Korean J. Chem. Eng., **21**(6), 1087-1097 (2004)

Application of Fuzzy Partial Least Squares (FPLS) Modeling Nonlinear Biological Processes

Chang Kyoo Yoo†, Yoon Ho Bang, In-Beum Lee*, Peter A. Vanrolleghem and Christian Rosén*****

BIOMATH: Department of Applied Mathematics, Biometrics and Process Control,
Ghent University, Coupure Links 653, B-9000 Gent, Belgium
*School of Environmental Science and Engineering/Department of Chemical Engineering,
Pohang University of Science and Technology, San 31 Hyoja Dong, Pohang 790-784, Korea
**LG Environmental Strategy Insitute (LGESI), Yonsei Univ., 134 Shinchon-dong, Seoul 120-749, Korea
***IEA: Department of Industrial Electrical Engineering and Automation,
Lund University, LTH, Box 118, SE-221 00, Lund, Sweden
(*Received 23 May 2003 • accepted 10 July 2004*)

Abstract−We applied a nonlinear fuzzy partial least squares (FPLS) algorithm for modeling a biological wastewater treatment plant. FPLS embeds the Takagi-Sugeno-Kang (TSK) fuzzy model into the regression framework of the partial least squares (PLS) method, in which FPLS utilizes a TSK fuzzy model for nonlinear characteristics of the PLS inner regression. Using this approach, the interpretability of the TSK fuzzy model overcomes some of the handicaps of previous nonlinear PLS (NLPLS) algorithms. As a result, the FPLS model gives a more favorable modeling environment in which the knowledge of experts can be easily applied. Results from applications show that FPLS has the ability to model the nonlinear process and multiple operating conditions and is able to identify various operating regions in a simulation benchmark of biological process as well as in a full-scale wastewater treatment process. The result shows that it has the ability to model the nonlinear process and handle multiple operating conditions and is able to predict the key components of nonlinear biological processes.

Key words: Fuzzy Partial Least Squares (FPLS), Multivariate Statistical Analysis, Nonlinear Modeling, Nonlinear PLS (NLPLS), Partial Least Squares (PLS), Wastewater Treatment Process (WWTP)

INTRODUCTION

Due to increasing environmental constraints and the necessity of reliable wastewater treatment, efficient modeling and monitoring methods are becoming more and more important. Reliable modeling and monitoring techniques of biological wastewater treatment plants (WWTP) are necessary to maintain the system performance as close as possible to optimal conditions. An adequate model enhances the understanding of the biological processes and it can be a basis for better process design, control, and operation. Also, process monitoring and early fault detection methods are efficient to execute corrective actions well before a dangerous situation occurs in biological processes.

The underlying point is that improving process monitoring and control necessarily means ensuring better knowledge of the process: which variables characterize the process, what are their internal interactions and what degree of confidence can be attributed to the measurements? All these questions are concerned with the characterization of a process, which involves several fundamental stages: the description of the process, the listing of the variables characterizing the process, the establishment of models between the variables, the identification of parameters which intervene in these models, the simplification of models to make them compatible with real-time use and the validation of models. It is generally recognized that, depending on the complexity of the process, two approaches can be adopted to tackle this modeling problem. The first is based on the description of the physical phenomena which enables a mechanistic or first principles model. The second uses only statistical processing of data to obtain 'black-box' type models, which do not take into account the nature and intensity of the physical interactions between the variables. The 'best choice' often seems to be a trade-off between these two viewpoints, leading to a 'grey-box' model which uses simplified hypotheses on the fundamental equations of physics, for example, in the form of matter balances and energy balances, statistics and data processing tools [Ragot et al., 2001; Yoo et al., 2001].

To date, the most successful model and the industrial standard in biological WWTP has been the deterministic mechanistic model, Activated Sludge Model no. 1 or ASM1 [Henze et al., 1987]. It has proven to be an effective model for carbonaceous and nitrogenous substrate removal processes in WWTPs. However, because the ASM model is high-dimensional and contains a large number of kinetic and stoichiometric parameters, which should be determined by using information on specific plant data and process operation, it is not omnipotent in every situation of model application. As a result, the general application of such a complex model to, for instance, process control and the development of operational strategies has been limited [Yoo et al., 2001, 2002].

Today, empirical data-based modeling is a widely used alternative to mechanistic modeling since it requires less specific knowledge of the process being studied compared to a first principles model. Empirical modeling techniques require data (measurements) which are collected on those variables believed to be representative of the

†To whom correspondence should be addressed.
E-mail: ChangKyoo.Yoo@biomath.ugent.be or ckyoo@postech.edu

process behavior and of the properties of the product or system output. Statistical regression techniques and neural networks are now routinely used in the process industries for building empirical models. Statistical regression techniques, based upon least squares methodology, have been used extensively for developing linear empirical models for prediction from historical data. However, it is well known that when dealing with highly correlated multivariate problems, the traditional least-squares approach can lead to singular solutions or imprecise parameter estimation. Limitations due to measurement noise, correlated variables, unknown variable and noise distribution, and data set dimensionality can be overcome by applying multivariate statistical projection based regression techniques such as principal component regression (PCR) and projection to latent structure (PLS). These two techniques provide a solution to both the dimensionality and the correlation problems and can also perform filtering of measurement noise. Projection-based techniques can handle highly correlated, noise-corrupted data sets since they are based upon the assumption of dependency (correlation) between the variables and, consequently, provide the capability to estimate the main underlying structure in terms of a number of latent variables which are linear combination of the original variables [Wold et al., 1989; Baffi et al., 1999].

However, many chemical and biological processes display a nonlinear behavior, which cannot be reliably modeled by means of linear regression techniques. A number of methodologies have been proposed to integrate non-linear features within the linear PLS framework. In particular, when linear PLS is applied to non-linear problems, the minor latent variables cannot always be discarded since they may not only describe noise or negligible variancecovariance structures in the data, they may encapsulate significant information about the non-linear nature of the problem. In fact, non-linear structures may be modeled by using a combination of higher-order and lower-order latent variables calculated from linear PLS [Wold et al., 1989; Qin and McAvoy, 1992; Baffi et al., 1999; Liu et al., 2000; Bang et al., 2003].

Biological treatment plants have different behavior patterns depending on the influent loads, temperature and the activity of microorganisms. The models used for the various operating conditions must generally be different. The challenge is, however, to build a single model framework for all conditions. One solution consists of representing the process by a suite of several models, each one being valid only in a specific operating domain. Another way of representing the process model consists of using a single structure resulting from the aggregation of several sub-models such as fuzzy modeling. Weighting functions are used to reflect the domains of influence of each model [Yen et al., 1998; Tay and Zhang, 1999].

In recent years, Bang et al. [2003] suggested a novel nonlinear fuzzy partial least squares (FPLS) modeling method, which integrated multiple fuzzy modeling capability for aggregation of several sub-models within the linear PLS framework while retaining the orthogonal properties of the linear methodology and keeping a good visualization capability. In this paper, we applied a fuzzy partial least squares (FPLS) for modeling of a biological process with nonlinear features. The outline of this paper is as follows. First, we briefly present PLS and TSK fuzzy modeling. Second, we introduce a nonlinear FPLS modeling and prediction method. Third, the FPLS method is applied to predict the important output variables in a simulation benchmark of biological process and a full-scale wastewater treatment process and the results are discussed. Finally, the conclusion of this work is given.

METHODS

1. PLS Modeling Method

The PLS method is a multivariable linear regression algorithm that can handle correlated inputs and limited data. The algorithm reduces the dimension of the predictor variables (input matrix, **X**) and response variables (output matrix, **Y**) by projecting them to the directions (input weight **w** and output weight **c**) that maximize the covariance between input and output variables. This projection decomposes variables of high collinearity into one-dimensional variables (input score vector **t** and output score vector **u**). The decomposition of **X** and **Y** by score vectors is formulated as follows:

$$\mathbf{X}=\sum_{h=1}^{m}\mathbf{t}_h\mathbf{p}_h^T+\mathbf{E} \tag{1}$$

$$\mathbf{Y}=\sum_{h=1}^{m}\mathbf{u}_h\mathbf{q}_h^T+\mathbf{F} \tag{2}$$

where **p** and **q** are loading vectors, and **E** and **F** are residuals.

2. TSK Fuzzy Modeling

The fuzzy inference system proposed by Takagi, Sugeno and Kang, known as the TSK model, provides a powerful tool for modeling complex nonlinear systems [Yen et al., 1998]. Typically, a TSK model consists of IF-THEN rules of the form

$$\begin{aligned}&R_i\text{: if } x_1 \text{ is } A_{i1} \text{ and } \cdots \text{ and } x_r \text{ is } A_{ir} \text{ then } y_i\\ &\quad =b_{i0}+b_{i1}x_1+\cdots+b_{ir}x_r \quad \text{for} \quad i=1, 2, \cdots, L\end{aligned} \tag{3}$$

where L is the number of rules, $\mathbf{x}_i=[x_1\ x_2\ \cdots\ x_r]^T$ are input variables, y_i are local output variables, A_{ij} are the fuzzy sets that are characterized by the membership function $A_{ij}(x_j)$, and $\mathbf{b}_i=[b_{i0}\ b_{i1}\cdots\ b_{ir}]^T$ are real-valued parameters. The overall output of the model is computed by

$$y=\frac{\sum_{i=1}^{L}\tau_i y_i}{\sum_{i=1}^{L}\tau_i}=\frac{\sum_{i=1}^{L}\tau_i(b_{i0}+b_{i1}x_1+\cdots+b_{ir}x_r)}{\sum_{i=1}^{L}\tau_i} \tag{4}$$

where τ_i is the firing strength of rule R_i, which is defined as

$$\tau_i=A_{i1}(x_1)\times A_{i2}(x_2)\times\cdots\times A_{ir}(x_r) \tag{5}$$

Fig. 1 shows a schematic block diagram of the TSK fuzzy model.

In general, Gaussian-type membership functions are used to build the model. They are defined by

$$A_{ir}(x_r)=\exp\left(-\frac{(x_r-c_{ir})^2}{2\sigma_i^2}\right),\ i=1, 2, \cdots, L \tag{6}$$

where c_{ir} is the center of the *i*th Gaussian membership function of the *r*th input variable x_r and σ_i is the standard deviation of the membership function.

The great advantage of the TSK fuzzy model is its representative power, which stems from its ability to describe complex nonlinear systems using a small number of rules. Moreover, the output of the model has an explicit functional form (Eq. (4)), and the individual rules give insights into the local behavior of the model.

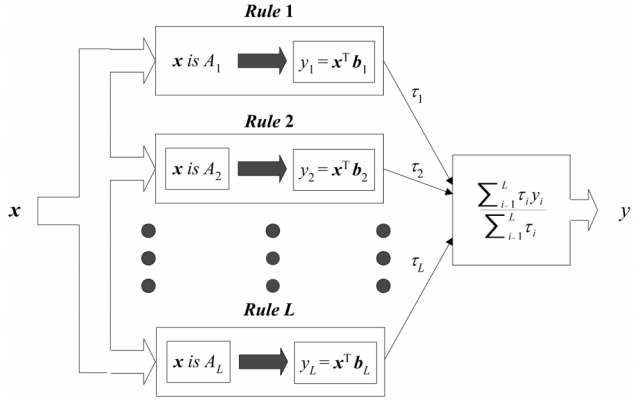

Fig. 1. Block diagram of the TSK fuzzy model.

The good interpretability of the fuzzy system may match the utility of the PLS method in intuitive data analysis [Yen et al., 1998].

3. Nonlinear Fuzzy Partial Least Squares (FPLS) Modeling

As described previously, biological treatment plants have different and nonlinear dynamics depending on, for instance, the influent loads, temperature and the activity of microorganisms, which may call for multiple (nonlinear) models for the various operating conditions. In this case, fuzzy modeling, which aggregates several sub-models and integrates weighting functions that reflect the domains of influence of each model, could be an alternative.

Bang et al. [2003] proposed a fuzzy partial least squares FPLS modeling which applies the TSK fuzzy model to the PLS inner regression. The FPLS method is basically a combination of the PLS method and the TSK fuzzy model. The PLS outer projection is used as a dimension reduction tool to remove collinearity, and the TSK fuzzy inner model is used to capture the nonlinearity in the projected latent space. An advantage of using the TSK fuzzy model as the inner regressor is its interpretability, which facilitates the design of the FPLS model structure by allowing human experts to participate in the design process.

4. FPLS Algorithm

Fig. 2 shows a schematic of the basic FPLS method, which uses the PLS outer transform to generate score variables from the data. Score vectors ($\mathbf{t}_h$ and $\mathbf{u}_h$) of the same factor h are used to train the inner TSK fuzzy model $f_h(\cdot)$, which obeys the following relation

$$u_h = f_h(t_h) + e_h \tag{7}$$

where $\mathbf{e}_h$ represents the regression error. The parameters of $f_h(\cdot)$ should

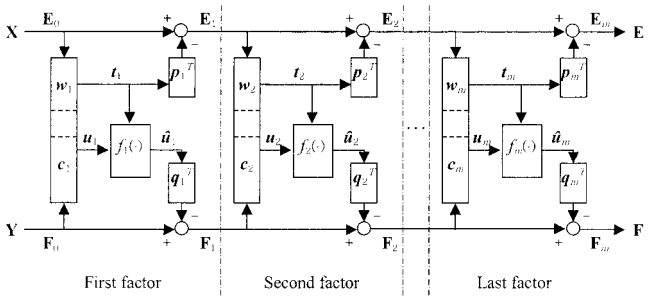

Fig. 2. Block diagram of the FPLS method.

be selected to minimize $\mathbf{e}_h$ without over-fitting. To summarize, by not updating the outer relation FPLS keeps the linear PLS property that variables are projected into the directions maximizing the covariance, and it captures nonlinearity through the nonlinear modeling capacity of the TSK model [Bang et al., 2003].

The FPLS algorithm can be formulated as follows.

1. Scale **X** and **Y** to have zero-mean and unit-variance.

Let $\mathbf{E}_0 = \mathbf{X}$, $\mathbf{F}_0 = \mathbf{Y}$ and h=1.

2. For each factor h, take $\mathbf{u}_h$ from one of the columns of $\mathbf{F}_{h-1}$.

3. PLS outer transform:

$$\mathbf{w}_h^T = \mathbf{u}_h^T \mathbf{E}_{h-1} / (\mathbf{u}_h^T \mathbf{u}_h) \tag{8}$$

$$\mathbf{w}_h = \mathbf{w}_h / \|\mathbf{w}_h\| \tag{9}$$

$$\mathbf{t}_h = \mathbf{E}_{h-1} \mathbf{w}_h \tag{10}$$

$$\mathbf{c}_h^T = \mathbf{t}_h^T \mathbf{F}_{h-1} / (\mathbf{t}_h^T \mathbf{t}_h) \tag{11}$$

$$\mathbf{c}_h = \mathbf{c}_h / \|\mathbf{c}_h\| \tag{12}$$

$$\mathbf{u}_h = \mathbf{F}_{h-1} \mathbf{c}_h \tag{13}$$

Iterate this step until it converges. This step is called the nonlinear iterative partial least squares (NIPALS) algorithm. Although there exists a faster and more stable algorithm using eigenvectors, we use NIPALS to give readers a clearer picture of PLS outer projection.

4. Find the TSK fuzzy-type inner relation function, $f_h(\cdot)$, which predicts the output score u_h with the input score t_h. $f_h(\cdot)$, has the functional form

$$f_h(t) = \sum_{i=1}^{L} G_i (b_{i0} + b_{i1} t) \tag{14}$$

where

$$G_i = \frac{\tau_i}{\sum_{i=1}^{L} \tau_i} \tag{15}$$

$$\tau_i(t) = \exp\left(-\frac{(t - c_i)^2}{2\sigma_i^2} \right), \; i = 1, 2, \cdots, L \tag{16}$$

G_i is the normalized firing strength and τ_i is a Gaussian-type firing strength for the ith rule. First, the number of fuzzy rules, L, should be estimated by the model designer at an integer value that minimizes the regression error of $f_h(\cdot)$ without creating an over-fitted model. The designer may use information gained from the score plot or some numerical criteria such as the sum of squared errors (SSE) for cross validation. The designer can then decide the other parameters, such as c_i, σ_i and $\mathbf{b}_i$, using a numerical curve fitting function to minimize the SSE.

5. Calculate the **X** and **Y** Loadings

$$\mathbf{p}_h^T = \mathbf{t}_h^T \mathbf{E}_{h-1} / (\mathbf{t}_h^T \mathbf{t}_h) \tag{17}$$

$$\mathbf{q}_h^T = \hat{\mathbf{u}}_h^T \mathbf{F}_{h-1} / (\hat{\mathbf{u}}_h^T \hat{\mathbf{u}}_h) \tag{18}$$

where $\hat{\mathbf{u}}_h = \mathbf{f}_h(\mathbf{t}_h) = [f_h(t_h(1)), f_h(t_h(2)), \cdots, f_h(t_h(N))]^T$ for N samples.

6. Calculate the residuals for factor h.

$$\mathbf{E}_h = \mathbf{E}_{h-1} - \mathbf{t}_h \mathbf{p}_h^T \tag{19}$$

$$\mathbf{F}_h = \mathbf{F}_{h-1} - \hat{\mathbf{u}}_h \mathbf{q}_h^T \tag{20}$$

7. Let h=h+1, then return to step 2 until all m principal factors are calculated. The number of factors m is decided by the designer. The designer may use information gained from the score plot or some numerical criteria such as SSE for cross validation.

The parameters of the fuzzy-type inner relation function $f_h(\cdot)$ (i.e., the values of c_i, σ_i and $\mathbf{b}_i$) can be decided by various heuristic rules. In this work, the c_i, σ_i and $\mathbf{b}_i$ values are determined by using the fuzzy c-means (FCM) algorithm [Jang et al., 1997], the nearest neighbor heuristic rule suggested by Moody and Darken [1989] and a global learning procedure (see the appendix for the mathematical formulations of these methods). Then a numerical nonlinear least squares curve fitting function is applied for the optimization of the parameters with respect to minimizing the SSE. However, if the optimized model shows signs of over-fitting, such as very steep changes in its trend, the designer can change and fix some parameters and then optimize the other parameters to make a smoother and more reliable model within the criteria of his or her expertise.

As is shown in the algorithm, the designer's decisions are emphasized in the calibration of an FPLS model. This aspect of FPLS represents an improvement over other PLS algorithms. Generally, structural parameters such as L and m are selected by using a cross validation method to avoid the problem of over-fitting. Cross validation is often a must for high dimensional models, because the model shape cannot be well presented in visible form. Although the fuzzy modeling process gives particular weight to the application of the expert's knowledge in the modeling process, it is also hindered by the problem of high dimensionality. Regardless of the type of modeling, designers should check the validity of their model. The FPLS method aids designers in model validation by providing a simple modeling interface for visual checking, in addition to the typical cross validation method. The visual check comprises checks of the error correlation, high leverage data treatment, local minimum, over-fitting and lower fitting. Checking using visualization is possible because of the robust data reduction and the two-dimensional presentation properties of PLS. Other PLS methods such as quadratic PLS (QPLS) and neural net PLS (NNPLS) also have these properties, but they lack the interpretability and high nonlinear regression capacity of the TSK inner relation function. The fuzzy rules of the TSK function provide insights into the model that allow us to make simple linear predictions of its behavior even in the extrapolation range and to interactively change its parameters. These capabilities make FPLS a promising modeling and monitoring method.

5. Prediction Method with FPLS Model

The FPLS model trained on a calibration data set can be used to predict the test data, Let us denote the outer projection vectors of the m factors by matrix form, i.e., **P**, **Q** and **W**. Then, for a new input data set **X** the output data set **Y** can be predicted by using the following steps.

1. Scale **X** by the mean and variance of $\mathbf{X}_0$.

2. Calculate the input score matrix

$$\mathbf{T}=\mathbf{X}\mathbf{W}(\mathbf{P}^T\mathbf{W})^{-1} \tag{21}$$

where $\mathbf{T}=[\mathbf{t}_1, \mathbf{t}_2, \cdots, \mathbf{t}_m]$

3. Predict output score vectors using the TSK inner model defined in Eq. (14), with $\mathbf{c}_h$, $\boldsymbol{\sigma}_h$ and $\mathbf{b}_h$ for each factor h.

$$\hat{\mathbf{u}}_h = f_h(\mathbf{t}_h) \tag{22}$$

4. Predict the scaled **Y**

$$\hat{\mathbf{Y}} = \hat{\mathbf{U}}\mathbf{Q}^T \tag{23}$$

where $\hat{\mathbf{U}} = [\hat{\mathbf{u}}_1, \hat{\mathbf{u}}_2, \cdots, \hat{\mathbf{u}}_m]$ for i=1, 2, ⋯, m.

5. Rescale $\hat{\mathbf{Y}}$ by the mean and variance of $\mathbf{Y}_0$

Using the PLS outer relation and the TSK fuzzy-type inner model, the FPLS method is capable of robustly describing any complex nonlinear system and provides informative biplots. Because FPLS uses the outer relation of PLS, the analytical meaning of the outer projection vectors remains valid. Hence, various PLS monitoring methods are still applicable to FPLS. Moreover, the interpretation based on fuzzy rules gives a new way of monitoring nonlinear systems. For an example, each sample of a system modeled by FPLS can be classified according to the fuzzy rule that has the largest firing strength value.

RESULTS AND DISCUSSION

The FPLS algorithm was applied to two data sets: a simulation data set of a benchmark plant and a real data set from a full-scale biological wastewater treatment plant. Fuzzy model parameters of FPLS were built by using three heuristics rules, where the parameters of FPLS, that is, c_i, σ_i and $\mathbf{b}_i$, are determined by using FCM, the Moody and Darken rule and a global learning procedure, respectively (see the appendix). To compare with other linear and nonlinear PLS, prediction performances of FPLS are compared with linear PLS (LPLS) and quadratic PLS (QPLS).

1. Simulation Benchmark

Eight variables were used to build the X-block in the simulation benchmark [Spanjers et al., 1998; Yoo et al., 2001, 2002]: the influent ammonia concentration ($S_{NH,in}$), the influent flow rate (Q_{in}), the nitrate concentration in the second aerator ($S_{NO,2}$), the total suspended solid concentration in aerator 4 (TSS_4), the DO concentration in aerated tanks 3 and 4 ($S_{O,3}$ and $S_{O,4}$), the oxygen transfer coefficient in aerated tank 5 (K_La_5), and the internal recirculation rate (Q_{int}). The quality variables are the effluent ammonia ($S_{NH,e}$) and nitrate ($S_{NO,e}$). We used data from 14 days of normal (dry weather) operation for the development of the model., The first seven days were used for training the model and the remaining seven days were used for validation.

A comparison of the results of three PLS models is represented in Table 1, where four LVs are selected for each PLS model. Table 1 lists the percent variances captured of training data (%) and mean squared error (MSE) of the validation data set, which shows the regression performance of all the PLS models. Explained variances

Table 1. Percent variance captured (%) and MSE of several PLS models in benchmark

LV	LPLS		QPLS		FPLS	
	X	Y	X	Y	X	Y
1	64.49	40.60	64.49	43.03	64.49	43.68
2	88.96	60.06	88.96	67.85	88.97	71.61
3	91.40	71.04	91.28	77.12	91.45	78.51
4	96.85	72.25	97.04	78.66	97.10	80.00
MSE	0.60		0.46		0.44	

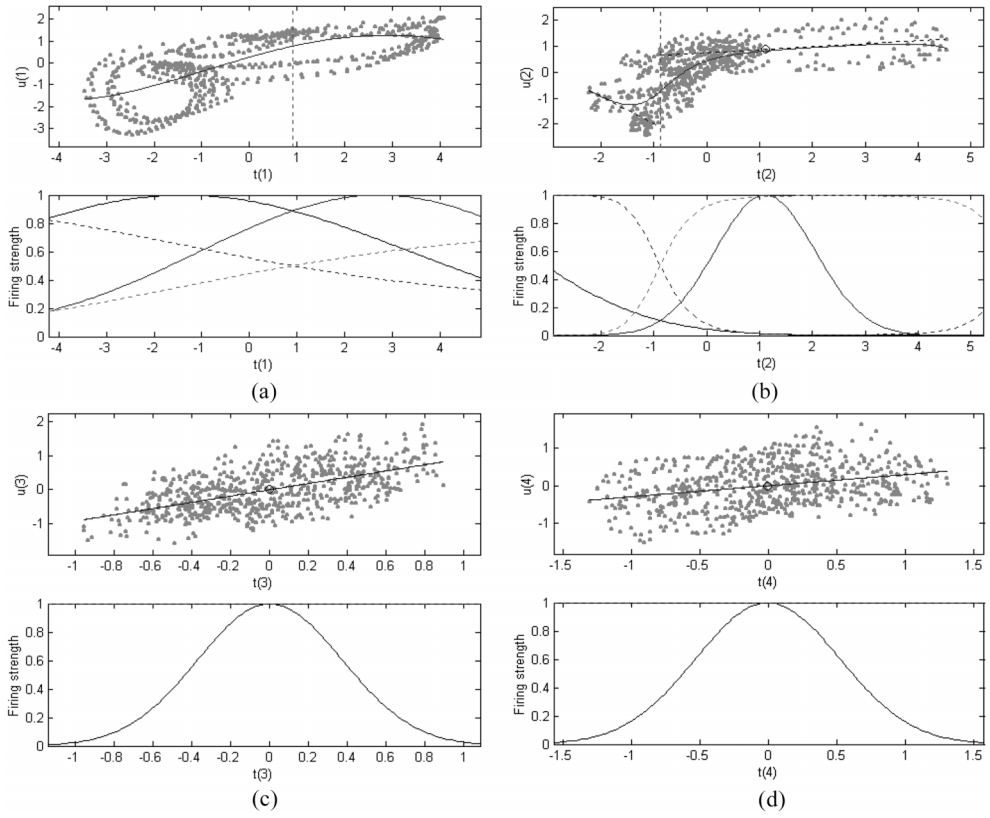

Fig. 3. Scatter plots (upper plot) and firing strength plot (lower plot) of four latent variables in FPLS model (benchmark) (a) first LV (b) second LV (c) third LV (d) fourth LV.

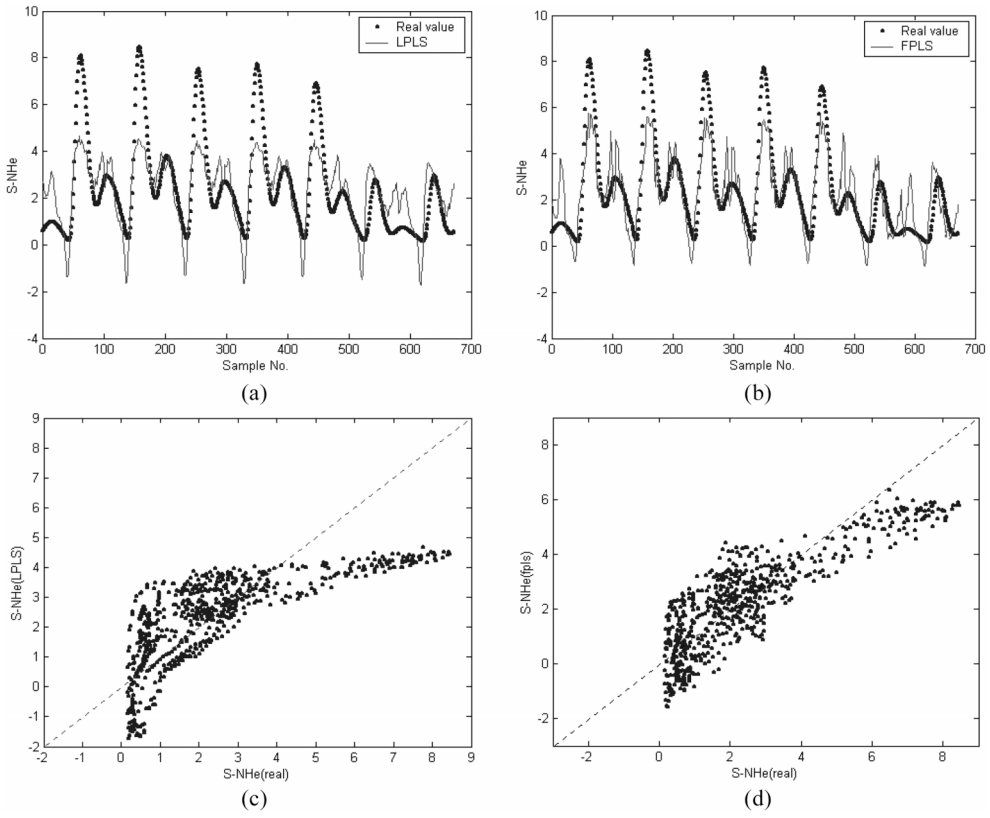

Fig. 4. Comparisons of LPLS and FPLS for the predicted and real value of effluent ammonium S_{NHe} (a) Time series plot of LPLS (b) Time series plot of FPLS (c) Scatter plot of LPLS (d) Scatter plot of FPLS.

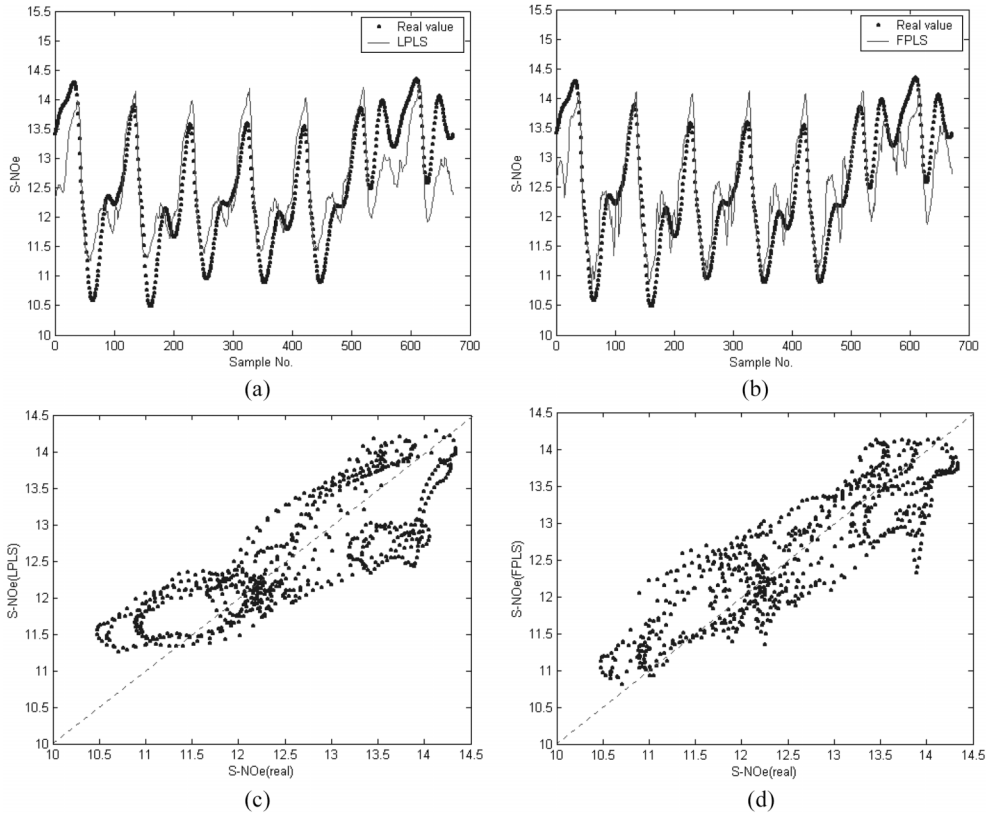

Fig. 5. Comparisons of LPLS and FPLS for the predicted and real value of effluent nitrate, S_{NOe} (a) Time series plot of LPLS, (b) Time series plot of FPLS (c) Scatter plot of LPLS (d) Scatter plot of FPLS.

of X-block using LPLS, QPLS and FPLS model do not show any particular difference. The values of Y-variance captured by the nonlinear models are larger than linear method. The mean squared error (MSE) of the validation data set indicates that best prediction performance is achieved by the FPLS method.

Fig. 3 shows the score plots (upper plot) and firing strength plot of four latent variables (lower plot) in the FPLS model. In the score plots, the small circle represents the center c_i of a membership function shown in the lower plot and the dashed line crossing the circle is its fuzzy rule. In the lower plot, the solid lines represent the firing strength τ_i and the dashed lines represent the normalized firing strength. These plots clearly show the nonlinear nature of the benchmark plant. LPLS gives no direct way to cope with this nonlinearity; however, FPLS can give a direct and interactive way of treating such nonlinearities. To decide the number of fuzzy rules, we applied various numbers of fuzzy rules and heuristic rules to each LV. Then, we found that '2-2-1-1' fuzzy rules for each LV and determining the center of the fuzzy rule of the first LV by FCM yielded the best regression performance on training and validation data sets. The score plots of the third and fourth LVs displayed almost no nonlinearity; hence, we used only one fuzzy rule for each of these LVs. Compared with other nonlinear PLS methods, the FPLS model gives a visual and interactive design capability which can treat such nonlinearities and avoid the over-fitting problem. Figs. 4 and 5 show the prediction results of effluent ammonia and nitrate, $S_{NH,e}$ and $S_{NO,e}$, in the validation data set for LPLS and FPLS method. Time series plots and scatter plots illustrate the prediction improvements that are achievable through the fuzzy regression approach. The scatter plots certify the modeling capability of FPLS.

These results are not surprising since the FPLS model is designed to capture the main variability of the training data set, and the validation data set is generated with the similar statistical properties to the training data. However, the above results are valid on only the normal data set. In other situations, such as other disturbances cases, other models may be better than the FPLS model. The situation and the aim of the models must determine the best model structure.

2. Full-scale WWTP

Process data were collected from a biological WWTP treating coke wastewater from an iron and steel producing plant in Korea, so-called biological effluent treatment (BET). Fig. 6 shows the layout of the studied full-scale plant. This treatment plant uses an ac-

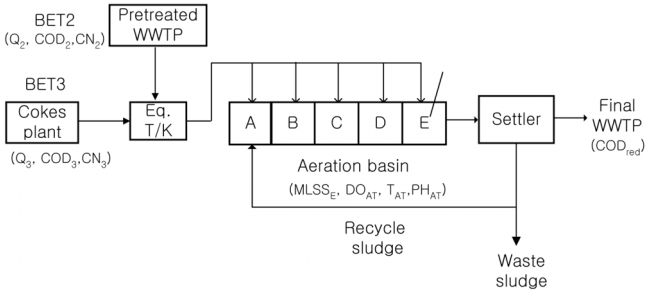

Fig. 6. Plant layout of cokes wastewater treatment plant (BET).

Table 2. Process input and output variables in full-scale WWTP

No	Variable	Description
X_1	Q_2	Influent flow rate from BET2
X_2	Q_3	Influent flow rate from BET2
X_3	CN_2	Cyanide influent from BET2
X_4	CN_3	Cyanide influent from BET3
X_5	COD_2	COD influent from BET2
X_6	COD_3	COD influent from BET3
X_7	$MLSS_{at}$	MLSS concentration at final aeration basin
X_8	$MLSS_r$	MLSS concentration in the returned sludge
X_9	DO_{at}	DO concentration at final aeration basin
X_{10}	T_{inf}	Influent temperature
X_{11}	$T_{aerator}$	Temperature in the final aeration tank
X_{12}	pH_{at}	pH in the final aeration tank
Y_1	SVI_r	SVI in the returned sludge
Y_2	CN_{red}	The reduction of cyanide concentration in the effluent
Y_3	COD_{red}	The reduction of COD in the effluent

Table 3. Percent variance captured (%) and MSE of several PLS models in BET

LV	LPLS		QPLS		FPLS	
	X	Y	X	Y	X	Y
1	17.75	31.76	17.75	31.92	17.75	31.76
2	33.32	47.85	33.32	48.29	33.32	48.79
3	44.01	58.32	43.96	58.70	44.01	59.27
4	52.96	60.61	52.89	60.83	53.00	61.55
5	59.61	62.20	59.55	62.90	60.37	63.00
6	64.64	63.43	65.10	63.90	65.57	64.21
MSE of validation data	1.13		1.12		1.11	
MSE of test data	1.67		1.68		1.71	

tivated sludge process with five aeration basins (each of size 900 m^3) and a secondary clarifier (1,200 m^3). The treatment plant has two influent streams: wastewater arrives either directly from a coke making plant (called BET3) or as pretreated wastewater from an upstream WWTP at another coke making plant (called BET2). The coke-oven plant wastewater is produced during the conversion of coal to coke. This type of wastewater is extremely difficult to treat because it is highly polluted and most of the chemical oxygen demand (COD) contains large quantities of toxic, inhibitory compounds and coal-derived wastewaters that contain, e.g., phenolics, thiocyanate, cyanides, poly-hydrocarbons and ammonium.

Table 2 describes the process variables of **X** and **Y** blocks. Twelve process and manipulated variables, the **X** block, were used to mod-

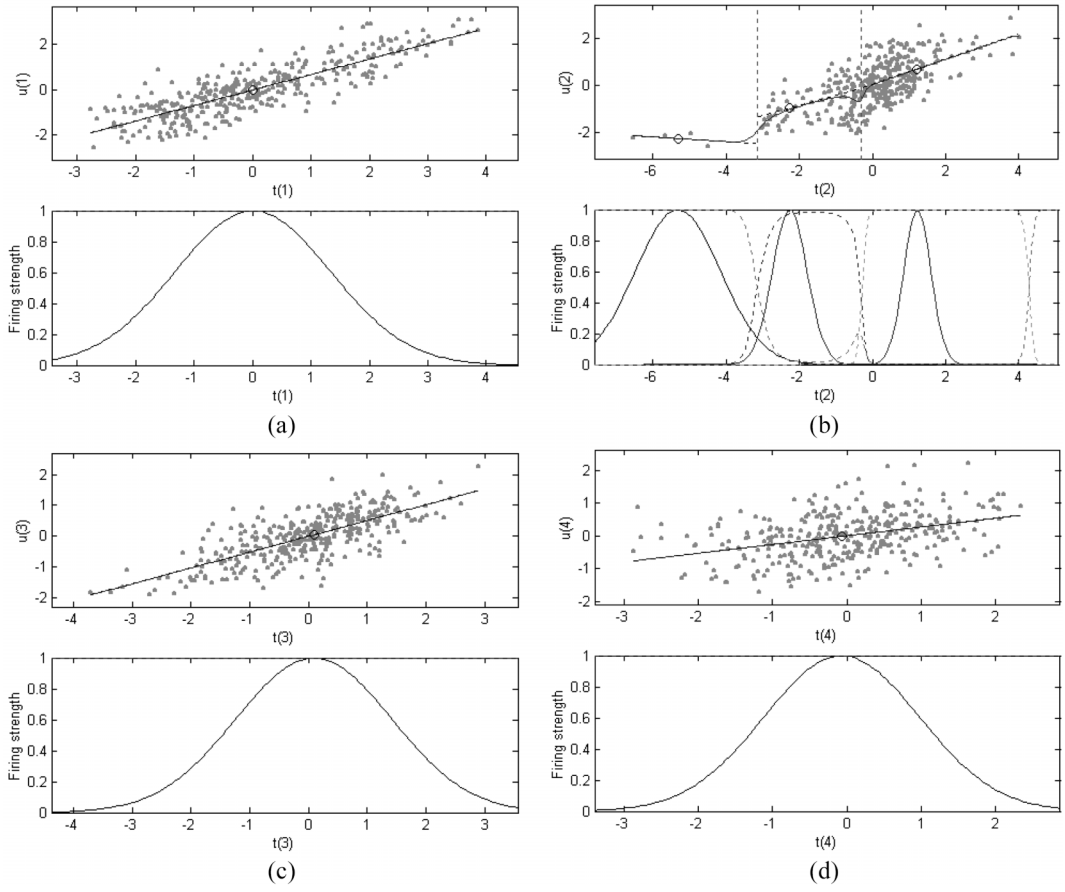

Fig. 7. Scatter plots (upper plot) and firing strength plot (lower plot) of four latent variables in FPLS model in full-scale WWTP (a) first LV (b) second LV (c) third LV (d) fourth LV.

el three process output variables, the **Y** block. The **Y** block consists of the sludge volume index (SVI), the reduction of cyanide (ΔCN), and the reduction of COD (ΔCOD). The process data consisted of daily mean values from 1 January, 1998 to 9 November, 2000 with a total number of 1034 observations. The first 720 observations were used for the calibration of the PLS models. Odd sample numbers were used as the training set and even sample numbers were used as the validation set. The remaining 314 observations were used as a test data set.

The comparison results of three PLS models are represented in Table 3. Six LVs are selected for each PLS model since the **Y**-variance captured by the smaller factors was less than 1%. The MSE of the cross validation data set was calculated with the six LV. Fig. 7 shows the scatter plot and firing strength of the FPLS model with six LVs (the fifth and sixth LV are not shown), which shows the inferred relation between input and output latent variables. Unexpectedly, the data from BET showed no obvious nonlinearity. However, we did find some nonlinear characteristics at the second LV, which leads us to use three fuzzy rules for this factor. The first and later factors showed almost no nonlinearity and, hence, one fuzzy rule was used for each of these LVs.

The value of X and Y-variance captured by the FPLS model is larger than that of LPLS and QPLS methods, and the mean squared error (MSE) of the validation set is smallest for FPLS. However, contrary to our expectation, the MSE for the test data set shows that LPLS and QPLS have better prediction performance than FPLS. During the test data set, WWTP received a large influent load and experienced a significant change in the operating condition. These process transitions altered the sludge, which changed the process dynamics in BET. Because the FPLS model is designed to capture

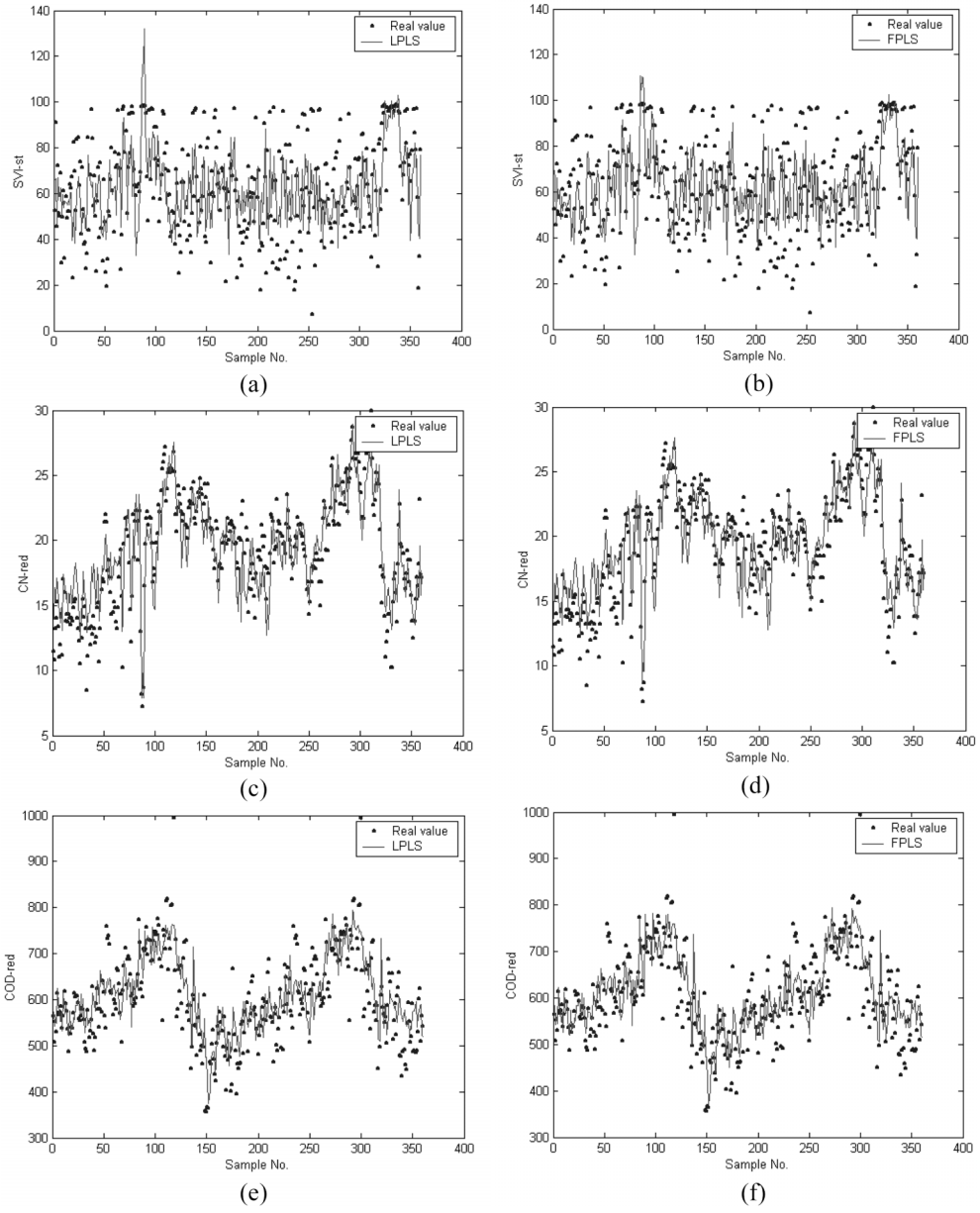

Fig. 8. Time series plots of predicted and actual output in full-scale WWTP (a) SVI with LPLS, (b) SVI with FPLS, (c) ΔCN with LPLS, (d) ΔCN with FPLS (e) ΔCOD with LPLS (f) ΔCOD with FPLS.

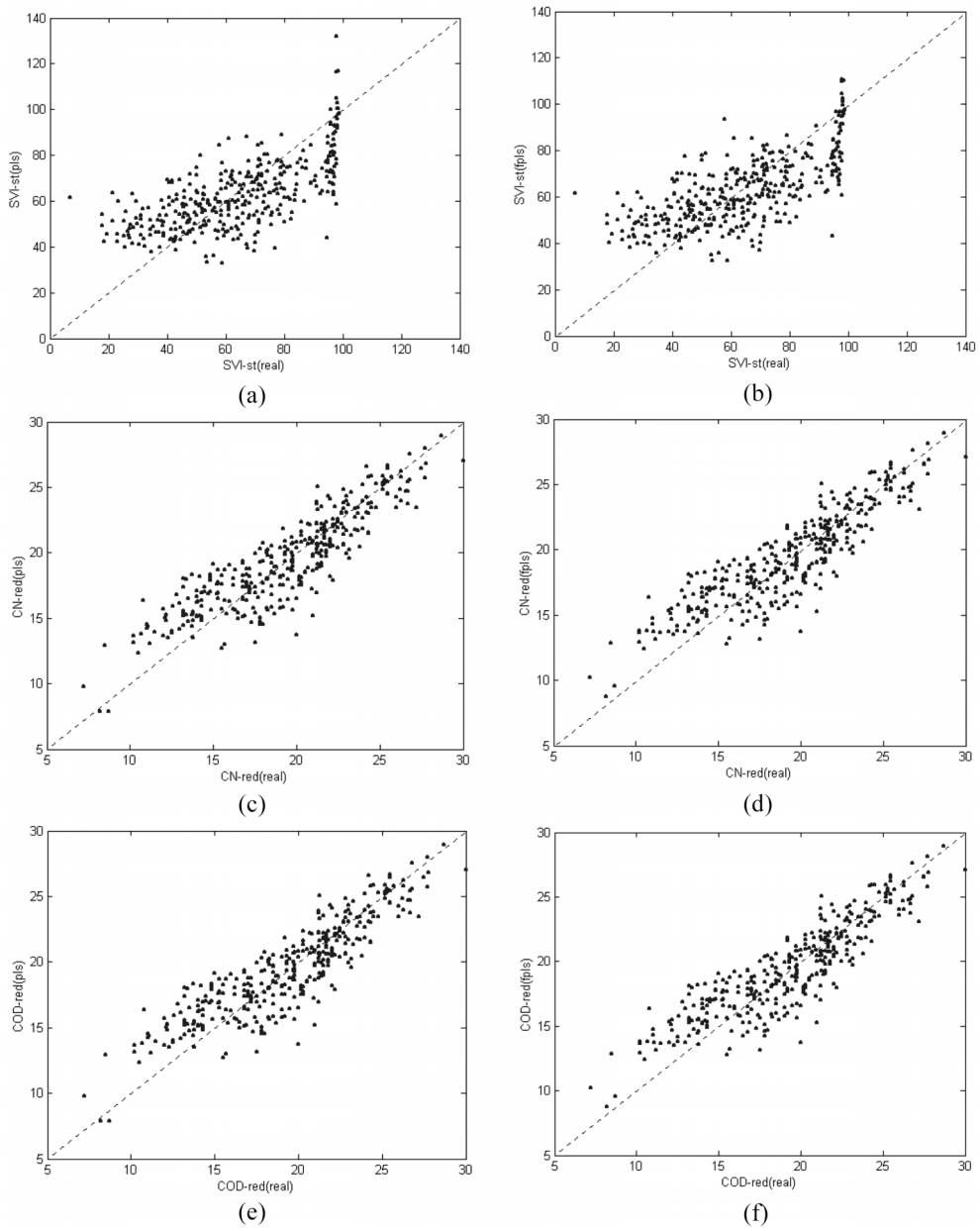

Fig. 9. Scatter plots of predicted and actual output in full-scale WWTP (a) SVI with LPLS, (b) SVI with FPLS, (c) ΔCN with LPLS, (d) ΔCN with FPLS (e) ΔCOD with LPLS (f) ΔCOD with FPLS.

the nonlinear behavior and statistical properties of the training data set, the FPLS model showed poorer prediction result in these disturbances cases. Figs. 8 and 9 show the time series and scatter plots of real and predicted values with LPLS and FPLS model during the validation periods. The prediction performances of COD and CN reduction are satisfactory. But, the prediction of SVI of secondary settler is less good compared to those of the other process quality variables. LPLS and FPLS show a similar prediction performance.

Since it is difficult to make a fair comparison between models whose algorithm has its own characteristics, we will not present a detailed comparison between models, but below we will outline the difference between FPLS and the other nonlinear PLS (NLPLS) methods in two aspects. First, inner relation models of FPLS usually take on a gentler curvature than those of NLPLS, as they are locally weighted averages of linear fuzzy rules and model designers would not favor highly nonlinear shapes of inner relation models whose variables are the results of linear computations. In contrast, other NLPLS models can take on any nonlinear shape to minimize the SSE, providing this shape is permitted by cross validation. If an FPLS model were built only according to the cross-validation result, with no input from the experts, it could have greater curvature. Hence, it ultimately depends on the experts' decision whether to use a conservative model or a sum of square error (SSE) minimizing model [Bang et al., 2003].

Second, the number of regression parameters estimated for each nonlinear PLS inner model depends on a few structural parameters,

such as the order of a polynomial for QPLS, the order of polynomials and the number of knots for the spline PLS (SPLS), the number of neurons for neural net PLS (NNPLS) and the number of rules for FPLS. They also vary depending on the nonlinearity of the modeled system. If the value of the structural parameters is increased, the regression SSE of the model will decrease and the model will take on a more nonlinear shape. Because these structural parameters have different physical meanings, their values cannot be compared with those of another NLPLS. However, if the values are the same, FPLS generally uses more parameters than other NLPLS methods. For example, if the values of the structural parameter are L for both NNPLS and FPLS, an inner model of NNPLS needs 2L+1 regression parameters for the input and output weights of the neurons plus a bias term, whereas that of FPLS needs 4L parameters for c, σ and b. However, this does not mean that FPLS is a more complex model to interpret. Because FPLS analyzes the system using sub models represented by fuzzy rules, the 2L parameters used for b help in the preparation of sub models and the 2L parameters used for c and σ help to interpret the relationship between the input data and the sub models. Therefore, although FPLS uses more regression parameters than other NLPLS methods for the same structural parameter, its superiority as an informative model will rate it highly among the elemental NLPLS methods [Bang et al., 2003; Yoo et al., 2003].

CONCLUSION

The FPLS model was applied to two nonlinear biological processes and the experimental results show the application of the algorithm. The FPLS model not only possesses nonlinear modeling ability, but also the robustness and interpretability of the PLS and fuzzy methods. Moreover, because the TSK fuzzy model is a combination of linear sub-models, it causes the FPLS model to provide more stable estimations of output on extrapolation. The case study clearly showed that it gave good modeling performance and higher interpretability than any other nonlinear PLS modeling method.

ACKNOWLEDGMENT

This work was supported by the Post-doctoral Fellowship Program of Korea Science & Engineering Foundation (KOSEF). This research was also financially supported by the Visiting Postdoctoral Fellowship of the Fund for Scientific Research-Flanders (FWO) which is the Belgium National Fund for Scientific Research.

APPENDICES - IDENTIFICATION OF THE PARAMETERS OF FPLS MODEL

1. The Center of a TSK Fuzzy Model (c_i) using FCM

The center of a TSK fuzzy model c_i in each rule can be decided on the basis of the clusters of CFCM algorithm, which is previously described.

$$c_i = \frac{\sum_{j=1}^{N} u_{i,j}^m t_j}{\sum_{j=1}^{N} u_{i,j}^m}, \; i = 1, 2, \cdots, L \tag{A1}$$

where $u_{i,j}$ is the membership function of each rule i. This clustering method essentially deals with the task of splitting a set of patterns into a number of clusters with respect to a suitable similarity measure. It is able to identify regions where the system can be locally approximated by the TSK model. So, it is applied to obtain a rule-based model focusing on compactness and transparency. As a result, each fuzzy rule built at this point can become a representative regression model of its cluster.

2. The Width of a TSK Fuzzy Model (σ_i) using Moody and Darken' Rule [1989]

The widths of a TSK fuzzy model, σ_i, are determined by using the nearest neighbor heuristic suggested by Moody and Darken, that is,

$$\sigma_i = \left[\frac{1}{p} \sum_{l=1}^{p} (\mathbf{c}_i - \mathbf{c}_l)^2 \right]^{1/2}, \; i = 1, 2, \cdots, L \tag{A2}$$

where c_l (l=1, 2, ⋯, p) are the p (typically p=2) nearest neighbors of the center c_i. In this paper, we assume that all Gaussian membership functions have the same width σ, which is obtained by averaging σ_i in equation (A2) over all L centers.

3. The Linear Parameters of a TSK Fuzzy Model (b_i) using Global Learning Algorithm [Yen et al., 1998]

The parameters, $\mathbf{b}_i$, of the TSK fuzzy model can be determined by using a global learning method. Global learning chooses the parameters of the fuzzy rules that minimize the objective function J_G.

$$J = (y - Xb)^T (y - Xb) \tag{A3}$$

where

$$\mathbf{X} = \begin{bmatrix} w_1(1)w_1(1)x_1(1)w_1(2)x_1(2)\cdots w_1(1)x_r(1)\cdots \\ w_1(2)w_1(2)x_1(2)w_1(2)x_1(2)\cdots w_1(2)x_r(2)\cdots \\ \vdots \\ w_1(N)w_1(N)x_1(N)w_1(N)x_1(N)\cdots w_1(N)x_r(N)\cdots \end{bmatrix}$$
$$\begin{bmatrix} w_L(1)w_L(1)x_1(1)\ldots w_L(1)x_r(1) \\ w_L(2)w_L(2)x_1(2)\ldots w_L(2)x_r(2) \\ \\ w_L(N)w_L(N)x_1(N)\ldots w_L(N)x_r(N) \end{bmatrix} \tag{A4}$$

$$\mathbf{b} = [b_{10}\ b_{11}\ \cdots\ b_{1r}\ \ldots\ b_{L0}\ b_{L1}\ \ldots\ b_{Lr}] \tag{A5}$$

$$\mathbf{y} = [y(1)\ y(2)\ \cdots\ y(N)]^T \tag{A6}$$

, w_i is the normalized firing strength and N is the number of training datasets. If the parameters of the antecedent membership functions are predetermined, the only unknown component in J is the parameter vector b whose elements are the parameters in the linear regression equations of the TSK model. We can use the well-known least squares estimation (LSE) method to solve the parameter vector.

$$\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{A7}$$

Or we can use a computationally efficient method, such as singular value decomposition (SVD), to solve the singularity problem in computation of the inverse of X^TX. Applying SVD to X yields

$$X = U\Sigma V^T \tag{A8}$$

where $\mathbf{U}=[u_1\ u_2\ \cdots\ u_N]^T \in R^{N\times N}$ and $\mathbf{V}=[v_1\ v_2\ \cdots\ v_{2L}]T \in R^{2L\times 2L}$ are

orthogonal matrices, and $\mathbf{\Sigma}=\text{diag}(\sigma_1, \sigma_2, \cdots, \sigma_{2L}) \in R^{N\times 2L}$ is a diagonal matrix with $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{2L} \geq L$. Substituting (A8) into (A3) and after simple manipulations, the minimum Euclidean norm solution of the fuzzy rule parameters, b, is computed as

$$\mathbf{b} = \sum_{i=1}^{s} \frac{\mathbf{u}_i^T \mathbf{y}}{\sigma_i} \mathbf{v}_i \tag{A9}$$

where s is the number of nonzero singular values in $\mathbf{\Sigma}$.

NOMENCLATURE

A_{ij} : fuzzy sets that are characterized by the membership function $A_{ij}(x_j)$
E : residual matrices of the predictor variables
F : residual matrices of the response variables
$f_h(\cdot)$: inner function of TSK fuzzy model
L : the number of rules
m : number of latent variables
P : loading matrix
$\mathbf{p}_h$: loading vector
R_i : the ith fuzzy rule
T : score matrix
$\mathbf{t}_h$: score vector
$\mathbf{x}_i$: $[x_1\ x_2\ \cdots\ x_r]^T$ input variables
X : input data matrix

Greek Letters

c_{ir} : the center of the *i*th Gaussian membership function of the *r*th input variable x_r
t_i : the firing strength of rule R_i
σ_i : the width of the membership function

Abbreviations

FPLS : fuzzy partial least squares
PCA : principal component analysis
PLS : partial least squares
QPLS : quadratic partial least squares
TSK : Takagi-Sugeno-Kang
WWTP : wastewater treatment plant

REFERENCES

Baffi, G., Martin, E. B. and Morris, A. J., "Non-linear Projection to Latent Structures Revisited: the Quadratic PLS Algorithm," *Comp. & Chem. Eng.*, **23**, 395 (1999).

Bang, Y. H., Yoo, C. K. and Lee, I., "Nonlinear PLS Modeling with Fuzzy Inference System," *Chem. Int. Lab. Sys.*, **64**(2), 137 (2003).

Henze, M., Grady, C. P., Gujer, W., Marais, G. R. and Matsuo, T., "A General Model for Single-sludge Wastewater Treatment Systems," IAWPRC Scientific and Technical Report No. 1. International Water Association, UK (1987).

Jang, J. R., Sun, C. and Mizutani, E., "Neuro-fuzzy and Soft Computing," Prentice-Hall, USA (1997).

Liu, J., Min, K. G., Han, C. H. and Chang, K. S., "Robust Nonlinear PLS Based on Neural Networks and Application to Composition Estimator for High-purity Distillation Columns," *Korean J. Chem. Eng.*, **17**, 184 (2000).

Moody, J. and Darken, C. J., "Fast Learning in Networks of Locally-tuned Processing Units," *Neural Computat.*, **1**, 281 (1989).

Qin, S. J. and McAvoy, T. J., "Nonlinear PLS Modeling using Neural Networks," *Comp. & Chem. Eng.*, **16**(4), 379 (1992).

Ragot, J., Grapin, G., Chatellier, P. and Colin, F., "Modeling of a Water Treatment Plant. A Multi-model Representation," *Environmetrics*, **12**, 599 (2001).

Spanjers, H., Vanrolleghem, P. A., Nguyen, K., Vanhooren, H. and Patry, G. G., "Towards a Simulation-benchmark for Evaluating Respirometry-based Control Strategies," *Wat. Sci. Tech.*, **37**(12), 219 (1998).

Tay, J. and Zhang, X., "Neural Fuzzy Modelling of Anaerobic Biological Wastewater Treatment Systems," *J. Env. Eng.*, **125**(12), 1149 (1999).

Wold, S., Kettaneh-Wold, N. and Skagerberg, B., "Non-linear PLS Modeling," *Chemom. Int. Lab. and Sys.*, **7**, 53 (1989).

Yen, J., Wang, L. and Gillespie, W., "Improving the Interpretability of TSK Fuzzy Models by Combining Global Learning and Local Learning," *IEEE Trans. on Fuzzy System.*, **6**(4), 530 (1998).

Yoo, C. K., Kim, D. S., Cho, J. H., Choi, S. W. and Lee, I., "Process System Engineering in Wastewater Treatment Process," *Korean J. Chem. Eng.*, **18**, 408 (2001).

Yoo, C. K., Choi, S. W. and Lee, I., "Disturbance Detection and Isolation in the Activated Sludge Process," *Wat. Sci. Tech.*, **45**(4-5), 217 (2002).

Yoo, C. K., Vanrolleghem, P. A. and Lee, I., "Nonlinear Modeling and Adaptive Monitoring with Fuzzy and Multivariate Statistical Method in Biological Wastewater Treatment Plant," *J. Biotechnology*, **105**(1-2), 135 (2003).