ISSN 1001—0742
Journal of Environmental Sciences Vol. 18, No.4, pp. 827—831, 2006 CN 11—2629/X

Article ID: 1001-0742(2006)04-0827-05 **CLC number**: X131.2 **Document code**: A

Prediction of the concentration of chlorophyll-*a* for Liuhai urban lakes in Beijing City

ZENG Yong[1,2], YANG Zhi-feng[1,2,*], LIU Jing-ling[1,2]

(1. State Key Laboratory of Water Environment Simulation, School of Environmental, Beijing Normal University, Beijing 100875, China. E-mail: zfyang@bnu.edu.cn; 2. Key Laboratory for Water and Sediment Sciences of Ministry of Education, School of Environment, Beijing Normal University, Beijing 100875, China)

Abstract: The weekly water quality monitor data of Liuhai lakes between April 2003 and November 2004 in Beijing City were used as an example to build an artificial neural networks (ANN) model and a multi-varieties regression model respectively for predicting the fresh water algae bloom. The different predicted abilities of the two methods in Liuhai lakes were compared. A principle analysis method was first used to select the input variables of the models to avoid the phenomenon of collinearity in the data. The results showed that the input variables for the artificial neural networks were T, TP, transparency(SD), DO, chlorophyll-*a* (Chl-*a*), pH and the output variable was Chl-*a*. A three layer Levenberg-Marguardt feed forward learning algorithm in ANN was used to model the eutrophication process of Liuhai lakes. 20 nodes in hidden layer and 1 node of output for the ANN model had been optimized by trial and error method. A sensitivity analysis of the input variables was performed to evaluate their relative significance in determining the predicted values. The correlation coefficient between predicted value and observed value in all data and in test data were 0.717 and 0.816 respectively in the artificial neural networks. The stepwise regression method was used to simulate the linear relation between Chl-*a* and temperature, of which the correlation coefficient was 0.213. By comparing the results of the two models, it was found that neural network models were able to simulate non-linear behavior in the water eutrophication process of Liuhai lakes reasonably and could successfully estimate some extreme values from calibration and test data sets.

Keywords: artificial neural networks; eutrophication; multi-varieties regression; forecast; Liuhai lakes; Beijing City

Introduction

The eutrophication of fresh water has become a main water environmental problem in the world. The main negative impacts of fresh water eutrophication are water quality deterioration and the decrease of hydrophytes and aquatic species due to the bloom of phytoplankton. Water use, human health and social development have been seriously impacted by fresh water eutrophication. It is generally accepted that the blooming of phytoplankton is caused by the co-effect of physical, chemical and biological processes in the fresh water. However, the relations between phytoplankton bloom and the various factors are complex. The stochastic, uncertain and nonlinear characters in the phytoplankton bloom mechanism exist and are still not fully realized today (Lu *et al.*, 2003).

The techniques used to model the incidence of phytoplankton in freshwater can be divided into two main categories: process-based models and statistically-based models. Process-based simulation models are based on basic physical and chemical theory to maximize the use of scientific knowledge. This models include: QUAL-II, WASP, and SALMO (Xia and Du, 2000; Peng and Guo, 2002; Jing and Xu, 2004; Walter, 2004). The process-based models have a wider domain of application than the statistical models, as they are based on fundamental physical relationships. However they required large quantities of data to build and calibrate the parameters. Statistical models make use of historical data to obtain the best possible relationship between phytoplankton biomass and a number of environmental variables. They include the traditional multivariable regression methods and artificial neural network methods. Zhou *et al.* (1999) suggested using multivariable regression methods to predict the environmental change of Donghu Lake in Wuhan City. However, selection of prediction model structures and nonlinear problems exist in the use of traditional multivariable regression models. Recently, artificial neural network models have been widely used to predict freshwater eutrophication because of its adaptive, learning ability and because it acts as a real multi-input and output system (Karul, 2000). Maier *et al.* (1998) used artificial neural networks to predict the relation between the concentration of phytoplankton and various input data under different lag times, which were used as model inputs to select the best fit model by sensitivity analysis. To avoid the over-fit problem ANN training process, Gurbuz (2003) trained and calibrated artificial neural networks by first the time terminate method. Pei *et al.* (2004) searched and optimized the structure of artificial neural networks by trial and error methods in algae bloom prediction for the Xihu Lake in Hangzhou City. Walter *et al.* (2001) compared the prediction abilities of a determinative model of SALMO and a recurrent neural network model of ANNA. The results indicated that ANNA is

Foundation item: The Hi-Tech Research and Development Program(863) of China (No. 2003AA601010) and the National Natural Science Foundation of China(No. 50239020); *Corresponding author

suitable for short time prediction. Wu *et al.* (2000) contrasted the precision of different prediction abilities of the neural network method and the single variable regression method in predicting the relation between water quality deterioration and population increase or fishery yield, but it was not applied in prediction of algae bloom by using the multi-variable regress ion method.

Among the statistical models, the multi-variable regression models and neural network models have been used to predict the concentration of Chlorophyll-*a* (Chl-*a*) among existing research results. However, they have not been used in Liuhai lakes. So it is necessary to contrast these two methods and selected the better one to predict the water algae bloom for Liuhai lakes with available data. The weekly water quality monitor data between April 2003 and November 2004 were used as an example. Two methods are contrasted in order to provide detailed information to explore a fresh water eutropication management tool for Liuhai lakes in Beijing City.

1 Research areas

Liuhai lakes are important scenic water bodies in Beijing City, and consist of Xihai, Houhai, Qianhai, Beihai, Zhonghai and Nanhai lakes. Its source water comes from Guanting Reservoir and runoffs along the way. The inflow of the lake begins at Tielin gate. It runs through the sequence of sub-lakes and later outflows to Tongzhihe River. The total area is 142.11×10^4 m^2. The length is 3.5 km and the volume is 209.34×10^4 m^3. The map of research areas is shown in Fig.1.

The water quality parameters of the lakes are shown in Table 1. Its total phosphourous and total nitrogen content exceeds the water catalog according to water quality standards (GB3838-2002). The serious quantity of fresh water algae bloom has occurred since 1999, in which Xihai Lake is the most seriously affected among the six sub-lakes. It can be used as a representative sample of Liuhai lakes.

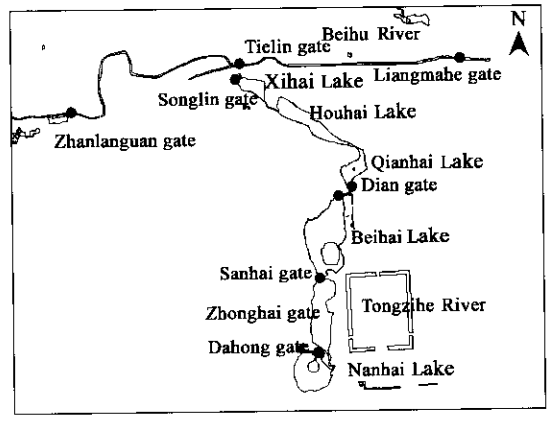

Fig.1 The map of research areas

2 Methods

One question when constructing an artificial neural networks model is how to select the input varieties of the model. If the input varieties contain repeated information, this information should be removed. The principle analysis method is used to select the input varieties of neural network models to avoid the phenomenon of collinearity among input varieties.

Table 1 Water quality of Liuhai lakes in year 2003

Sampling station		*T*, ℃	pH	DO, mg/L	COD_{Mn}, mg/L	NH_3-N, mg/L	TP, mg/L	TN, mg/L	Chl-*a*, μg/L	SD, m
Xihai	Average	24.4	8.1	12.0	8.6	0.80	0.21	1.60	71.6	0.5
	Scope	2.5—33.0	7.4—8.8	5.2—23.4	3.2—18.1	0.10—2.30	0.11—0.37	0.60—3.40	2.7—347.5	0.3—0.7
Houhai	Average	23.2	8.2	11.0	9.2	0.39	0.19	0.93	38.80	0.38
	Scope	2—32	7.5—9.1	3.3—18.6	4.3—31.1	0.13—1.17	0.046—0.37	0.26—2.53	1.34—200.5	0.15—1.0
Qianhai	Average	23.1	8.2	8.8	8.0	0.39	0.19	0.80	22.94	0.39
	Scope	2—33	7.5—9	2—16	3.8—19.6	0.15—1.25	0.05—0.60	0.21—2.3	1.34—81.5	0.1—0.8
Beihai	Average	20.2	8.2	7.0	7.2	0.36	0.14	0.82	19.04	0.37
	Scope	2—29.5	7.5—9.0	2—16.0	2.8—14.0	0.09—1.41	0.04—0.25	0.189—2.28	0—38.1	0.2—0.7
Zhonghai	Average	22.1	8.5	10.0	6.3	0.28	0.06	0.41	11.97	0.57
	Scope	1.3—30.5	7.8—9.4	7.8—12.6	3.7—8.0	0.1—0.49	0.02—0.10	0.03—1.44	0.67—44.8	0.35—1.5
Nanhai	Average	22.3	8.6	11.0	6.8	0.35	0.05	0.58	12.21	0.62
	Scope	1.5—31	7.8—9.4	8.6—14.4	2.9—11.8	0.11—0.99	0.02—0.10	0.06—2.35	1.34—29.4	0.4—2.0

2.1 Principle analysis method

The principle analysis method was used to reduce the original data by using the linear combination of original data (He, 2004). The method is to find eigenvalues and eigenvectors to substitute the original data. Assuming n samples, each having p indexes, an original matrix is acquired with $n \times p$ dimensions. After the standardization of the original matrix, the eigenvalues and eigenvectors can be achieved by solution of Eq.(1), which is as follows.

$$|R-\lambda I| = 0 \tag{1}$$

Where R is the correlation matrix; I is the identity matrix; λ is the eigenvalues.

If the eigenvector are labeled as $\gamma_1, \gamma_2, \gamma_i$, then the principle Y_p is given in Eq.(2):

$$Y_p = \gamma_i X;\ i = 1, 2, \Lambda, p \tag{2}$$

where X is the original variables.

2.2 Back propagation artificial neural networks

A back propagation artificial neural network model is used in this paper (Wen *et al.*, 2003). Firstly, data standardization is given by Eq.(3):

$$A = (A_n - u)/S_d \tag{3}$$

Where A is the data after standardization; A_n is the original data; u and S_d are the average and standard deviation of the original data. Data is nearly ± 1 after standardization. Thus it is advantageous for network training. The reverse process can be achieved according to the converse function of Eq.(3).

A tangent sigmoid transfer function is selected between input layer and hidden layer. A linear transfer function is selected between hidden layer and output layer. A tangent sigmoid transfer function is shown as Eq.(4).

$$f(x) = \frac{1}{1+e^{-x}} \tag{4}$$

the output of the hidden layer is shown as Eq.(5):

$$y_i = f(\sum_j w_{ij}x_j - \theta_i) \tag{5}$$

Where x_j are input layers; y_i are hidden layers; w_{ij} is the weight between input and hidden layer; θ_i is threshold value.

The output of the output layer is shown as Eq.(6):

$$O_t = f(\sum_i T_{it}y_i - \theta_t) \tag{6}$$

where O_t are output layers; T_{it} is the weight between hidden and output layers.

The prediction error is calculated by Eq.(7):

$$E = \frac{1}{2}\sum_t (t_t - O_t)^2 \tag{7}$$

where E is the standard error; t_t is the expectation output.

The network learning process is to reduce the standard error(E) between the prediction value and the observed value by changing the weight of the input layer (w_{ij}, T_{it}) and threshold value (θ_t) along gradient direction. The training process will be repeated until the training error has satisfied the required precision specified before training.

2.3 Multiple varieties regression model

A multiple varieties regression model is shown as follows (Zhang, 2002):

$$y_i = a + b_1x_{1i} + \cdots\cdots + b_nx_{ni} + e_i \tag{8}$$

where y_i is the estimated value of the dependent variety of y; a is the interception of the line; b_n are the regression coefficients; e is the random error.

The selection of independent varieties is done using the Stepwise method. It first calculates and contrasts the contribution of all independent varieties to dependent varieties. The highest one is first selected to enter the model. The process is then repeated as the contributions of the other varieties are calculated again and the highest one enters the model. If the varieties already in the model after a new variety enters have not satisfied the statistic signification, they should be excluded from the model. The process is repeated again until the varieties can no longer be excluded from the models.

3 Results

Data used in the studied areas come from the weekly water quality monitor data from April 2003 to November 2004. The monitored items include temperature, pH, DO, NH_4-N, COD_{Mn}, TP, TN, Chl-*a*, and transparency (SD). Among them, the P and N nutrition are the confined substances of algae growth. While pH and COD_{Mn} are the chemical characters of research water, which have direct and indirect impact on the water ecosystem. The temperature of water can represent the climate impact in the studied areas. SD can represent the light condition, which is an important factor for hydrophyte growth. While DO is an important factor for the water biology and chemical reactions.

After the data are tested by the correlation coefficient test as shown in Table 2, the concentration of Chl-*a* is found to have a significant relation with the varieties of T, DO, COD_{Mn}, TP and SD. It is therefore necessary to exclude some repeat information of the original data.

Factor analysis of the software SPSS is used to analyze the data (Table 3). The three principle factors are achieved after the data is dealt with by the varimax rotation method in the factor analysis method. The total contribution of the three factors is above 72.5%, which is $\sum_{i=1}^{m}\frac{\lambda_i}{P} \geqslant 72.5\%$. The load coefficients of the index exceeding 0.770 among the three factors are selected as input varieties for the ANN models. The detailed indices of selected input varieties are T, TP, SD, DO, Chl-*a* and pH. While the output variety is Chl-a for the next period. A three-layer feed forward neural network model was used in the paper and the construct of 20 nodes in hidden layer and 1 node of output for the ANN model is optimized by the trial and error method. Among the many available training methods, Levenberg-Marguardt algorithm was used in the paper because it was reported to have the fastest convergence for medium sized neural networks that contain up to a few hundred nodes. Neural Network Toolbox of Matlab by Math works Co. was used in all calculations.

In order to avoid the over-fitting problem, the early stopping method was used in this study. To decide when to stop the training process, the data was

Table 2 Correlation coefficient matrix

Item	*T*	pH	DO	NH_3-N	COD_{Mn}	TP	TN	Chl-*a*	SD
T	1.000	-0.225	0.130	0.025	0.631	0.666	-0.187	0.297	-0.622
pH	-0.225	1.000	0.182	-0.284	0.030	-0.335	-0.281	-0.225	0.000
DO	0.130	0.182	1.000	-0.465	0.465	0.146	-0.367	0.563	-0.329
NH_3-N	0.025	-0.284	-0.465	1.000	-0.096	0.268	0.332	-0.174	0.021
COD	0.631	0.030	0.465	-0.096	1.000	0.528	-0.086	0.488	-0.645
TP	0.666	-0.335	0.146	0.268	0.528	1.000	0.060	0.370	-0.596
TN	-0.187	-0.281	-0.367	0.332	-0.086	0.060	1.000	-0.033	0.030
Chl-*a*	0.297	-0.225	0.563	-0.174	0.488	0.370	-0.033	1.000	-0.490
SD	-0.622	0.000	-0.329	0.021	-0.645	-0.596	0.030	-0.490	1.000

Table 3 Rotated component matrix

Item	Principle 1	Principle 2	Principle 3
T	0.891	-2.439E-02	-4.511E-02
TP	0.836	-9.513E-03	0.287
SD	-0.807	-0.243	1.330E-02
COD	0.770	0.349	-9.102E-02
DO	0.207	0.832	-0.264
Chl-*a*	0.380	0.771	0.303
NH_3-N	0.183	-0.630	0.467
pH	-0.147	-3.224E-03	-0.792
TN	-0.134	-0.164	0.732

randomly divided into three subsets. Half the data was used for training, one quarter for validation and the last quarter for testing.

After two periods of the training process, the training process was stopped for the error was increasing. The training error, validation error and test error are given by Fig.2. This figure belongs to a well generalized training session because all of the error plots show a similar behavior and the neural network training is stopped before starting to over-fit the data. The correlation between predicted value and observed value is 0.771 and that in the testing subset is 0.816, which are as shown in Fig.3.

As part of the sensitivity analyses, each of the inputs is increased by 5% in turn, and the change of the output caused by that of the input is calculated. The sensitivity analysis of each input variety reflects its importance to the output variety, which is as follows.

$$S=\frac{\Delta_{Out}}{\Delta_{In}}\times 100\% \qquad (9)$$

where S is the sensitivity of varieties (%); Δ_{Out} is the change of output (%); Δ_{In} is the change of input (%).

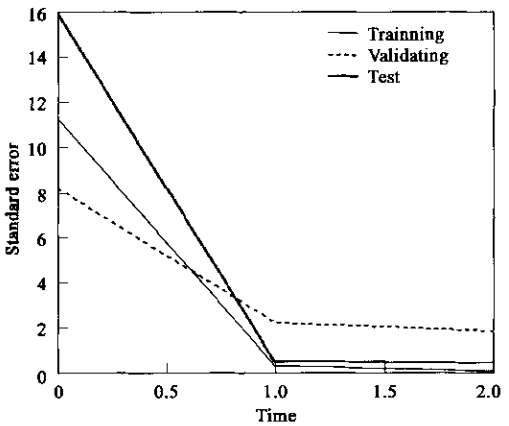

Fig.2 Error test of training, validating and testing

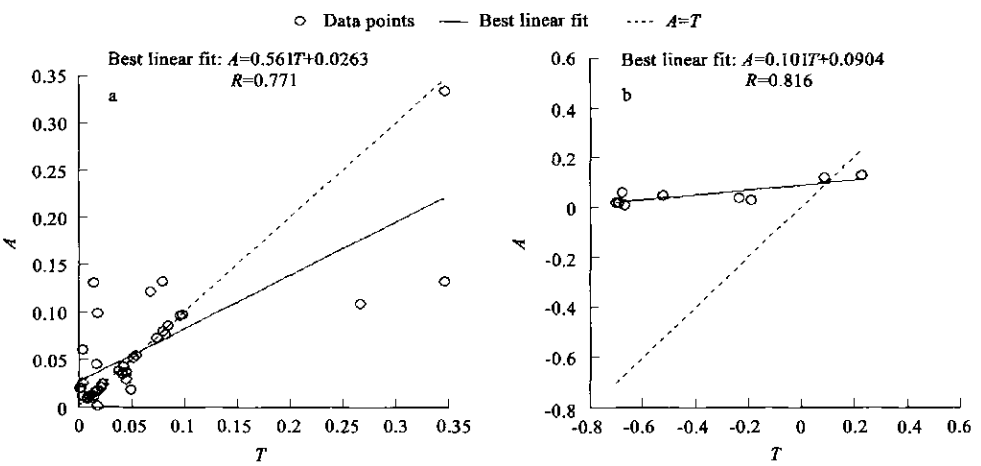

Fig.3 Comparison of predicted value and observed value of Chl-*a*(a) and Chl-*a* in testing subset(b)

The sensitivity analysis of the varieties is given by Fig.4. It shows that *T* and Chl-*a* are the most important factors for the concentration of Chl-*a* in the next period. It can be inferred that *T* is the most import environmental factor and the concentration of Chl-*a* reflecting the physical, chemical and biologic co-efficient results for algae bloom for the last period.

The linear regression method is also used to contrast the different prediction abilities of Chl-*a* for the two methods. The enter varieties are the same as

that of the ANN model. The Stepwise method (the enter probability is 0.05 and out probability is 0.1) is used to construct the model structures.

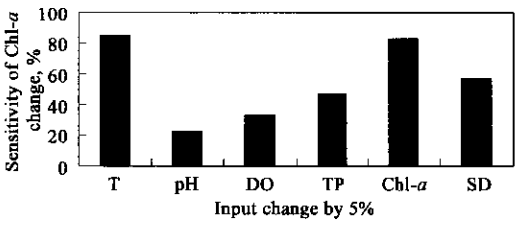

Fig.4 Sensitivity analysis of the input variables of ANNA

The parameters and test results of the regression model are shown in Table 4.

The multi-varieties regression model is given by Eq.(10).

$$\text{Chl-}a = -25.128 + 4.427T \quad (10)$$

Where Chl-*a* is the concentration of Chl-*a* (μg/L); *T* is the temperature (℃). The correlation of the model is 0.213.

Other varieties are excluded for not reaching the entering standards, which are shown in Table 5. The significant condition of the *T* test for pH, DO, TP and SD have exceeded the entering probability of 0.05 and are removed from Eq.(10).

Table 4 Coefficients of models

Model	Un-standardized coefficients		Standardized coefficients	*t*-test	*P*
	B	Std.error	*B*		
Constant	-25.128	47.172		-0.533	0.599
Temperature	4.427	2.036	0.392	2.174	0.039

The predicted value of the multi-varieties regression method, artificial neural networks and the observed value in the random test subset are shown in Fig.5. The predicted value by artificial network has more accuracy than that of the regression method as the average relative error rate is 0.67 to 4.19 respectively. The change trend and the four extreme values, which are never introduced to the system before, can be well simulated by the ANN model. However that of the multi-varieties regression method can only simulate three extreme values.

Table 5 Excluded Variables of models

Model	Coefficient *B*	*t*-test	*P*	Partial correlation	Co-linearity statistics tolerance
pH	-0.242	-1.279	0.213	-0.248	0.885
DO	-0.051	-0.277	0.784	-0.055	0.980
TP	-0.228	-0.942	0.355	-0.185	0.556
SD	-0.010	-0.044	0.965	-0.009	0.630

4 Conclusions

Due to the complex and heterogeneous character of Chl-*a* in Liuhai lakes, it can not be approximately predicted by a linear function of input variables. This

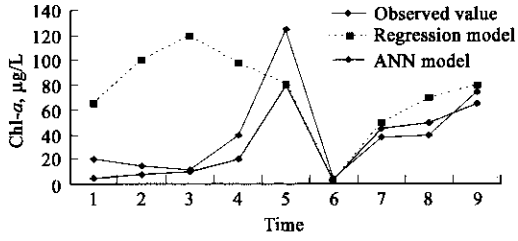

Fig.5 Comparison of predicted and observed curve of Chl-*a* in test subset

study showed that non-linear relationships between the related variables and Chl-*a* in the algae bloom can be modeled reasonably well in Liuhai lakes.

A neural network model can estimate values that lie outside the boundaries of the training set i.e. never introduced to the system before. From the study results, the neural network model can estimate all extreme values from the test data.

ANN models can be used as forecast tools for algae bloom in Liuhai lakes. The decision makers of related management department can take actions to reduce the intensity and time period of algae bloom occurrence.

References:

Gurbuz H, 2003. Predicting dominant phytoplankton quantities in a reservoir by using neural networks[J]. Hydrobiologia, 504: 133—141.

He X Q, 2004. Multi-variables statistical analysis [M]. Beijing: Renmin University of China press.

Jing N H, Xu F J, 2004. Water environment numerical simulation and visibility[M]. Beijing: Chemical Industry Press.

Karul C, 2000. Case studies on the use of neural networks in eutrophication modeling [J]. Ecological Modelling, 134: 145—152.

Lu X Y, Xu F L, Zhan W *et al.*, 2003. Current situation and development trend in lake eutrophication models [J]. Advance in Water Science, 14(6): 792—798.

Maier H R, Dandy G C, Burch M D, 1998. Use of artificial neural networks for modeling cyanobacteria *Anabaena* spp. In the river murray, south Australia[J]. Ecological Modeling, 105: 257—272.

Pei H P , Luo N N, Jiang Y, 2004. Applications of back propagation neural network for predicting the concentration of chlorophyll-a in West Lake[J]. Acta Ecologica Sinica, 24(2): 246—251.

Peng H, Guo S L, 2002. Numerical modeling of the hydrodynamic and water quality in the low reaches of the Hanjiang River [J]. Resource and Environment in the Yangtze Basin, 11 (4): 363—369.

Walter M, 2001. Predicting eutrophication effects in the Burrinjuck reservoir (Australia) by means of the deterministic model SALMO and the recurrent neural network model ANNA [J]. Ecological Modeling, 146: 97—113.

Wen X, Zhou L, Li X *et al.*, 2003. MATLAB neural network simulation and application[M]. Beijing: Science Press.

Wu H J, Lin Z Y, Gao S L, 2000. Application of artificial neural networks in the resources and environment management [J]. Resource and Environment in the Yangtze Basin, 9 (2): 237—241.

Xia J, Du M, 2000. Study on eutrophication synthetic water quality model and its application [J]. Shanghai Environment Science, 19 (7): 302—308.

Zhang W T, 2002. Advanced tutorial of statistic analysis for SPSS11 [M]. Beijing: Hope Electronic Press.

Zhou Y, Liu F, Wu D *et al.*, 1999. On the principle and method of lake water environment prediction [J]. Resource and Environment in the Yangtze Basin, 8(3): 305—311.

(Received for review July 6, 2005. Accepted October 8, 2005)