compares the DO forecast values with the actual measured values, and Fig.8 shows the DO forecast error rate. Correlative analysis between two groups of data indicates that the DO correlative coefficient is 0.9418, and analysis of the forecast error rate shows that the average forecast error is 1.68%, with maximum at 4.72% and minimum at 0.26%.

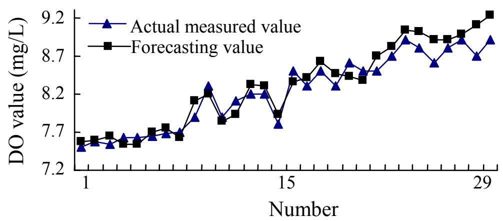

Fig.7 DO forecast value and actual measured value

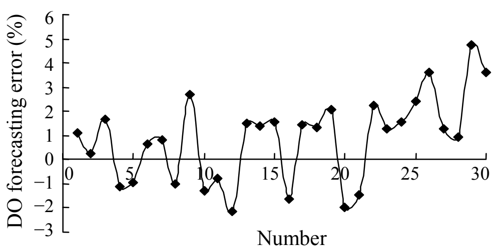

Fig.8 DO forecast error rate

From Figs.5 and 7, it can be seen that the forecast results of the first half of November are better than that of the second half. Since the sample data is selected from August through October of 2005, it indicates that the forecast model is relatively time-bound. As time goes by, the more deviated is the forecast result. Therefore it is quite necessary to update the model from time to time with new measured values. And the fact that the DO forecast result comes out better than that of COD shows that the selected affecting factors have greater impact on COD than on DO and that selection of affecting factors might affect the forecast results of the model remarkably. Several forecast values in the figures are more 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 average forecast error rate indicates that the overall forecast results are fairly good with error rate controlled within an acceptable range, proving the viability of the forecast model.

CONCLUSION

This paper adopts the BP neural network technology and GUI function of MATLAB to achieve easier and faster water quality forecast at Yuqiao reservoir. A water quality forecast model is created with application of neural network, in which LM algorithm is used for its faster convergence speed and lower error rate to overcome the shortcomings of traditional BP algorithm as slow to converge and easy to reach extreme minimum value. The study has the 2005 actual measured values of the reservoir as sample data to train the model. By changing the number of neurons and the type of transfer function of the hidden layer, the best working forecast model was obtained. Simulation testing showed that forecast model is time-bound and therefore it is necessary to update the model from time to time with actual measured values. Selection of affecting factors also plays a key role since they can have great impact on the forecast results. It concludes that the model can produce good forecast results in general and can be used for water quality forecast of the reservoir.

Moreover, this model can be extended to further applications. Provided that other water quality variables of this reservoir are to be forecasted, or a variety of variables to be forecasted simultaneously, it only needs to identify the affecting factors first, then create the forecast model in the way mentioned above, decide the input and output variables for the model, use existing measured values as sample data for training, obtain the best working model by changing the network's internal structure and variables, and complete the model with simulation testing. If the correlation between the forecast and actual measured values is fairly good, the forecast model is viable and can be applied to real practice. In this sense the forecast model is characterized by its extendability and practibility. Nevertheless, due to the fact that water quality forecast can be easily affected by external environment (Vandenberghe *et al*., 2007), the obtained model sometimes produces results much deviated from the actual values, therefore further study needs to be done in future work to identify the suitable forecast model, understand its laws of changes and solve the problem of forecast deviation.

References

Chang, T.C., Chao, R.J., 2006. Application of backpropagation networks in debris flow prediction. *Engineering Geology*, **85**(3-4):270-280. [doi:10.1016/j.enggeo.2006.02.007]

Chen, L.H., Chang, Q.C., Chen, X.G., Hu, Z.D., 2003. Using BP neural network to predict the water quality of Yellow River. *Journal of Lanzhou University (Natural Sciences)*, **39**(2):53-56 (in Chinese).

Guo, Z.Y., Chen, Z.Y., Li, L.Q., Song, B.P., Lu, Y., 2001. Artificial neural network and its application in regime prediction of groundwater quality. *Journal of East China Normal University (Natural Sciences)*, (1):84-89 (in Chinese).

Kuo, Y.M., Liu, C.W., Lin, K.H., 2004. 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*, **38**(1):148-158. [doi:10.1016/j.watres.2003.09.026]

Kurunç, A., Yürekli, K., Çevik, O., 2005. Performance of two stochastic approaches for forecasting water quality and streamflow data from Yeşilırmak River, Turkey. *Environmental Modelling & Software*, **20**(9):1195-1200. [doi:10.1016/j.envsoft.2004.11.001]

Lee, J.H.W., Huang, Y., Dickman, M., Jayawardena, A.W., 2003. Neural network modeling of coastal algal blooms. *Ecological Modelling*, **159**(2-3):179-201. [doi:10.1016/S0304-3800(02)00281-8]

Li, R.Z., 2006. Advance and trend analysis of theoretical methodology for water quality forecast. *Journal of Hefei University of Technology*, **29**(1):26-30.

Mo, H.F., Gu, A.Y., Zhang, X.Z., Zhang, J.C., 2004. Research on a method of BP neural network in water quality evaluation. *Control Engineering of China*, **11**:9-10,19.

Niu, Z.G., Zhang, H.W., Liu, H.B., 2006. Application of neural network to prediction of coastal water quality. *Journal of Tianjin Polytechnic University*, **25**(2):89-92.

Shu, J., 2006. Using neutral network model to predict water quality. *North Environment*, **31**(1):44-46.

Vandenberghe, V., Bauwens, W., Vanrolleghem, P.A., 2007. Evaluation of uncertainty propagation into river water quality predictions to guide future monitoring campaigns. *Environmental Modelling & Software*, **22**(5):725-732. [doi:10.1016/j.envsoft.2005.12.019]

Wang, Q.H., 2004. Improvement on BP algorithm in artificial neural network. *Journal of Qinghai University*, **22**(3): 82-84.

Wu, H.J., Lin, Z.Y., Guo, S.L., 2000. The application of artificial neural networks in the resources and environment. *Resources and Environment in the Yangtze Basin*, **9**(2):237-241 (in Chinese).

Xiang, S.L., Liu, Z.M., Ma, L.P., 2006. Study of multivariate linear regression analysis model for ground water quality prediction. *Guizhou Science*, **24**(1):60-62.

Xu, D., Wu, Z., 2002. Neural Network-system Design and Analysis Based on MATLAB6.X. University of Xi'an Electronics Technology Press, Xi'an, p.2 (in Chinese).

Xu, L.J., Xing, J.D., Wei, S.Z., Zhang, Y.Z., Long, R., 2007. Optimization of heat treatment technique of high-vanadium high-speed steel based on back-propagation neural networks. *Materials & Design*, **28**(5):1425-1432. [doi:10.1016/j.matdes.2006.03.022]

Zhang, Y.H., 1999. Mastering MATLAB5. Tsinghua University Press, Beijing, p.1-2 (in Chinese).