A systematic approach to data-driven modeling and soft sensing in a full-scale plant

M. H. Kim*, Y. S. Kim, A. A. Prabu* and C. K. Yoo

ABSTRACT

The well-known mathematical modeling and neural networks (NNs) methods have limitations to incorporate the key process characteristics at the wastewater treatment plants (WWTPs) which are complex, non-stationary, temporal correlation, and nonlinear systems. In this study, a systematic methodology of NNs modeling which can be efficiently included in the key modeling information of the WWTPs is performed by selecting the temporal effect of the hydraulics based on multi-way principal components analysis (MPCA). The proposed method is applied for modeling wastewater quality of a full-scale plant, which is a Daewoo nutrient removal (DNR) process. Through the experimental results in a full-scale plant, the efficiency of the proposed method is evaluated and the prediction capability is highly improved by the inclusion of the hydraulics term due to the optimized structure of neural networks.

Key words | hydraulic characteristics, multiway principal components analysis (MPCA), neural networks (NNs), nonlinear modeling, soft sensing, temporal correlation

M. H. Kim
Y. S. Kim
C. K. Yoo (corresponding author)
College of Environmental and Applied Chemistry,
Center for Environmental Studies/Green Energy Center,
Kyung Hee University,
Seocheon-dong 1,
Gyeonggi-Do 446-701,
South Korea
E-mail: *ckyoo@khu.ac.kr*

A. A. Prabu
Department of Chemistry,
School of Science and Humanities,
Vellore Institute of Technology University,
Vellore 632 014,
India
E-mail: *werver0311@khu.ac.kr*;
bamsu83@hanmail.net;
anandprabu@vit.co.in

INTRODUCTION

In recent years, the increase in environmental restrictions has led to an increase in efforts aimed at attaining higher effluent quality from wastewater treatment plants (WWTPs). In order to meet these demands, the use of advanced monitoring and modeling methods is required. The biological nutrient removal (BNR) process is the most widely used form of wastewater treatment. However, it is a very complicated process for modeling due to the nonlinear, time-varying characteristics of the microorganisms and the variations of incoming wastewater flow and its composition (Lee *et al.* 2002). Existing modeling methods such as mathematical models have been used to estimate the model equation and predict the effluent concentration. However, it is difficult to reflect hydraulic characteristics as time-variable, since WWTPs have large fluctuations and temporal correlation between the time series data (Hack & Kohne 1996; Olsson & Newell 1999; Kim & Yoo 2008). The removal of these temporal correlations can contribute to ensure consistent and long-term performance of the models.

Neural networks (NNs) in recent times have been successfully applied in various fields including environmental engineering, such as a modeling technique for nonlinear systems. NNs are able to learn nonlinear or dynamic behavior exclusively and predict the performance of the WWTPs. The NNs model serves as a black box model and does not require the knowledge of any parameters. It can represent a highly nonlinear process with a complex structure in some instances better than many other empirical models (Hong *et al.* 1997; Demuth *et al.* 2007; Pai 2008). However, NNs have been criticized for a lack of interpretation upon physical relationships and a difficulty of its structure determination.

In this paper, we propose a systematic approach for the determination of NNs structure that uses the multi-way principal component analysis (MPCA) method to interpret the variable relation of input variables, and to select the key

*These authors contributed equally to this work.

doi: 10.2166/wst.2009.346

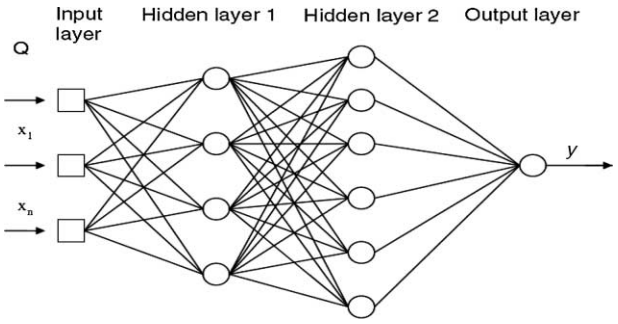

Figure 1 | The structure of multi-layer neural networks.

temporal hydraulic information by its variable importance in the projection (VIP) and soft sensing which can predict the effluent concentration of the treated plant.

METHODS

Neural networks (NNs)

The NN modeling in which the important operational features of the human nervous system are simulated can be applied to solve problems by nonlinear states. To operate analogous to a human brain, many simple computational elements called artificial neurons that are connected by variable weights are used in NNs. Multi-layer feed-forward NNs used this study can be used to solve complicated and nonlinear problems. A typical multi-layer feed-forward NNs model consists of several layers: input, hidden layers 1 and 2, and an output layer. Each layer is comprised of several operating neurons. The number of neurons in the input layer is fixed by the number of input variables. Also, the number of neurons of the output layer is determined by the number of output variables. The hidden layers perform an interface to fully interconnect input and output layers. The number of neurons in the hidden layers depends on the complexity of the correlations to be detected by the networks (Choi & Park 2001; Himmelblau 2008).

Figure 1 shows the structure of multi-layer neural networks. In general, the structure of NNs depends on the number of hidden layer and the number of neurons present in input, hidden and output layers. Each neuron is connected to every neuron in adjacent layers before being introduced as input to the neuron in the next layer by connection weight, which determines the strength of the relationship between two connected neurons. Each neuron sums all the inputs that it receives, and the sum is converted to an output value based on a predefined activation, or transfer function. The sum of x_i $(i = 1,2,\ldots,n)$ multiplied with corresponded weight factor w_i and critical value, b formed the neuron output y through transformed function, f, following Equation (1) (Lee *et al.* 2002; Pai 2008).

$$y = f(\text{net}) \quad \text{net} = \sum_{i=1}^{n} x_i w_i + b \tag{1}$$

where n is the number of neurons of input, hidden and output layers, y is model output, f is transformed function, w_i is weight and b indicates critical value.

The NNs modeling for the correlation of input and output in real systems depend on the structure, transformed function and learning rule. Supervised learning rule is often adopted for training the networks on how to relate input data to output data. The back-propagation learning rule which is used in this study is a generalized delta rule including momentum coefficient to minimize the mean

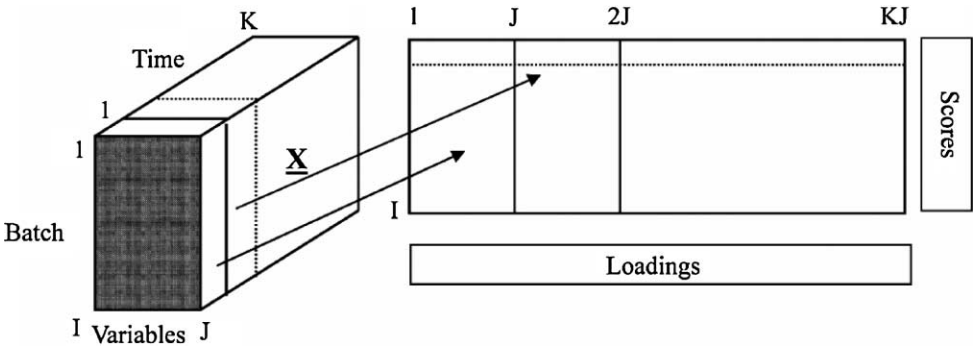

Figure 2 | Unfolding of three dimensional process data.

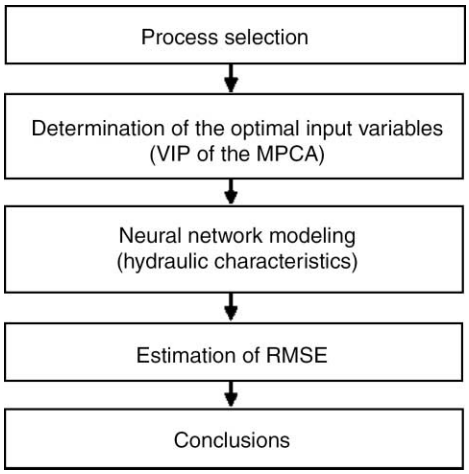

Figure 3 | The systematic procedure for data-driven NNs model and soft sensing.

square difference between the predicted and actual networks outputs. Multi-layer feed-forward NNs with back-propagation learning rule is implemented in MATLAB 7.1. All of the variables were normalized in scope of −1 to 1 using Equation (2) because of different units and long training time.

$$X_i = \frac{(x_i - x_0)}{\delta x}. \quad (2)$$

Multi-way principal component analysis (MPCA)

Multi-way Principal Component Analysis (MPCA) is an extension of Principal Components Analysis (PCA) to handle data in three-dimensional matrix. The three-dimensional matrix can be unfolded in three ways, yielding the two dimensional matrices on which PCA can be performed (Mu 2003; Lee *et al.* 2003; Villez *et al.* 2008). In this study, once it is focused on analyzing the variability among the variables, batch process data can be expressed as a three dimensional matrix $X(I \times J \times K)$, where I is the number of batches, J is the number of variables and K is the number of samples in a given batch. The batch-wise unfolding is used to analyze the data set, which enables simply integrating the off-line data records with the unfolded on-line type data as MacGregor's unfolding approach (Figure 2). To avoid the problems created by nonlinearities in the data, the major nonlinear behavior of the process is eliminated from the unfolded matrix by removing the mean trajectory of each variable. This is achieved by subtracting off the mean of each column from the corresponding column in the unfolded matrix. Once the matrix is mean-centered and scaled, PCA is performed. From the results from PCA, the loading vectors and the scores for each batch were calculated. The loading vectors provide a weighting for each variable at each time. Based on the loading vectors, MPCA compresses the normal batch data and extracts information by projecting the data onto a low-dimensional space that summarizes both the variables and their time histories (Nomikos & Macgregor 1994; Lee *et al.* 2003).

Optimal variable selection method of NNs model by MPCA Information

There is now is a great need for methods capable of selecting a parsimony structure of NNs model for predicting the output variables. However, the biological process data are characterized by irrelevant features as well as collinear and multivariate characteristics. To solve the aforementioned problems, the first step in creating such a modeling method is to extract the fundamental features (or inputs) of the process data set and the second step is to make a model for the data. Here, the information of MPCA can be used

Table 1 | Input variables of the NNs model

Symbol	Description	Symbol	Description	Symbol	Description	Symbol	Description
X_1	Flow rate	X_7	$Flowrate_{t-1}$	X_{13}	$Flowrate_{t-2}$	X_{19}	$CODe_{t-1}$
X_2	TSS	X_8	TSS_{t-1}	X_{14}	TSS_{t-2}	X_{20}	TNe_{t-1}
X_3	BOD	X_9	BOD_{t-1}	X_{15}	BOD_{t-2}	X_{21}	TPe_{t-1}
X_4	COD	X_{10}	COD_{t-1}	X_{16}	COD_{t-2}	X_{22}	$CODe_{t-2}$
X_5	TN	X_{11}	TN_{t-1}	X_{17}	TN_{t-2}	X_{23}	TNe_{t-2}
X_6	TP_t	X_{12}	TP_{t-1}	X_{18}	TP_{t-2}	X_{24}	TPe_{t-2}

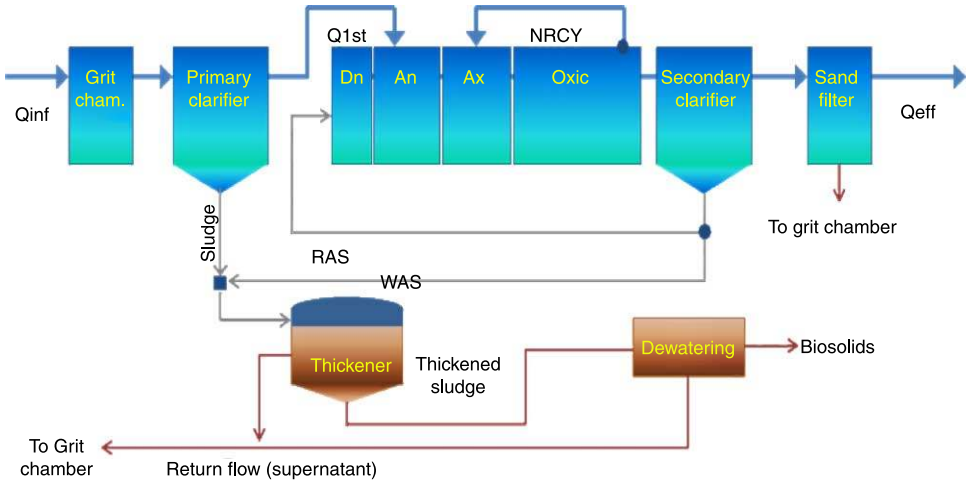

Figure 4 | The plant layout with DNR process.

for the variable selection of NNs model. Note that after the MPCA weight vectors are computed, input variables are selected via the variable importance in the projection (VIP) of MPCA, which is defined as follows:

$$\mathrm{VIP} = \sum_{a} (w_{ak})^2. \tag{3}$$

The VIP is calculated from the weight vector of the MPCA model and the percentage that is explained by the dimension of the model (Nguyen & Rocke (2002). The VIP can be considered as a measure of how much a certain input corresponds to the samples. Thus, important inputs based on the VIP value can be selected. The VIP is the sum over all model dimensions of the contributions. The VIP is a good measure of the influence of all input variables in the model on the response variables, such as COD, TN and TP.

Systematic method for data-driven NNs model and soft sensing

Figure 3 shows the systematic procedure for data-driven NNs model and soft sensing. First, the process variables which are suitable for the model purpose are selected. Second, optimal input variables in a plant are determined by using the VIP of the MPCA. To consider hydraulic characteristics in WWTPs, input variables for the NNs model with the influent and effluent variables of one and two days ago are pooled to the data set in Table 1. The VIP value of the MPCA is used for selecting the key input variables for the NNs model. Here, the normalized data are used. Third, the optimal structure of the NNs is found. Finally, the prediction results can be compared using the root mean square error (RMSE) criteria by following Equation (4) between predicted and actual data.

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (X_i - Y_i)^2}{n-1}}. \tag{4}$$

RESULTS AND DISCUSSION

Process data were collected from a biological wastewater treatment plant which is operated with Daewoo nutrient removal (DNR) process. This plant is used an advanced biological treatment process that has four basins of denitrification, anaerobic, anoxic, oxic processes and a secondary clarifier. The plant layout is shown in Figure 4. Initially, six input variables of the flow rate, total suspended solid (TSS), biological oxygen demand (BOD), chemical oxygen demand (COD), total nitrogen (TN) and total phosphorous (TP) of the influent are used to model three output variables which consists of COD, TN and TP of the effluent. To solve the missing and outlier sample problems, pretreatment of data is performed using 3-sigma technique and interpolation. The data contained daily average values measured between 9 Mar. 2007 and 29 Feb. 2008. The 70% of total data are used for training the NNs model and the remaining data of 30% are used in a test data for the validation.

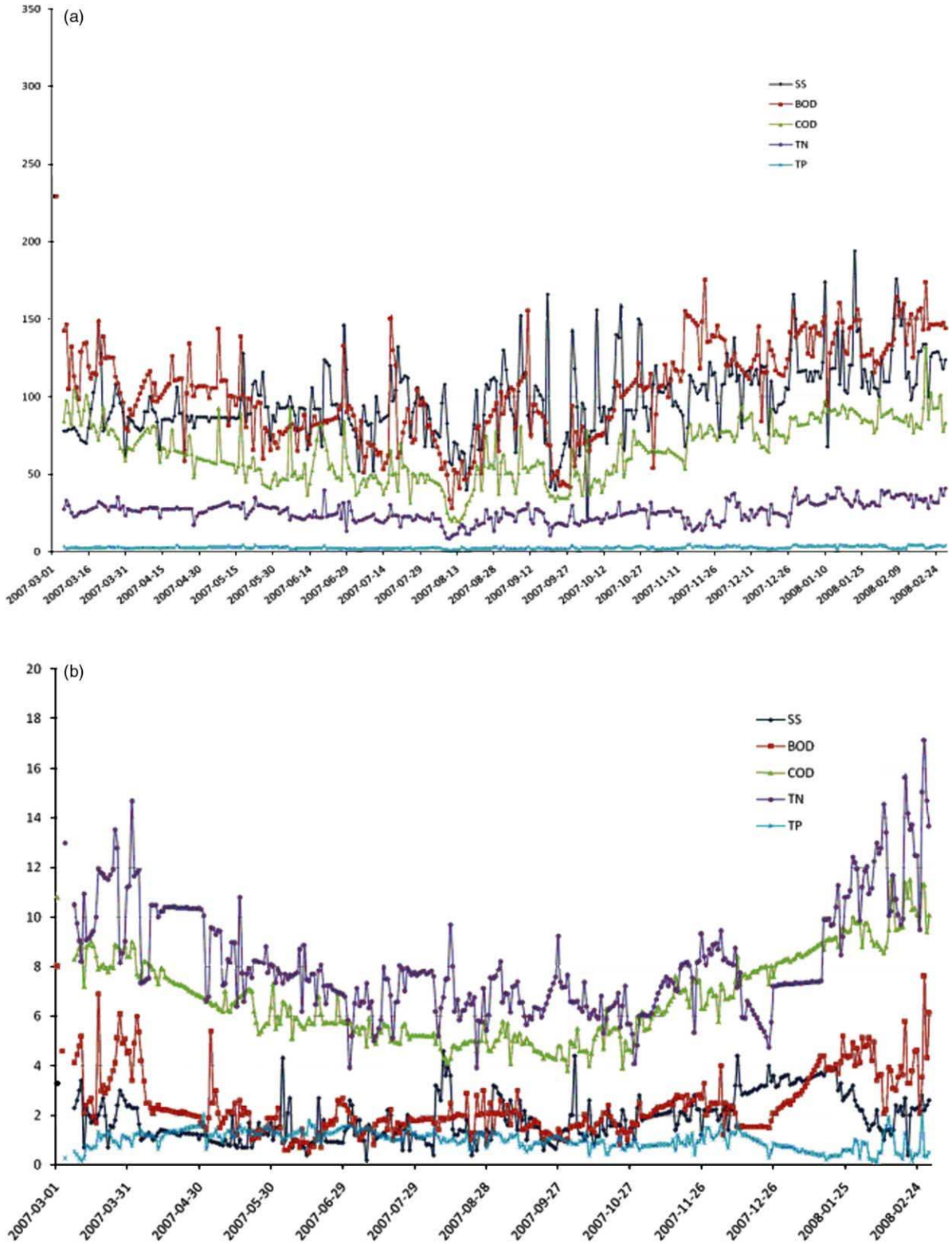

Figure 5 | Time series plot of influents and effluents of H-WWTP, (a) Influents (b) Effluents. Subscribers to the online version of *Water Science and Technology* can access the colour version of this figure from http://www.iwaponline.com/wst

Figure 5 shows the time series plot of influents and effluents for a year. TN and TP concentrations of influents shows usually regular patterns for a year and the influent and effluent of TSS, BOD and COD have relatively a lower concentration profile in the summer and a higher concentration profile in the winter. To consider hydraulic characteristics in WWTPs, influent and effluent data of one and two days ago are included in the input variables. Therefore, 24 input variables and 3 output variables are used to model.

MPCA is performed on three-dimensional matrix of the variables ($I = x_1, x_2, \ldots, x_{24}$), the output ($J = y_1, y_2, y_3$) and the samples ($K = 1, 2, \ldots, 356$). Figure 6 shows the results of MPCA with the score plot of the first two principal components and the VIP plot which is determined by the importance of input variables. The VIP can be considered as a measure of how much a certain influent variable corresponds to the samples. Thus, we can select important influent variables based on the VIP value. Here, seven influent variables (x_8, x_{14}, x_9, x_{15}, x_4, x_6 and x_{23}) are selected for input variables for NNs, which is shown in Figure 6(a). The 7 selected variables are revealed to be sensitive to parameters which are influential with effluents than other variables. Also, the other variables can possibly to exert a negative influence with effluent. Therefore, the VIP of MPCA is important to predict better efficiency though select

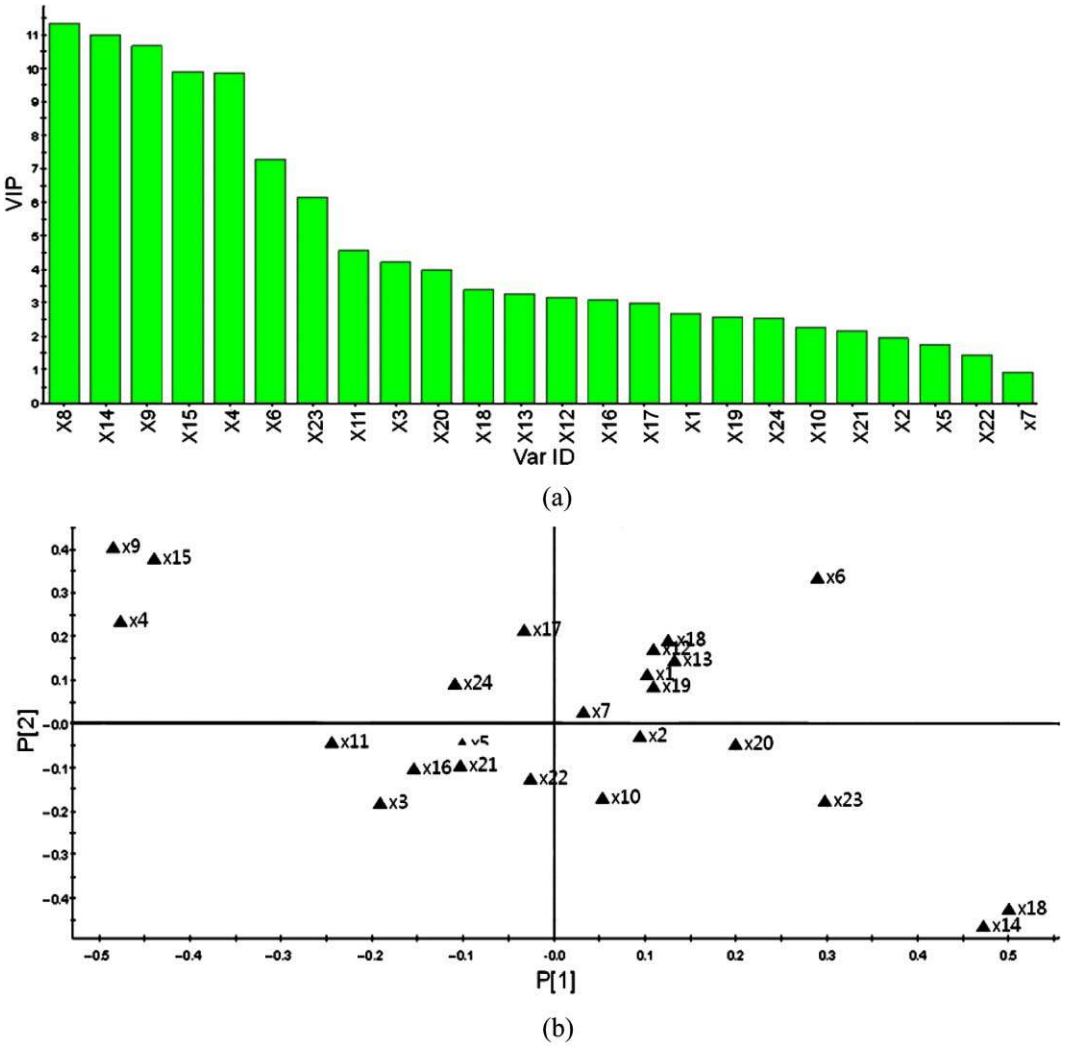

Figure 6 | MPCA results with (a) VIP plot, (b) score plot of the first two principal components.

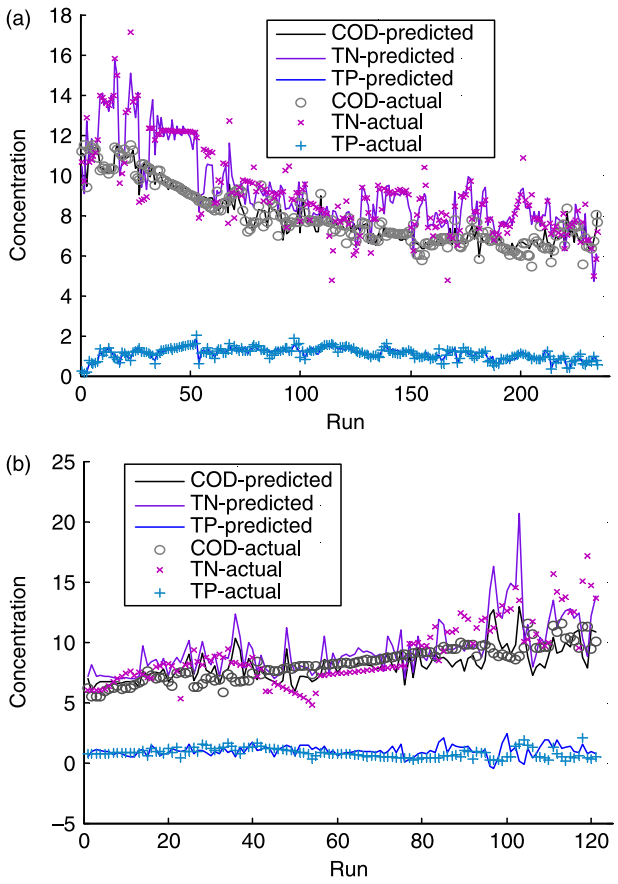

Figure 7 | The NNs results for (a) training data and (b) test data. Subscribers to the online version of *Water Science and Technology* can access the colour version of this figure from http://www.iwaponline.com/wst

key parameters. It can be seen that the importance variables (x_4, x_9, $x_{15}/x_{8,\ 14}$) are separated small groups by Figure 6(b). It means that these variables have strong correlation between one another.

NNs consist of 7 variables in input layer, two hidden layers and 3 variables in output layer. 60% of data is used for training data and 40% is used for test data. Figure 7(a) is the predicted results of training data, and Figure 7(b) shows the results of test data. Training data is possible to predict approximately because training data indicates prediction about the used data during the learning phase of the NNs model. For test data, the prediction results are a little worse than the training data but confirmed trend can be possible. Table 2 compares the prediction performances of general NNs model and the proposed method. When the general NNs model is used, RMSE values for training data and test data is 0.484 and 1.854 on COD, 0.768 and 2.581 on TN, 0.081 and 0.530 on TP, respectively. While using the proposed method, RMSE values of COD, TN and TP of the training data are 0.469, 0.843 and 0.126, respectively. Also test data have RMSE values for COD 1.337, TN 2.096 and TP 0.517, respectively. The proposed model shows a more accurate prediction capability and has a lower RMSE than the simple NNs model because it has already incorporated the key process information of the hydraulics by the VIP of the MPCA model. In the proposed method, the co-linearity problem between the original variables and temporal correlation is eliminated. That is, there are many variables and the co-linearity between the different variables increases the inverse problem of the NNs model. Overall, the predicted results of the NNs model using a systematic approach gives good modeling performance and higher interpretability than any other data-driven modeling method.

Table 2 | The RMSE values of general NNs model and the proposed method

	RMSE (Training data)			RMSE (Test data)		
	COD	TN	TP	COD	TN	TP
General NNs model	0.484	0.768	0.081	1.854	2.581	0.530
Proposed method	0.469	0.843	0.126	1.337	2.096	0.517

CONCLUSIONS

A systematic methodology for determining the structure of neural networks by combining MPCA on hydraulic characteristics and soft sensing for predicting the effluent concentrations are proposed and evaluated in a full-scale waste water treatment plant. To raise the efficiency of prediction, some additional input variables (influent and effluent historical variables) of the NNs model are added to the NNs input variables which were collected on two continuous days. The best structure of NNs model with the plant hydraulics information is formulated and validated in a plant. For the parsimony model of NNs, the VIP of the MPCA is used to select the key input variables for the optimal structure of the NNs model. When more hydraulic characteristics are considered for the model structure, the prediction model shows better results than the general NNs model. However, the proposed method is not able to predict effluents exactly in real full-scale WWTPs because; there are

much of variations, errors and irregularity in actual data. Since the influents information is not enough to predict the model, the other information such as temperature which represents the growth of microorganisms in biological WWTPs can be used. Also, advanced NNs method such as recurrent neural networks (RNNs) which are involved the previous states as well as current states can be used to predict higher efficiency. Moreover, our ongoing research is focused on inspecting the data collection and using RNNs in full-scale WWTPs. If some improved methods are developed, it can be applied to other WWTPs including engaged in activated sludge processes and advanced biological nutrient removal processes.

ACKNOWLEDGEMENTS

This work was supported by the second phase of the Brain Korea 21 project, the Korea Research Foundation by Grant funded by the Korean Government (MOEHRD) (KRF-2007-331-D00089) and funded by Seoul Development Institute (CS070160).

REFERENCES

Choi, D. J. & Park, H. K. 2001 Estimation of a wastewater component using a hybrid artificial neural network in a wastewater treatment process. *J. Korean Soc. Water Qual.* **17**(1), 87–98.

Demuth, H., Beale, M. & Kagan, M. 2007 *Neural Network Toolbox 5–User's Guide*. Mathworks, USA.

Hack, M. & Kohne, M. 1996 Estimation of wastewater process parameters using neural networks. *Water Sci. Technol.* **33**(1), 101–115.

Himmelblau, D. M. 2008 Accounts of experience in the application of artificial neural networks in chemical engineering. *Ind. Eng. Chem. Res.* **47**, 5782–5796.

Hong, Z., Hao, O. J., McAvoy, T. J. & Chang, C. H. 1997 Modeling nutrient dynamics in sequencing batch reactor. *J. Environ. Eng.* **123**(4), 311–319.

Kim, M. H. & Yoo, C. K. 2008 Design and environmental/economic performance evaluation of wastewater treatment plants using modeling methodology. *Korean Chem. Eng. Res.* **46**(3), 610–618.

Lee, D. S., Jeon, C. O., Park, J. M. & Chang, K. S. 2002 Hybrid neural network modeling of a full-scale industrial wastewater treatment process. *Biotechnol. Bioeng.* **78**(6), 671–682.

Lee, J. M., Yoo, C. K. & Lee, I. B. 2003 On-line batch process monitoring using a consecutively updated multiway principal component analysis model. *Comput. Chem. Eng.* **27**, 1903–1912.

Mu, F. 2003 *Multivariate statistical process monitoring and its integration with HAZOP analysis for abnormal event management*. PhD Thesis, Purdue University, USA.

Nguyen, D. V. & Rocke, D. M. 2002 Multi-class cancer classification via partial least squares with gene expression profiles. *Bioinformatics* **18**(9), 1216–1226.

Nomikos, P. & Macgregor, J. F. 1994 Monitoring batch processes using multiway principal component analysis. *AIChE J.* **40**(8), 1361–1375.

Olsson, G. & Newell, B. 1999 *Wastewater Treatment System–Modelling, Diagnosis and Control*. IWA, UK.

Pai, T. Y. 2008 Gray and neural network prediction of effluent from the wastewater treatment plant of industrial park using influent quality. *Environ. Eng. Sci.* **25**(5), 757–766.

Pai, T. Y., Chuang, S. H., Ho, H. H., Yu, L. F., Su, H. C. & Hu, H. C. 2008 Predicting performance of grey and neural network in industrial effluent using online monitoring parameters. *Process Biochem.* **32**, 199–205.

Villez, K., Rosén, C., Anctil, F., Duchesne, C. & Vanrolleghem, P. A. 2008 Combining multiway principal component analysis (MPCA) and clustering for efficient data mining of historical data sets of SBR processes. *Water Sci. Technol.* **57**(10), 1659–1666.