Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2012, Article ID 237693, 17 pages
doi:10.1155/2012/237693

Research Article

Sludge Bulking Prediction Using Principle Component Regression and Artificial Neural Network

Inchio Lou[1] and Yuchao Zhao[2]

[1] *Department of Civil and Environmental Engineering, Faculty of Science and Technology, University of Macau, Avenue Padre Tomás Pereira, Taipa 999078, Macau*
[2] *State Key Joint Laboratory of Environmental Simulation and Pollution Control, School of Environment, Beijing Normal University, Beijing 100875, China*

Correspondence should be addressed to Inchio Lou, iclou@umac.mo
and Yuchao Zhao, zhaoy@bnu.edu.cn

Received 5 August 2012; Revised 22 October 2012; Accepted 25 October 2012

Academic Editor: Siamak Talatahari

Sludge bulking is the most common solids settling problem in wastewater treatment plants, which is caused by the excessive growth of filamentous bacteria extending outside the flocs, resulting in decreasing the wastewater treatment efficiency and deteriorating the water quality in the effluent. Previous studies using molecular techniques have been widely used from the microbiological aspects, while the mechanisms have not yet been completely understood to form the deterministic cause-effect relationship. In this study, system identification techniques based on the analysis of the inputs and outputs of the activated sludge system are applied to the data-driven modeling. Principle component regression (PCR) and artificial neural network (ANN) were identified using the data from Chongqing wastewater treatment plant (CQWWTP), including temperature, pH, biochemical oxygen demand (BOD), chemical oxygen demand (COD), suspended solids (SSs), ammonia ($NH_4{}^+$), total nitrogen (TN), total phosphorus (TP), and mixed liquor suspended solids (MLSSs). The models were subsequently used to predict the sludge volume index (SVI), the indicator of the bulking occurrence. Comparison of the results obtained by both models is also presented. The results showed that ANN has better prediction power ($R^2 = 0.9$) than PCR ($R^2 = 0.7$) and thus provides a useful guide for practical sludge bulking control.

1. Introduction

Sludge bulking is the most common solid separation problem in activated sludge problem, which is caused by the excessive growth of filamentous bacteria extending outside the flocs, thus interfering with the settling of activated sludge. It has been reported that over 50%

of the wastewater treatments in US experience bulking [1]. Bulking leads to high level of total suspended solids in effluent that exceeds the discharge permit limitation and subsequently loses activated sludge in the aeration basin, resulting in the deterioration of wastewater treatment process [1]. Sludge setting and compaction are often quantified using sludge volume index (SVI). When SVI reaches 150 mL/g, bulking can be considered to happen.

Various theories and factors, such as kinetic selection theory [2–4], filamentous backbone theory [5], substrate diffusion limitation [6], storage phenomena [7, 8], and the difference in the decay rates between filaments and floc formers [9], have been proposed and extensively studied for explaining the competition between filaments and floc formers. However, no single or combined proposed mechanisms can explain completely the sludge bulking problem; for example, the uncertainty about the factors triggering the filaments, growth is still unclear. The current efforts to study sludge bulking problem rely mostly on experimental observation of filamentous bacteria population in the system, while some experimental results could lead to contradictory conclusions. Thus it is difficult to formulate deterministic mathematical models for predicting the filaments population, though some existing models were developed [9, 10].

Developing a model that could predict in real time with reasonable accuracy the potential for bulking is of great practical importance, as it can be used to improve the treatment plant efficiency and cost saving [11]. The complexity of the problem can be overcome by applying data-driven model for the whole system, rather than the breaking down of the system into small components described individually, in which only the inputs and outputs of the system are taken into consideration. One major advantage of the data-driven models over mechanistic models is that they require minimal information of the intrinsic processes of the system.

In PCR, PCA is first used to convert a set of observations of possibly correlated var-iables by orthogonal transformation into a set of values of uncorrelated variables called PCs, thus reducing the complexity of multidimensional system by maximization of component loadings variance and elimination of invalid components. PCA has been used alone or in combination with other methods, such as MLR, to model aquatic environmental and ecological processes including algal blooms problem in freshwater reservoirs [12–14]. From these studies, only the PCs with eigenvalues greater than 1 were selected for MLR, which can explain the high percentage of total variation of the environmental variables in PCA. It is followed by the MLR to check if the chlorophyll-a, cyanobacteria abundance, or microcystin concentrations could be explained by environmental variables and to be used for further prediction. On the other hand, ANN is regarded as an efficient tool for modeling and forecasting due to its wide range of applicability and capability to treat complicated nonlinear problems. After training, ANN can be used to predict the output with the new independent input parameter; thus, it is appropriate for modeling the water parameters data [15]. It was reported that ANN provided better results than PCR model, particularly in handing collinearity and nonlinear structures of forecast problems [12, 16–18].

In the aspect of wastewater treatment processes, the application of PCR and ANN as popular data-driven approaches has been widely researched in the literature. Belanche et al. [19] used ANN as error predictor to improve the accuracy of an existing mechanistic model of activated sludge process by coupling both techniques. Five key variables including effluent SS, COD, ammonia, mixed liquor oxygen, and volatile SS were simulated and predicted. Two steps were involved: optimization of model parameters was first investigated using

the downhill simplex method to minimize the sum of the square errors between observation and prediction; then ANN models were used to predict the remaining errors of the optimized mechanistic model. This study used over 10 days' 6–9 h measurements from the activated sludge treatment plant located at Norwich, England. Though the study is based on the real data, the models were not used for predicting bulking phenomena.

Côté et al. [20] developed and applied ANN models for the prediction of future bulking episodes. The simple prediction models based on the online data (flow rates), analytical data (COD, BOD), and qualitative data (presence of foam, filamentous bacteria, microfauna, and appearance) were developed for the effluent TSS that is an indicator of plant performance. The data came from the WWTP from Catalonia, Spain, in 609 consecutive days. Through the combined use of the rough set theory and ANN, the reasonable prediction models are found to show the different importance of variables and provide insight into the processes' dynamics. However, compared to SVI, TSS was not a good indicator for bulking, though during bulking episodes, the effluent TSS undoubtedly increases. Besides, the parameters sets used for selecting the significant variables are incomplete. For example, the important variables including temperature, pH, TN, and TP are not included in the selection, resulting in losing the key information for explaining the bulking phenomena.

A recent study [21] utilized a self-organizing radial basis function (SORBF) neural network method to predict the evolution of SVI. The hidden nodes in the SORBF neural network can be grown or pruned based on the node activity and mutuality to achieve the appropriate network complexity and overall computational efficiency. The performance of this method was verified in a real WWTP. This method enhanced the capacity of the RBF model to adapt to nonlinear dynamic system and thus yielded more accurate predictions than the other method. However, in this study only limited input parameters, influent flow rate, DO, pH, BOD, COD, and TN were included, which are not enough to explain sludge bulking mechanisms.

Considering the drawbacks of previous studies using ANN in wastewater treatment system, the purpose of the present study is to analyze bulking problems of CQWWTP that used the A/A/O treatment process that has not been discussed before, based on more complete daily variables including temperature, pH, BOD, COD, SS, NH_4^+, TN, TP, and MLSS for the whole year. These variables provide more complete data input to explain the bulking mechanisms, in spite of applying only the data-driven models developed in the study. The models can be used to evaluate the relative influence of the operational conditions, influents characteristics, and activated sludge concentrations on the SVI and to predict the SVI values using PCR and ANN. The comparisons of both models in this study and the prediction model developed by Han and Qiao [21] were made to select the best prediction method for wastewater treatment management. The key contributions of this paper not only focus on the mathematical modeling itself, but also take the complete main factors that affect the bulking into consideration, by integrating all of those potential mechanistic bulking causative variables into both models, though only the data-driven models were applied.

The rest of this paper was organized as follows. The study area and data source of CQWWTP were first introduced concisely, followed by modeling approaches (PCR and ANN) formulation and the performance indicators used for evaluation in Section 2. Section 3 presented and discussed the modeling results performed by PCR and ANN, respectively, and made comparisons between both methods. The conclusion was drawn finally in Section 4.

Figure 1: Location of CQWWTP.

2. Materials and Methods

2.1. Study Area

Chongqing is the biggest city in Western China. Like Beijing, Tianjin, and Shanghai, it is directly under the central government of China. The city has grown very quickly during the last 10 years, with the population of 31 millions and the area of 82,400 km^2. There is now a big effort to collect and treat the wastewater, due to the recent achievement of the three-gorge dam in the downstream. CQWWTP (29.601615 in latitude and 106.634133 E in longitude, Figure 1), one of biggest WWTP in Chongqing, is designed to have a capacity of an average flow rate of 300,000 m^3/d and about 750,000 person equivalents (in carbon, nitrogen, and phosphorus). CQWWTP uses conventional A/A/O (anaerobic/anoxic/aerobic) treatment processes (Figure 2) that are susceptible to sludge bulking. It was reported that 36% of sludge experience bulking in the year of 2010, and the situation appeared to be worsening in the recent years, particularly in the springs.

2.2. Data Source

Sludge samples were collected daily in the reaction tank over the year of 2010. The monitored parameters included operational conditions (temperature and pH), influent characteristics (BOD, COD, SS, NH_4^+, TN, and TP), and activated sludge concentrations (MLSS). Water samples were preserved, delivered, and analyzed using the standard methods of the American Public Health Association [22].

Figure 3 showed the changes of water parameters over the time, with the simple statistical analysis shown in Table 1. It was showed that the pH is maintained stable over the range of 7.6–8.2. The water temperatures matched the atmospheric temperatures that are low in the winter and high in the summer. The BOD and COD concentrations in

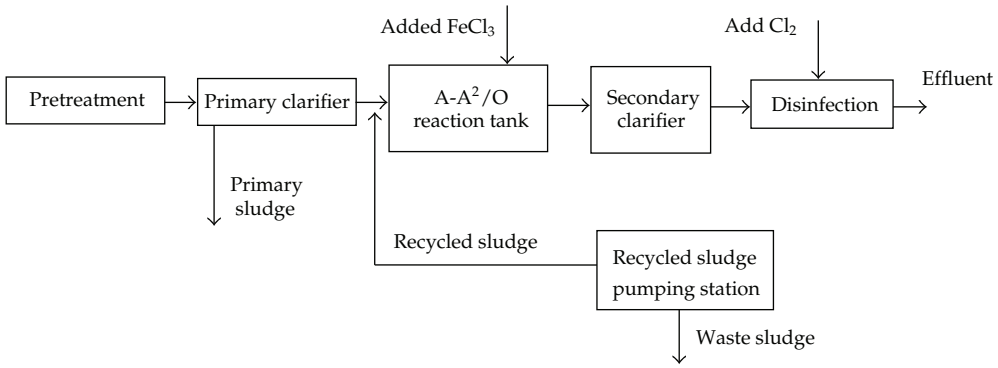

Figure 2: CQWWTP treatment processes.

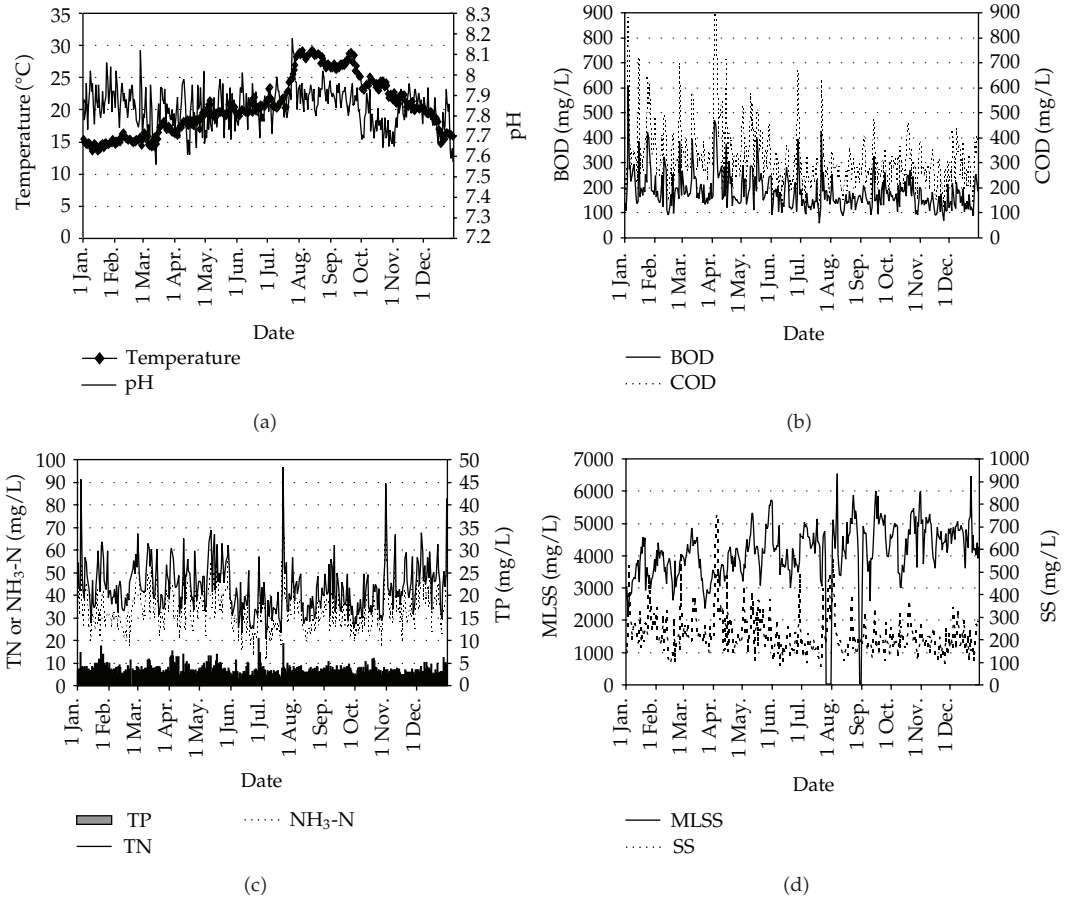

Figure 3: Change of water parameters over time in 2010. (a) Temperature and pH; (b) MLSS and SS; (c) BOD and COD; (d) TN/NH_3-N (lines) and TP (bar chart).

Table 1: Water quality characteristics of CQ from Jan. 1, 2010 to Dec. 31, 2010.

Parameter	Max.	Min.	Mean	Std. dev.
Temp.	29.2	13.5	20.499	4.395
MLSS	6530	14	4088.363	999.860
BOD	609	58	184.975	74.802
COD	899	113	305.094	122.215
SS	755	80	224.798	99.811
pH	8.18	7.56	7.840	0.099
TP	10.4	1.38	3.772	1.220
TN	96.9	20.6	43.460	11.295
NH_4^+-N	82	11.2	32.875	9.334
SVI	298	27	123.141	67.893

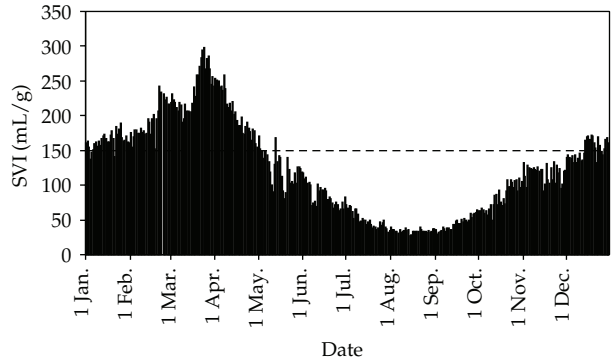

Figure 4: Change of SVIs over time in 2010.

the influents fluctuated from time to time, with high standard deviations of 74.8 mg/L for BOD and 113 mg/L for COD. However the BOD/COD ratios were within 0.53–0.8 for 85% of data, which were within the normal range of municipal wastewater, indicating that it is readily biodegradable wastewater. Similarly, the nitrogen and phosphorus concentrations in the influents fluctuated with 2-3 times higher or lower than the average values, which were believed to be the highly possible reasons that affected the growth of the bulking-causing filamentous bacteria in the reaction tank afterward. Due to the instability of the wastewater characteristics and the occurrence of bulking, the MLSS in the aeration tank cannot keep stable, ranging from 2000 mg/L to more than 6500 mg/L. It was also noted that the closely zero concentrations at the end of July and August were due to the measurement errors.

Figure 4 showed the change of SVIs over time, which clearly indicates that the bulking mostly happened in the springs from Jan. to April, with the SVIs greater than 150 mL/g. On the other hand, bulking levels were low in the summers from July to September, with the SVI around 50 mL/g. When Figure 4 was compared with Figure 3, it was found that there is a correlation between temperature and SVI, showing that bulking in CQWWTP mostly happened in the spring and nonbulking occurred in the summer. This relationship would be further investigated in the following statistical studies.

2.3. Modeling Approaches

Two different modeling techniques, PCR and ANN, were analyzed and applied to model the SVI data from CQWWTP. The measured SVI values in the reaction tank reflected the bulking levels, which in turn depended on the various variables including physical, chemical, and biological water parameters and the interaction among them. They all affect the growth of the filamentous bacteria in the biological WWTP. From the literature review [23, 24], those important parameters include temperature, pH, BOD, COD, SS, NH_4^+, TN, TP, and MLSS. Temperature and pH are the growth environment for microorganisms. The temperature increases the growth of floc-forming bacteria and filamentous bacteria and strengthen, their interaction and competition. The optimum pH in the reaction tank is 7–7.5, and pH below 6.0 would favor the growth of fungi that induces filamentous bulking. SS and MLSS is the indicator of the amount of activated sludge. The wastewater compositions, BOD/COD, NH_4^+/TN, and TP are the carbon source, nitrogen source and phosphorus source for microorganisms, respectively. High carbohydrate components and low substrate concentrations with low F/M (food/microorganism) ratios appear to be conducive to sludge bulking [1]. Besides, the deficiency of nitrogen and phosphorus results in the production of nutrient-deficient floc particles and loss of settleability in reaction tanks. Thus all these parameters were taken as the input of the models.

2.3.1. PCR

PCR is divided into two parts, principle component analysis (PCA) and multiple linear regressions (MLRs). PCA is a multivariate statistical method which uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of uncorrelated variables called principle components (PCs), thus reducing the complexity of multidimensional system by maximization of component loadings variance and elimination of invalid components. MLR attempts to model the relationship between two or more explanatory variables and a response variable by fitting a linear equation to observed data. The eigenvalues of the standardized matrix are calculated from following equation:

$$|C - \lambda I| = 0, \tag{2.1}$$

where C is the correlation matrix of the standardized data, λ is the eigenvalues, and I is the identity matrix. Then the weights of the variables in the PC are calculated by

$$|C - \lambda I|W = 0, \tag{2.2}$$

where W is the matrix of the weights.

Varimax rotation was used to obtain values of rotated factor loadings for evaluating the influence of each variable in the PC. These loadings represent the contribution of each variable in a specific principle component.

In this study, the PCA was performed on the water parameters to rank their relative significance and to describe their interrelation patterns as well as on the phytoplankton population levels. The stepwise option was used to choose the principle components, and the principle component scores of the selected parameters were used as independent variables in the MLR to check if the occurrences of phytoplankton could be explained by environmental

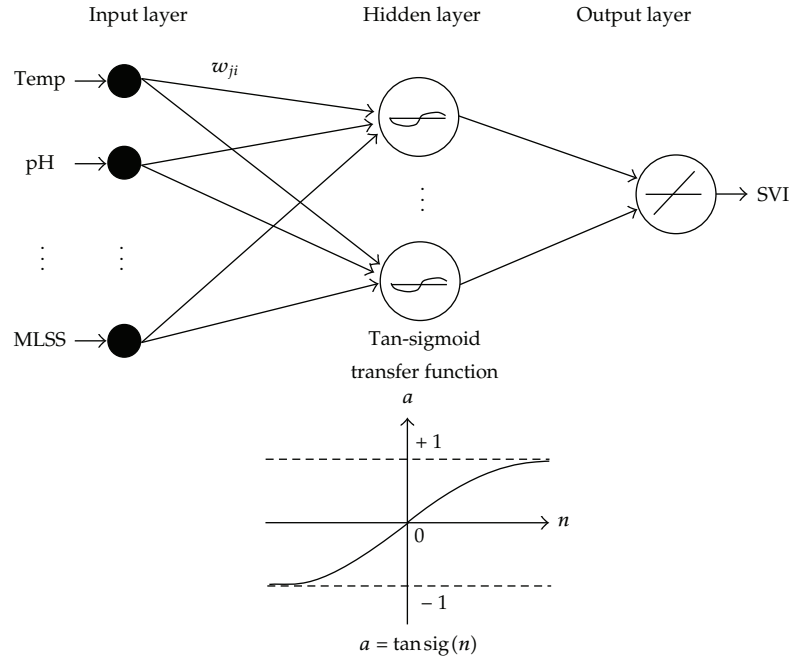

Figure 5: Three-layer feedforward network with Tan-sigmoid function.

variables as well as to predict the phytoplankton abundance. Since phytoplankton abundance did not show normal distribution, logarithmic transformation was applied to phytoplankton data to be used in PCA. Kaiser-Meyer-Olkin (KMO) measure of sample adequacy and Bartlett's test of Sphericity were used to verify the applicability of PCA [25]. PCA and MLR were carried out using PASW 19 software package (SPSS Inc.). The detail procedure of PCR has been described in our previous study [26].

2.3.2. ANN

ANN computing is a new approach to system modeling and identification, with the attractive self-learning system. Different from conventional computational methods to process information, ANN is a system based on the operation of biological neural networks. It has the advantage of being able to assign significance to the input parameters and map the inputs to outputs when the relationships between parameters are unknown.

The ANN model was built with a three-layered feedforward network (Figure 5): an input layer, one or more hidden layers, and an output layer. The nodes in each layer were connected by weights, which will be adjusted through the training process to obtain the optimum model. Tan-sigmoid transfer function was used in the hidden layer to give the nonlinear modeling capability. The neural network architecture consists of two or more layers of neurons connected by weights denoted as w_{ji}. Each neuron is used to calculate its output based on the amount of stimulation it receives from the individual input vector x_i

(where x_i is the input of neuron i). Then the net input of a neuron is calculated as the weighted sum of its inputs, and the output of the neuron are used to estimate the magnitude of this net input via the transfer. The net output u_j from a neuron can be expressed as

$$u_j = \sum_{i=1}^{p} w_{ji} x_i, \tag{2.3}$$

Sigmoid function was selected as the transfer function in this study, which is represented in the following equation:

$$\begin{aligned} \varphi(v) &= \frac{1}{1+\exp(-v)}, \\ y_j &= \varphi(u_j), \end{aligned} \tag{2.4}$$

where y_j is the output of the jth neuron in the layers.

The ANN is first to establish a relationship between a set of input variables and a set of output variables from the historical data sets. This is achieved by repeatedly presenting examples of the desired relationship to the network and adjusting the connection weights (i.e., the model coefficients) to reduce the mean-root-square error (RMSE) between the simulated outputs and the observed outputs. The weights of the network continually change until the total error of all the training set is below the acceptable error or other stop mechanism.

Backpropagation is most widely used due to its broad applicability to solve complex nonlinear problems in many domains, such as classification, prediction, and modeling. It works to determine the optimal weights and improve function approximation potential for complex nonlinear data by increasing the number of the hidden layers or the neuron in the hidden layers. Thus the new weights can be calculated by adding a modification to the old weights. The collected data is divided into two sets, one for training and the other for testing.

Determining the size of the hidden layer is a significant task in ANN. Some general rules for selecting the number of hidden nodes N^H in the ANN model suggest that it should be within N^I and $2N^I + 1$ [27], where N^I is the number of input nodes. Moreover, in order to prevent overfitting of the training data, Rogers and Dowla [28] also suggest that the condition $N^H \leqq N^{\mathrm{TR}}/(N^I + 1)$ needs to be satisfied, where N^{TR} is the number of training samples. In this study, a trial-and-error approach was carried out to find the optimum number of hidden nodes in the models. In general, a network structure with less hidden nodes is more preferable; this usually gives better generalization capabilities and fewer overfitting problems. To avoid the overfitting problem, which commonly occurs with the application of ANN, cross-validation tests were used. The selection of the network was performed by considering a minimum value of MSE for the cross-validation data set [29].

In this study, ANN development and simulation were conducted using ANN toolbox of Matlab 2011a (Matwork, NA). Batch gradient decent backpropagation training algorithm was adopted; the training stops when it hits one of the several stopping criteria, including

Table 2: Correlation coefficients between SVIs and water parameters in MSR.

Water parameter	Temp	MLSS	BOD	COD	SS	pH	TP	TN	N-NH_3
SVI	−0.82	−0.18	0.26	0.32	0.21	−0.23	0.18	0.28	0.19

maximum number of iteration, maximum training time, targeted total sum-squared error, and minimum gradient.

2.4. Performance Indicators

The performance of models was evaluated using the following indicators: coefficient of determination (R^2) that provides the variability measure for the data reproduced in the model. Prediction R^2 is a good measure for both comparison and seeing the model's prediction capability. The calculation method is also known as the cross-validation, in which we exclude the first observation, and build the model with the remaining ones, use this model to predict the excluded observation, and repeat for all observations. It is a good measure for out-of-sample accuracy. As this test cannot give the accuracy of the model, other statistical parameters should be reported. Mean absolute error (MAE) and root-mean-square error (RMSE) measure residual errors, providing a global idea of the difference between the observation and modeling. The indicators were defined as follow:

$$R^2 = \frac{\sum_{i=1}^{n}\left(Y_i - \overline{Y_i}\right)^2 - \sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2}{\sum_{i=1}^{n}\left(Y_i - \overline{Y_i}\right)^2},$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|\widehat{Y}_i - Y_i\right|, \tag{2.5}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\widehat{Y}_i - Y_i\right)^2},$$

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

3. Results and Discussion

Correlation between SVIs and water parameters were analyzed to evaluate the influence of each parameter on the bulking level, which provides a measure of linear relationship between SVI and each parameter. The results (Table 2) showed that all the coefficients were greater than 0.15, indicating that all these parameters had high correlation with SVIs and thus included in the models as input variables. It is noted that high correlation coefficient (0.82)

Table 3: Eigenvalue and percentage variance of the 9 principle components for the prediction model.

PC	1	2	3	4	5	6	7	8	9
Eigenvalue	4.3	1.3	1.1	0.9	0.7	0.3	0.2	0.1	0.1
% variance	47.8	14.8	12.5	10.3	7.4	3.6	1.9	1.1	0.6

Table 4: Composition of the principle components for the prediction model.

Variables	Component		
	PC1	PC2	PC3
COD	.938	.006	−.131
BOD	.913	−.055	−.100
SS	.897	−.049	.007
TP	.855	.019	.016
TN	.685	.551	−.110
NH3	.582	.653	−.048
MLSS	−.197	.729	−.004
pH	.114	−.265	.767
Temperature	−.246	.210	.762

was found between SVI and temperature, which was consistent with the observation that bulking of CQWWTP mostly occurs in the springs.

3.1. PCA

The values of KMO for both prediction and forecast models were above the criteria value of 0.6, indicating that the PCA was applicable [13]. PCA demonstrates the relative importance of each standardized variable in the PC calculations.

The PCA for the prediction model was performed using the 9 selected parameters from the result of correlation analysis. Table 3 showed that the first 3 principle components can explain 74.1% variation of the data variation. The scree test suggested only 3 components with the eigenvalues greater than 1 to be retained, in which all the 9 parameters were included. The composition of the 3 principle components are shown in Table 4, in which PC1 represented the component of water characteristics in the influent expressed as a function of COD, BOD, SS, TP, TN, and NH_3-N, PC2 represented the component of activated sludge mixed liquor concentration expressed as a function of MLSS, and the PC3 represented the component of environmental condition expressed as a function of temperature and pH.

3.2. MLR

The MLR results for the prediction model were shown in Table 5. Stepwise approach was adopted. A *t*-test (significance level of 0.05) was applied to calculate the statistically valid parameters. MLR result showed that all PCs were significant. Therefore, the prediction model for phytoplankton abundance can be written as SVI = 468.935 + 0.025(PC1) − 0.007(PC2) − 15.898(PC3).

Table 5: MLR result for prediction model.

Included independent variables	Regression coefficient (B)	Std. Error of B	Std. regression coefficient (β)	t	Sig.
(Constant)	468.935	16.782		27.942	.000
PC1	.025	.008	.097	3.238	.001
PC2	−.007	.003	−.070	−2.392	.017
PC3	−15.898	.603	−.794	−26.381	.000

Table 6: Performance indexes of the PCR and ANN prediction models.

Performance index	Accuracy performance (training set)		Generalization performance (testing set)	
	PCR	ANN	PCR	ANN
R^2	0.689	0.901	0.772	0.907
RMSE	37.360	21.141	31.673	20.258
MAE	28.524	16.375	24.490	15.899

3.3. ANN

To apply the ANN model, several network structures were tested to find the most appropriate topology. Using the 9 water parameters as inputs, the best architecture consisted of a three-layer network. Sigmoid and linear functions were used as activation function in the neurons of the hidden layer and output neuron, respectively. 70% of original data were used for training, among which 10% were randomly selected for cross-validation, and the remaining 30% of data were used for testing. The training was performed for a maximum of 30000 iterations. The detailed results were presented in Figures 6–9, and they are discussed in more detail in the next section.

3.4. Modeling Results Comparison

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 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 to prevent the overfitting issue of the model, both performance checks need to be considered. In the present research, the performance indexes for ANN's models were averaged with 50 runs.

The performance of prediction models were shown in Table 6. Using the PCR model, the performance indexes for the testing step were generally better than those for the training step, with the R^2 of 0.689 for training and 0.772 for testing. Compared to PCR model, the ANN model has the best performance, with R^2 (0.901, 0.907), RMSE (21.141, 20.258), and MAE (16.375, 15.899) for accuracy and generalization performance, indicating that instead of PCR, ANN can handle well the nonlinear relationship between SVIs and water parameters.

It was noted that the ANN model did not need to perform PCA to obtain the good results. The PCA-ANN results obtained with the R^2 of 0.9 (not shown here) cannot improve the prediction powers for testing and training data sets, confirming that ANN is a powerful tool for dealing with collinearity of data.

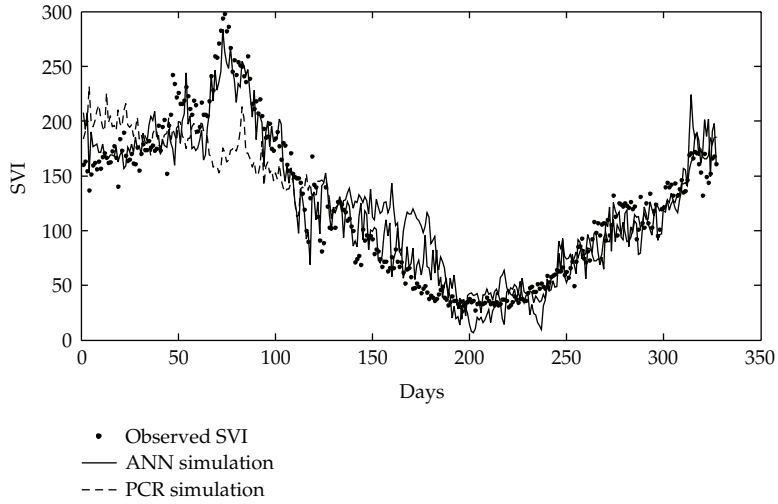

Figure 6: Observed and predicted SVIs for the training data set of the prediction model.

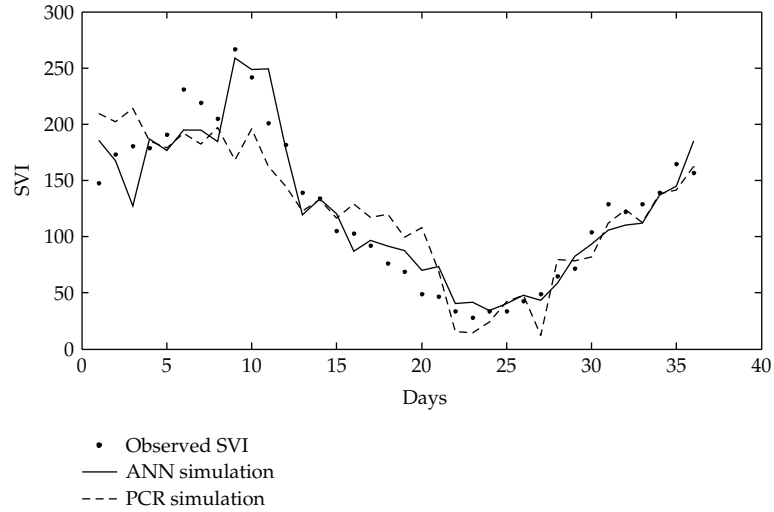

Figure 7: Observed and predicted SVIs for the testing data set of the prediction model.

In the prediction models, no delay was observed for the PCR model in the training set data (Figure 6), but the magnitude is more fluctuate than the ANN models. The prediction of the testing set in Figure 7 for both models exhibit over-estimates in the low SVI level region. In general, ANN was successful to predict the SVIs with a reasonable degree of accuracy for the forecast and the prediction model.

The modeling SVIs versus observed SVIs for PCR and ANN were showed in Figures 8 and 9, respectively. For both training and testing data, both models fitted the measured data well, with the slopes equal to 1 for both fitting curves, that is, the modeling results are equal to the measured data. However, compared to the PCR in which the measured data were distributed more scatter along the fitting curve, ANN models provided better simulation for the measurements, confirming that ANN fits better than PCR when used

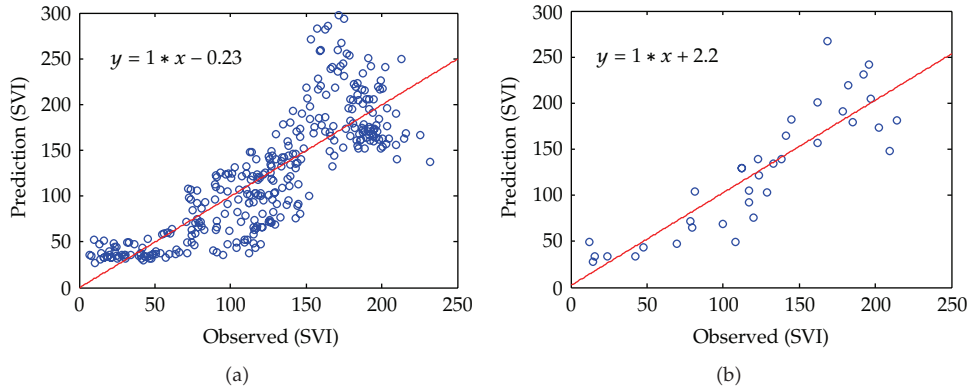

Figure 8: Training (a) and testing (b) results of PCR prediction model.

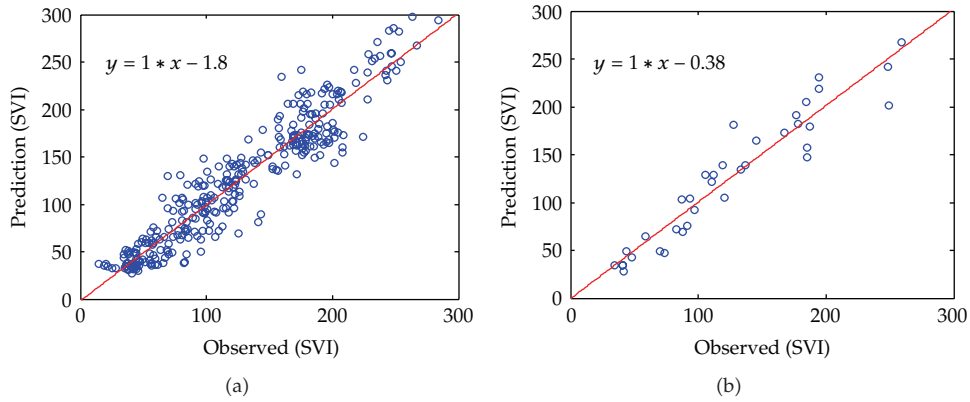

Figure 9: Training (a) and testing (b) results of ANN prediction model.

in predicting the SVIs of CQWWTP. From the modeling point of view, a disadvantage of ANN is that the mechanisms of the inner signal processing are unknown. However, it has provided enough information for CQWWTP to prevent the sludge bulking problems; for controlling the sludge bulking problem occurrence, the engineers can only control the predicted SVI $< 150\,\mathrm{mL/g}$ by adjusting the operational variables, such as MLSS, without understanding the complete mechanisms and the relationships among the variables. ANN was demonstrated to effectively solve the problems where response flexibility and constant tuning of the models are required.

When compared with the SORBF model recently developed by Han and Qiao [21], our ANN models showed similar values of RMSE and R^2 and simpler ANN algorithm, demonstrating that our ANN model is suitable and has more advantages for the SVI prediction. This is highly probably due to the more complete bulking causative variables involved in the ANN model thus providing more information in explaining the sludge bulking phenomena, despite that the complete mechanisms causing bulking and the relationships among variables are still unclear.

In summary, predicting sludge bulking using our ANN model can provide accurate prediction results. The fitting accuracy was found to improve with the increasing number

of bulking causative variables. The model has been tested in the CQWWTP using A/A/O processes, which are different from the traditional aerobic process. Thus, empirical studies will also be conducted in the future for additional data sets to demonstrate that the ANN model is generalizable to extensive data sets under different circumstances.

4. Conclusions

The econometric technique (PCR) and the artificial intelligence technique (ANN) applied in the study are powerful analysis tools that can be used to solve a problem that is poorly understood or difficult to solve with the traditional deterministic relationship. The updated knowledge on sludge bulking is still unclear, and thus the unconventional systematic data-driven modeling approaches could be used to improve the prediction. Prediction models with PCR and ANN were compared for simulating the SVIs in CQWWTP, using nine water parameters including environmental conditions of temperature and pH, wastewater characteristics of BOD, COD, SS, NH_4^+, TN, and TP, and activated sludge concentration of MLSS. PCA result indicated that only 3 PCs with eigenvalues greater than 1 were obtained, which can explain 74.1% variance of data. The application of PCA in the PCR model was considered better than using the original data, as it would eliminate the collinearity problem and reduce the number of inputs, thus decreasing the model complexity.

PCR showed worse prediction performance than ANN, indicating that the complex nonlinear relationship among the variables in the treatment systems cannot not be simulated using linear model alone. Besides, by using PCR, the highest SVI values were underestimated during the training step. On the other hand, ANN had better prediction power with the R^2 of 0.9 for both accuracy performance and generalization performance, implying that ANN is good to deal with the collinearity problem in the data without performing data pretreatment using PCA. Compared with the recently developed SORBF model, ANN model is suitable and has more advantages for the SVI prediction by using simpler ANN algorithm and including more bulking causative variables in the model. The ANN models established by this research project performed well to address the wastewater quality and sludge bulking problem of CQWWTP. The modeling approach described here for analyzing the bulking problem has yielded useful information for effective wastewater treatment management.

Though the ANN presented here is obtained from the CQWWTP, the technique can also be applied for the other WWTPs, as the input parameters and operational conditions are similar. The method can be used for control of wastewater treatment operation in order to improve the treatment performance.

Acknowledgments

The authors thank Mio Cheng (Alice) Chan, the undergraduate student in the Faculty of Science and Technology at the University of Macau, for assistance in performing ANN and the technical staff in CQWWTP for all water parameters measurement. The financial support from the Fundo para o Desenvolvimento das Ciências e da Tecnologia (FDCT) and the Research Committee at University of Macau under Grant no. MYRG106 (Y1-L3)-FST12-LIC is gratefully acknowledged.

References

[1] D. Jenkins, M. G. Richard, and G. T. Digger, *Manual on the Caused and Control of Activated Sludge Bulking, Foaming and other Solids Separation Problems*, Lewis Publishers, New York, NY, USA, 2003.

[2] J. Chudoba, J. Blaha, and V. Madera, "Control of activated sludge filamentous bulking. III. Effect of sludge loading," *Water Research*, vol. 8, no. 4, pp. 231–237, 1974.

[3] J. Chudoba, P. Grau, and V. Ottova, "Control of activated sludge filamentous bulking. II. Selection of microorganisms by means of a selector," *Water Research*, vol. 7, no. 10, pp. 1389–1406, 1973.

[4] J. Chudoba, V. Ottova, and V. Madera, "Control of activated sludge filamentous bulking: I. Effect of the hydraulic regime or degree of mixing in an aeration tank," *Water Research*, vol. 7, no. 8, pp. 1163–1182, 1973.

[5] M. Sezgin, D. Jenkins, and D. S. Parker, "A unified theory of filamentous activated sludge bulking," *Journal of the Water Pollution Control Federation*, vol. 50, no. 2, pp. 362–381, 1978.

[6] A. M. P. Martins, J. J. Heijnen, and M. C. M. Van Loosdrecht, "Effect of feeding pattern and storage on the sludge settleability under aerobic conditions," *Water Research*, vol. 37, no. 11, pp. 2555–2570, 2003.

[7] R. Goel, T. Mino, H. Satoh, and T. Matsuo, "Intracellular storage compounds, oxygen uptake rates and biomass yield with readily and slowly degradable substrates," *Water Science and Technology*, vol. 38, no. 8-9, pp. 85–93, 1998.

[8] M. C. M. Van Loosdrecht, M. A. Pot, and J. J. Heijnen, "Importance of bacterial storage polymers in bioprocesses," *Water Science and Technology*, vol. 35, no. 1, pp. 41–47, 1997.

[9] C. L. In and F. L. De Los Reyes, "Integrating decay, storage, kinetic selection, and filamentous backbone factors in a bacterial competition model," *Water Environment Research*, vol. 77, no. 3, pp. 287–296, 2005.

[10] I. Lou and F. L. De Los Reyes III, "Substrate uptake tests and quantitative FISH show differences in kinetic growth of bulking and non-bulking activated sludge," *Biotechnology and Bioengineering*, vol. 92, no. 6, pp. 729–739, 2005.

[11] A. G. Capodaglio, H. V. Jones, V. Novotny, and X. Feng, "Sludge bulking analysis and forecasting: application of system identification and artificial neural computing technologies," *Water Research*, vol. 25, no. 10, pp. 1217–1224, 1991.

[12] H. R. Maier, A. Jain, G. C. Dandy, and K. P. Sudheer, "Methods used for the development of neural networks for the prediction of water resource variables in river systems: current status and future directions," *Environmental Modelling and Software*, vol. 25, no. 8, pp. 891–909, 2010.

[13] H. Çamdevýren, N. Demýr, A. Kanik, and S. Keskýn, "Use of principal component scores in multiple linear regression models for prediction of Chlorophyll-a in reservoirs," *Ecological Modelling*, vol. 181, no. 4, pp. 581–589, 2005.

[14] S. H. Te and K. Y. H. Gin, "The dynamics of cyanobacteria and microcystin production in a tropical reservoir of Singapore," *Harmful Algae*, vol. 10, no. 3, pp. 319–329, 2011.

[15] J. T. Kuo, Y. Y. Wang, and W. S. Lung, "A hybrid neural-genetic algorithm for reservoir water quality management," *Water Research*, vol. 40, no. 7, pp. 1367–1376, 2006.

[16] F. Recknagel, M. French, P. Harkonen, and K. I. Yabunaka, "Artificial neural network approach for modelling and prediction of algal blooms," *Ecological Modelling*, vol. 96, no. 1–3, pp. 11–28, 1997.

[17] K. I. Yabunaka, M. Hosomi, and A. Murakami, "Novel application of a back-propagation artificial neural network model formulated to predict algal bloom," *Water Science and Technology*, vol. 36, no. 5, pp. 89–97, 1997.

[18] P. P. Zhang, "Time series forecasting using a hybrid ARIMA and neural network model," *Neurocomputing*, vol. 50, pp. 159–175, 2003.

[19] L. Belanche, J. J. Valdés, J. Comas, I. R. Roda, and M. Poch, "Prediction of the bulking phenomenon in wastewater treatment plants," *Artificial Intelligence in Engineering*, vol. 14, no. 4, pp. 307–317, 2000.

[20] M. Côté, B. P. A. Grandjean, P. Lessard, and J. Thibault, "Dynamic modelling of the activated sludge process: improving prediction using neural networks," *Water Research*, vol. 29, no. 4, pp. 995–1004, 1995.

[21] H. G. Han and J. F. Qiao, "Prediction of activated sludge bulking based on a self-organizing RBF neural network," *Journal of Process Control*, vol. 22, no. 6, pp. 1103–1112, 2012.

[22] APHA, AWWA, and WEF, *Standard Methods for the Examination of Water and Wastewater*, American Public Health Association, Washington, DC, USA, 2002.

[23] P. H. Nielsen, C. Kragelund, R. J. Seviour, and J. L. Nielsen, "Identity and ecophysiology of filamentous bacteria in activated sludge," *FEMS Microbiology Reviews*, vol. 33, no. 6, pp. 969–998, 2009.

[24] A. M. P. Martins, K. Pagilla, J. J. Heijnen, and M. C. M. Van Loosdrecht, "Filamentous bulking sludge—a critical review," *Water Research*, vol. 38, no. 4, pp. 793–817, 2004.

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

[26] W. . Zhang, I. Lou, W. K. Ung, Y. Kong, and K. M. Mok, "Eutrophication in Macau main storage reservoir," in *Proceedings of the 12th International Conference on Environmental Science and Technology,*, pp. 1114–1121, Rhodes Island, Greece, September 2011.

[27] R. . Hecht-Nielsen, "Kolmogorov's mapping neural network existence theorem," in *Proceedings of the 1st IEEE International Joint Conference of Neural Networks*, vol. 3, pp. 11–14, New York, NY, USA, 1987.

[28] 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, pp. 457–481, 1994.

[29] S. I. V. Sousa, F. G. Martins, M. C. M. Alvim-Ferraz, and M. C. Pereira, "Multiple linear regression and artificial neural networks based on principal components to predict ozone concentrations," *Environmental Modelling and Software*, vol. 22, no. 1, pp. 97–103, 2007.