Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2012, Article ID 397473, 12 pages
doi:10.1155/2012/397473

Research Article

Freshwater Algal Bloom Prediction by Support Vector Machine in Macau Storage Reservoirs

Zhengchao Xie,[1] Inchio Lou,[1] Wai Kin Ung,[2] and Kai Meng Mok[1]

[1] *Faculty of Science and Technology, University of Macau, Taipa, Macau*
[2] *Laboratory & Research Center, Macao Water Supply Co. Ltd., Conselheiro Borja, Macau*

Correspondence should be addressed to Inchio Lou, iclou@umac.mo

Received 26 August 2012; Accepted 11 November 2012

Academic Editor: Sheng-yong Chen

Understanding and predicting dynamic change of algae population in freshwater reservoirs is particularly important, as algae-releasing cyanotoxins are carcinogens that would affect the health of public. However, the high complex nonlinearity of water variables and their interactions makes it difficult to model the growth of algae species. Recently, support vector machine (SVM) was reported to have advantages of only requiring a small amount of samples, high degree of prediction accuracy, and long prediction period to solve the nonlinear problems. In this study, the SVM-based prediction and forecast models for phytoplankton abundance in Macau Storage Reservoir (MSR) are proposed, in which the water parameters of pH, SiO_2, alkalinity, bicarbonate (HCO_3^-), dissolved oxygen (DO), total nitrogen (TN), UV_{254}, turbidity, conductivity, nitrate, total nitrogen (TN), orthophosphate (PO_4^{3-}), total phosphorus (TP), suspended solid (SS) and total organic carbon (TOC) selected from the correlation analysis of the 23 monthly water variables were included, with 8-year (2001–2008) data for training and the most recent 3 years (2009–2011) for testing. The modeling results showed that the prediction and forecast powers were estimated as approximately 0.76 and 0.86, respectively, showing that the SVM is an effective new way that can be used for monitoring algal bloom in drinking water storage reservoir.

1. Introduction

Freshwater algal bloom is one of water pollution problems that occurs in eutrophic lakes or reservoirs due to the presence of excessive nutrients. It has been found that most species of algae (also called phytoplankton) can produce various cyanotoxins including *microcystins*, *cylindrospermopsis*, and *nodularin*, which have a direct impact on the water treatment processes and consequently the health of public [1]. Thus, it is of great importance to understand the population dynamics of algae in the raw water storage units. However, modeling the algae population in such a complicated system is a challenge, as the physical, chemical,

and biological processes as well as the interaction among them are involved, resulting in the highly nonlinear relationship between phytoplankton abundance and various water parameters.

Computational artificial intelligence techniques have been developed as the efficient tools in recent years for predicting (without considering time series effect) or forecasting (considering time series effect) algal bloom. Previous studies [2] have used the principle component regression (PCR), that is, principal component analysis (PCA) followed by multiple linear regression (MLR), to predict chlorophyll-a levels, the fundamental index of phytoplankton. However, the intrinsic problem of PCR is that the variables dataset used as the input of the model has high complex nonlinearity, expecting that PCR alone is inadequate for prediction, and the prediction results were unsatisfactory. With the development of artificial intelligence models, artificial neural network (ANN) such as backpropagation (BP) was applied to predict the algal bloom by assessing the eutrophication and simulating the chlorophyll-a concentration. ANN is a well-suited method with self-adaptability, self-organization, and error tolerance, which is better than PCR for nonlinear simulation. However, this method has such limitations as requirement of a great amount of training data, difficulty in tuning the structure parameter that is mainly based on experience, and its "black box" nature that makes it difficult to understand and interpret the data [2, 3].

Considering the drawbacks of both the methods, recently support vector machine (SVM) started to be used for predicting the chlorophyll concentration. It is a new machine-learning technology based on statistical theory and derived from instruction risk minimization, which can enhance the generalization ability and minimize the upper limit of generalization error. Compared to ANN, SVM has advantages of only requiring a small amount of samples, high degree of prediction accuracy, and long prediction period by using kernel function to solve the nonlinear problems. It is believed that SVM will provide a new approach for predicting the phytoplankton abundance in the reservoirs [4]. Also, this black box model can be applied in other locations and other cases such as red tide.

In this study, we attempted to develop an SVM-based predictive model to simulate the dynamic change of phytoplankton abundance in Macau Reservoir given a variety of water variables. The measured data from 2001 to 2011 were used to train and test the model. The present study will lead to a better understanding of the algal problems in Macau, which will help to develop later guidelines for forecasting the onset of algae blooms in raw water resources.

2. Materials and Methods

Macau is situated 60 km southwest of Hong Kong and experiences a subtropical seasonal climate that is greatly influenced by the monsoons. The difference of temperature and rainfall between summer and winter is significant though not great. Macau Main Storage Reservoir (MSR) (Figure 1), located in the east part of Macau peninsula, is the biggest reservoir in Macau with the capacity of about 1.9 million m^3 and the water surface area of 0.35 km^2. It is a pumped storage reservoir that receives raw water from the West River of the Pearl River network and can provide water supply to the whole areas of Macau for about one week. MSR is particularly important as the temporary water source during the salty tide period when high salinity concentration is caused by intrusion of sea water to the water intake location. In recent years, there were reports (Macao Water Supply Co. Ltd, unpublished data) that the reservoir experienced algal blooms, and the situation appeared to be worsening.

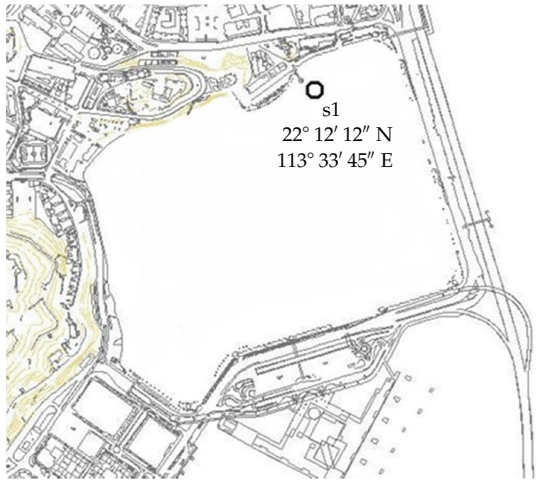

Figure 1: Location of the MSR.

Macau Water Supply Co. Ltd. is responsible for water-quality monitoring and management. Location in the inlet of the reservoir was selected for sampling. Samples were collected in duplicate monthly from May 2001 to February 2011 at 0.5 m from the water surface. A total of 23 water quality parameters, including hydrological, physical, chemical, and biological parameters, were monitored monthly. Precipitation was obtained from Macau Meteorological Center (http://www.smg.gov.mo/www/te_smgmail.php). Imported volume, exported volume, and water level were recorded by the inlet and outlet flow meters, based on which the hydraulic retention time (HRT) can be calculated. Turbidity, temperature, pH, conductivity, chloride (Cl^-), sulfate (SO_4^{2-}), silicon (SiO_2), alkalinity, bicarbonate (HCO_3^-), dissolved oxygen (DO), ammonium (NH_4^+), nitrite (NO_2^-), nitrate (NO_3^-), total nitrogen (TN), phosphorus (PO_4^{3-}), total phosphorus (TP), suspended solid, total organic carbon (TOC) and UV_{254}, and iron (Fe) were measured according to the standard methods [5]. The phytoplankton samples were fixed using 5% formaldehyde and transported to laboratory for microscopic counting.

In this work, correlation analysis was conducted to identify the water parameters which were significantly correlated with phytoplankton abundance (Table 1). Only the parameters with the correlation coefficients greater than 0.3 are selected as inputs in the SVM models. It was also noted that the parameters selected in forecast models are different from those in the prediction models, as the water parameters in previous data were also used in the correlation analysis.

As a prediction algorithm, SVM was firstly proposed by Vapnik [6] and is an effective tool for data classification and regression. The SVM is fundamentally based on Mercer core expansion theorem which maps sample space to a higher-dimension or even unlimited dimension feature space by nonlinear mapping functions (kernel function) [7]. In SVM, it transforms the problem of searching for an optimal linear regression hyperplane to a convex programming problem of solution for a convex restriction condition. Moreover, SVM can provide the global optimum solution because the problem in SVM is transformed to finding the solution to the quadratic programming.

SVM is selected in this work because of its advantages over other "black box" modeling approaches such as ANN as listed as follows [8].

(1) The architecture of the estimated function does not have to be determined before training. Input data of any arbitrary dimensionality can be treated with only linear costs in the number of input dimensions.

(2) SVM treats the regression as a quadratic programming problem of minimizing the data-fitting error plus regularization, which produces a global (or even unique) solution.

(3) SVM combines the advantages of multivariate nonlinear regression in that only a small amount of data is required to produce a good generalization. In addition, the weakness of the transformational models in multivariate nonlinear regression can be overcome by mapping the data points to a sufficiently high-dimensional feature space.

(4) Results obtained from SVM are easy to interpret.

In SVM, the whole process consists of several layers. The input vectors are put in the first layer. Suppose that the training datasets are

$$(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N). \tag{2.1}$$

A nonlinear mapping $\psi(\cdot)$ is used to map samples from former space R^n to feature space [9]:

$$\psi(x) = \big(\phi(x_1), \phi(x_2), \ldots, \phi(x_N)\big). \tag{2.2}$$

Then, in this higher-dimension feature space, optimal decisions function is

$$f(x) = w\phi(x) + b, \tag{2.3}$$

where b is the bias constant or the threshold which can be calculated as introduced in [8].

In this way, nonlinear prediction function is transformed to linear prediction function in higher-dimension feature space [9]. Note that parameters used in equations will be introduced later in this section. The SVM needs to find out the solution to minimize the following functional:

$$\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}(\xi_i + \xi_i^*),$$

$$\text{s.t.} \begin{cases} y_i - w^T\phi(x_i) - b \le \varepsilon + \xi_i \\ w^T\phi(x_i) + b - y_i \le \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \ge 0. \end{cases} \tag{2.4}$$

As introduced previously, SVM can provide the global optimum solution because the problem in SVM is transformed to finding the solution to the quadratic programming. So,

Table 1: Correlation analysis of prediction and forecast model.

Parameters	Prediction model	Forecast model Time lagged (month)		
		t-1	t-2	t-3
Turbidity	−0.03	0.00	−0.01	−0.06
Temperature	0.19	0.21	0.19	0.14
pH	**0.49**	**0.42**	**0.38**	**0.33**
Conductivity	−0.08	0.01	0.14	0.21
Cl^-	0.01	0.10	0.22	0.28
SO_4^{2-}	−0.03	0.03	0.14	0.22
SiO_2	**0.33**	**0.31**	0.16	0.04
Alkalinity	**−0.34**	**−0.30**	−0.21	−0.12
HCO_3^-	**−0.46**	**−0.40**	**−0.32**	−0.24
DO	**0.39**	**0.35**	**0.34**	**0.31**
NO_3^-	−0.29	−0.22	−0.22	−0.15
NO_2^-	−0.10	−0.08	−0.02	0.03
NH_4^+	0.11	0.10	0.08	0.25
TN	**0.68**	**0.60**	**0.53**	**0.46**
UV_{254}	**0.56**	**0.55**	**0.48**	**0.47**
Fe	−0.14	−0.06	−0.04	−0.08
PO_4^{3-}	0.02	0.06	0.06	0.03
TP	0.08	0.05	0.02	0.00
Suspended solid	**0.31**	**0.35**	**0.31**	0.23
TOC	**0.38**	**0.33**	0.29	**0.35**
HRT	−0.12	−0.11	−0.13	−0.16
Water level	0.13	0.05	0.01	−0.02
Precipitation	−0.09	0.05	0.11	0.06
Phytoplankton abundance	—	**0.82**	**0.71**	**0.62**

the minimization problem shown in (2.4) could be transformed to finding the solution to maximize the following equation [5, 9–11]:

$$\max_{\alpha,\alpha^*} = -\frac{1}{2}\sum_{i=1,l=1}^{N}(\alpha_i-\alpha_i^*)(\alpha_l-\alpha_l^*)\langle\phi(x_i),\phi(x_l)\rangle - \varepsilon\sum_{i=1}^{N}(\alpha_i+\alpha_i^*)+\sum_{i=1}^{N}y_i(\alpha_i-\alpha_i^*)$$
$$\text{s.t.}\begin{cases}\sum_{i=1}^{N}(\alpha_i-\alpha_i^*)=0,\\ \alpha_i,\alpha_i^*\in[0,C].\end{cases} \tag{2.5}$$

where $\alpha, \alpha^*, \eta, \eta^* \geq 0$ are Lagrange multipliers.

According to Mercer's condition, in SVM the inner product $\langle \phi(x), \phi(x_i) \rangle$ can be defined through a kernel function $K(x, x_i)$. There are several kernel functions that are available as follows:

(1) linear: $K(x_i, x_j) = x_i^T x_j$,

(2) polynomial: $K(x_i, x_j) = (\gamma x_i^T x_j + r)^d$,

(3) radial basis function: $K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2)$,

(4) sigmoid: $K(x_i, x_j) = \tanh(\gamma a_i^T + r)$.

For these four kernel functions, in general, the RBF kernel function is a reasonable first choice [9]. This kernel function nonlinearly maps samples into a higher-dimensional space. So, unlike the linear kernel, it can handle the case when the relation between class labels and attributes is nonlinear. The second reason is that the RBF kernel function has a less number of hyperparameters which influences the complexity of model selection. Finally, the RBF kernel has fewer numerical difficulties [12–16].

As shown in the kernel function mentioned previously, there are three parameters which need to be specified in the application of SVM: (1) capacity parameter C that controls the trade-off between maximizing the margin and minimizing the training error. If C is too small, then insufficient stress will be placed on fitting the training data. If C is too large, then the algorithm will overfit the training data. (2) RBF width parameter γ: the γ value is important in the RBF model and can lead to under- or over-fitting in prediction. A very large value of γ may lead to overfitting, and all the support vectors distances are taken into account, while in case of a very small γ, the machine will ignore most of the support vectors leading to failure in the trained point prediction [9]. (3) Insensitive loss function ε: if ε is too large, then it will result in less support vectors, and consequently, the resulting regression model may yield large prediction errors on unseen future data [10]. In this work, in order to prevent overtraining, an internal cross-validation [11] during construction of SVR models is adopted to have a good combination of the three parameters C, γ, and ε. Now, after the introduction of SVM, the following section gives the numerical results from the application of SVM.

With the above introduction of SVM, it is necessary to present performance indicators. The performance of models was evaluated using the following indicators: square of correlation coefficient (R^2) that provides the variability measure for the data reproduced in the model; mean absolute error (MAE) and root mean square error (RMSE) that measure residual errors, providing a global idea of the difference between the observation and modeling. The indicators were defined as follows:

$$
\begin{aligned}
R^2 &= 1 - \frac{F}{F_o}, \\
F &= \sum \left(Y_i - \widehat{Y}_i \right)^2, \\
F_o &= \sum \left(Y_i - \overline{Y_i} \right)^2, \\
\mathrm{MAE} &= \frac{1}{n} \sum_{i=1}^{n} \left(\widehat{Y}_i - Y_i \right)^2, \\
\mathrm{RMSE} &= \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(\widehat{Y}_i - Y_i \right)^2},
\end{aligned}
\tag{2.6}
$$

Table 2: Performance indexes of the prediction and forecast models.

Performance index	Prediction model				Forecast model			
	Accuracy performance (training set)		Generalization performance (testing set)		Accuracy performance (training set)		Generalization performance (testing set)	
	ANN	SVM	ANN	SVM	ANN	SVM	ANN	SVM
R^2	0.752	0.760	0.749	0.758	0.758	0.863	0.760	0.863
RMSE	0.307	0.307	0.316	0.351	0.299	0.229	0.306	0.264
MAE	0.238	0.243	0.243	0.274	0.229	0.127	0.247	0.226

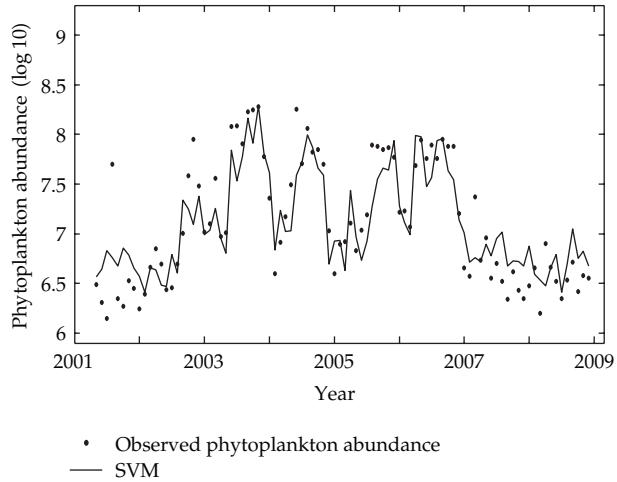

Figure 2: Observed and predicted phytoplankton level for the training and validation dataset of the prediction models.

where n is the number of data; Y_i and $\overline{Y_i}$ are observation data and the mean of observation data, respectively, and $\widehat{Y_i}$ is the modeling results.

3. Results and Discussion

The correlation of $\log_{10}$ phytoplankton and water parameters for forecast model and prediction model was shown in Table 2. Parameters with correlation coefficients greater than 0.3 (highlighted in bold) will be retained in the models. It was also noted that the parameters selected in forecast models are different from those in the prediction models, as the water parameters in previous data (past record) were also used in the correlation analysis. In the forecast models of SVM, phytoplankton abundance (t) is a function of water parameter (t-1), water parameter (t-2), and water parameters (t-3), where t-1, t-2, and t-3 represent the 1 month, 2 months, and 3 months prior to time t. Thus, there were only 9 parameters used in the prediction models and 23 time-lagged parameters selected for the forecast models.

After the correlation analysis, it comes to the testing of the models invoked two parts, the accuracy performance and the generalization performance. Accuracy performance is to test the capability of the model to predict the output for the given input set that is originally used to train the model, while generalization performance is to test the capability of the model to predict the output for the given input sets that were not in the training set. In order

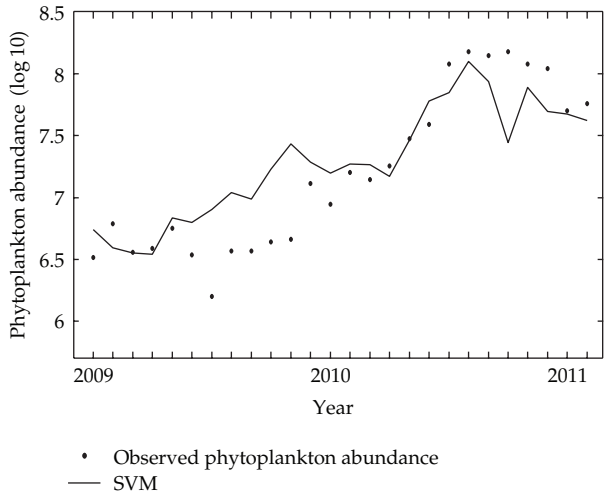

Figure 3: Observed and predicted phytoplankton level for the testing dataset of the prediction models.

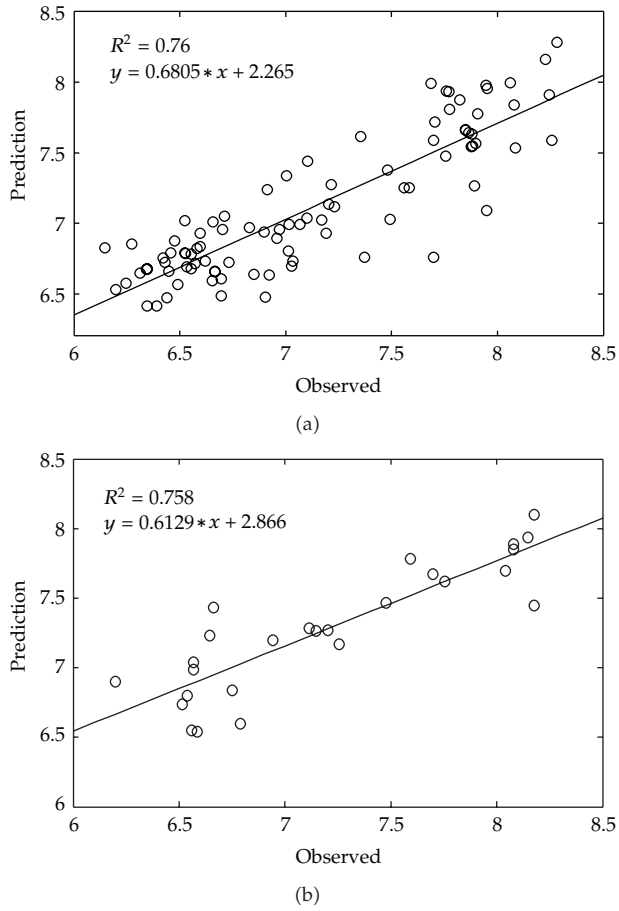

Figure 4: SVM result for the training and validation (a) and testing (b) data set.

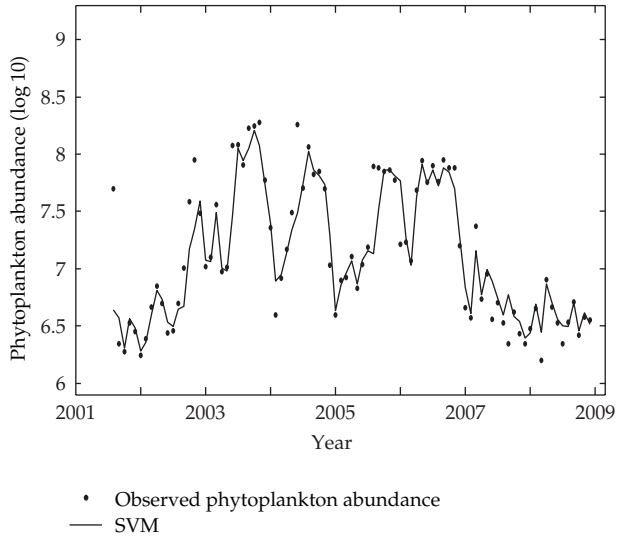

Figure 5: Observed and predicted phytoplankton level for the training and validation dataset of the forecast models.

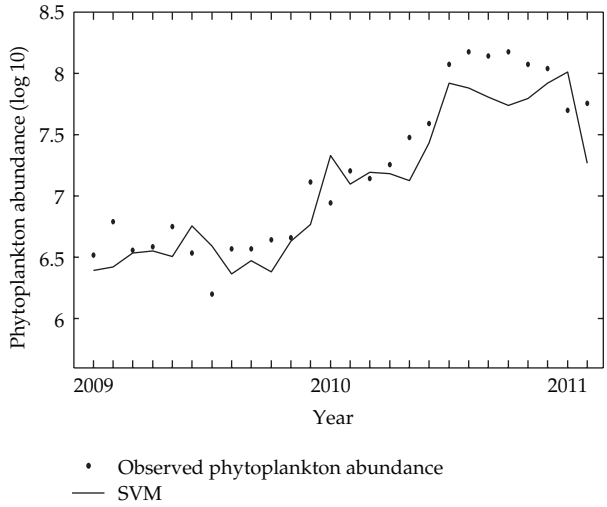

Figure 6: Observed and predicted phytoplankton level for the testing dataset of the forecast models.

to prevent the model that is memorizing the inputs instead of generalized learning, both performance checks need to be considered. In the present research, the performance indexes for SVM-based models were averaged with 50 runs.

In the application of SVM in this work, for the predication model, after the correlation analysis, 9 parameters such as pH, SiO_2 are selected as the independent variables, and phytoplankton abundance is selected as the induced variable (target value). Then, the data from May 2005 to December 2008 are used to train the model, and data from January 2009 to February 2011 are used to test the model. In the training process, the cross-validation approach as mentioned previously is adopted to obtain the optimal combination of parameters for the testing. Specifically, the training data are divided into 10 about the same

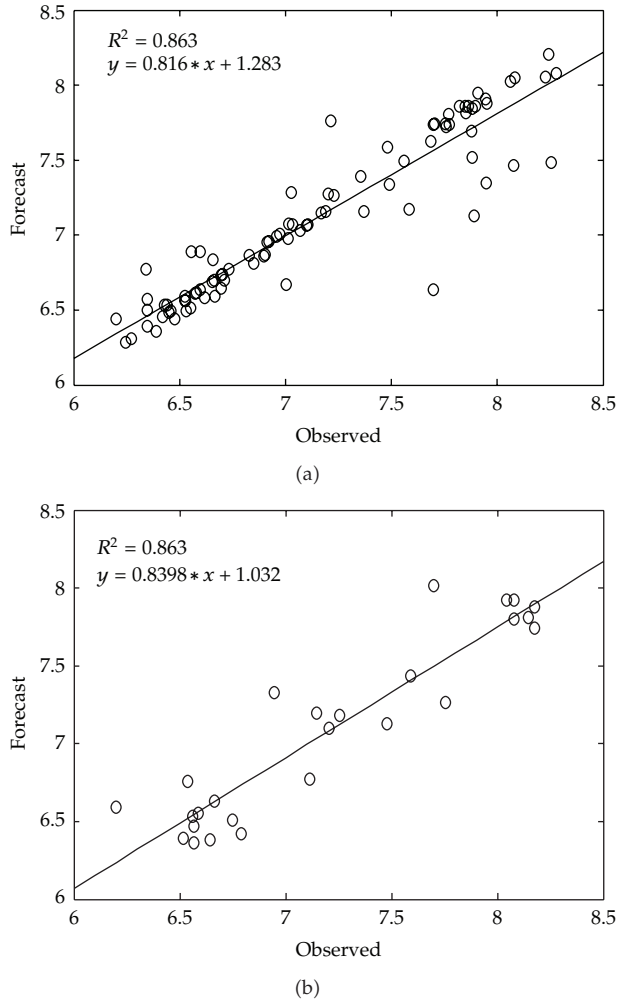

Figure 7: SVM result for the training and validation (a) and testing (b) data set.

size groups that are 9 groups for training, and the rest 1 group is used to test the model trained by the previous 9 groups' data. Then, this (9 groups training and 1 group testing) is repeated for 9 times (10 times in total). And then, parameters of the one process which has the best testing performance in these 10 repeats will be used as the optimal parameters combination in the "real" testing process which has the data from January 2009 to February 2011. The forecast model basically follows the same steps of the prediction model, while the only difference between these two models is that effect of time series which is included in the forecast model. So, in the forecast model, only the previous three months' data are included in the training process.

The performance of prediction and forecast models was shown in Table 2. Compared to our previous studies using ANN, the SVM has a similar performance for prediction model with R^2 of 0.758, RMSE of 0.351, and MAE of 0.274, while it has much better performance for forecast model with R^2 of 0.863, RMSE of 0.229, and MAE of 0.127, for testing. To balance the R^2 in training and testing, we defined the equal values for both data sets as the performance of the models. The observed data versus the modeling data were shown in Figures 4 and 7,

and the observed and modeling phytoplankton abundance changes over time were listed in Figures 2, 3, 5, and 6.

These results confirmed that SVM can handle well the nonlinear relationship between water parameters and phytoplankton abundance.

4. Conclusions

The SVM-based prediction and forecast models for phytoplankton abundance in MSR are proposed in this study. 15 water parameters with the correlation coefficients against phytoplankton abundance greater than 0.3 were selected, with 8-year (2001–2008) data for training and cross-validation and the most recent 3 years (2009–2011) for testing. The results showed that the forecast model has better performance with the R^2 of 0.863 than prediction model with the R^2 of 0.760, implying that the algal bloom problem is a complicated nonlinear dynamic system that is affected not only by the water variables in current month, but also by those in a couple of previous months. In addition, compared to ANN in our previous studies, SVM in the study showed superior forecast power, while similar prediction power in terms of regression coefficient. These results will provide an effective way for water quality monitoring and management of drinking water storage reservoirs. In addition, additional numerical approaches and optimization algorithms can be applied to enhance the performance [17–19].

Acknowledgments

The authors thank Macao Water Supply Co. Ltd. for providing historical data of water quality parameters and phytoplankton abundances. The financial support from the Fundo para o Desenvolvimento das Ciências e da Tecnologia (FDCT) (Grant no. FDCT/016/2011/A) and Research Committee at University of Macau is gratefully acknowledged.

References

[1] Z. Selman, S. Greenhalgh, and R. Diaz, *Eutrophication and Hypoxia in Coastal Areas: A Global Assessment of the State of Knowledge*, World Resources Institute, Washington, DC, USA, 2008.

[2] J. Pallant, I. Chorus, and J. Bartram, "Toxic cyanobacteria in water," in *SPSS Survival Manual*, 2007.

[3] R. Hecht-Nielsen, "Kolmogorov's mapping neural network existence theorem," in *Proceedings of the 1st IEEE Internetional Jopint Conference of Neural Networks*, New York, NY, USA, 1987.

[4] L. L. Rogers and F. U. Dowla, "Optimization of groundwater remediation using artificial neural networks with parallel solute transport modeling," *Water Resources Research*, vol. 30, no. 2, p. 457, 1994.

[5] APHA, *Standard Methods for the Examination of Water and Wastewater*, American Public Health Association (APHA), American Water Works Association (AWWA) & Water Environment Federation (WEF), 2002.

[6] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer, New York, NY, USA, 1995.

[7] T. A. Stolarski, "A system for wear prediction in lubricated sliding contacts," *Lubrication Science*, vol. 8, no. 4, pp. 315–351, 1996.

[8] K. Li, *Automotive engine tuning using least-square support vector machines and evolutionary optimization [Ph.D. thesis]*, University of Macau, 2011.

[9] Z. Liu, X. Wang, L. Cui, X. Lian, and J. Xu, "Research on water bloom prediction based on least squares support vector machine," in *Proceedings of the WRI World Congress on Computer Science and Information Engineering (CSIE '09)*, pp. 764–768, April 2009.

[10] A. J. Smola and B. Scholkopf, 2003, http://alex.smola.org/papers/2003/SmoSch03b.pdf.

[11] H. Wang and D. Hu, "Comparison of SVM and LS-SVM for regression," in *Proceedings of the International Conference on Neural Networks and Brain Proceedings (ICNNB '05)*, pp. 279–283, October 2005.

[12] C. W. Hsu and C. C. Chang, *A Practical Guide to Support Vector Classification*, 2003.

[13] U. Çaydaş and S. Ekici, "Support vector machines models for surface roughness prediction in CNC turning of AISI 304 austenitic stainless steel," *Journal of Intelligent Manufacturing*, vol. 23, pp. 639–650, 2012.

[14] E. Avci, "A new expert system for diagnosis of lung cancer: GDA-LS_SVM," *Journal of Medical Systems*, vol. 36, pp. 2005–2009, 2012.

[15] E. Çomak and A. Arslan, "A biomedical decision support system using LS-SVM classifier with an efficient and new parameter regularization procedure for diagnosis of heart valve diseases," *Journal of Medical Systems*, vol. 36, pp. 549–556, 2012.

[16] Y. Xu, X. Chen, and Q. Li, "INS/WSN-integrated navigation utilizing LS-SVM and H_∞ filtering," *Mathematical Problems in Engineering*, vol. 2012, Article ID 707326, 19 pages, 2012.

[17] C. Cattani, S. Chen, and G. Aldashev, "Information and modeling in complexity," *Mathematical Problems in Engineering*, vol. 2012, Article ID 868413, 3 pages, 2012.

[18] S. Chen, Y. Zheng, C. Cattani, and W. Wang, "Modeling of biological intelligence for SCM system optimization," *Computational and Mathematical Methods in Medicine*, vol. 2012, Article ID 769702, 10 pages, 2012.

[19] P. Lu, S. Chen, and Y. Zheng, "Artificial intelligence in civilengineering," *Mathematical Problems in Engineering*. In press.