Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Biotechnology 105 (2003) 135–163
JOURNAL OF Biotechnology
www.elsevier.com/locate/jbiotec

Nonlinear modeling and adaptive monitoring with fuzzy and multivariate statistical methods in biological wastewater treatment plants

Chang Kyoo Yoo [a,b,*,1], Peter A. Vanrolleghem [a], In-Beum Lee [b,1]

[a] *Department of Applied Mathematics, Biometrics and Process Control (BIOMATH), Ghent University, Coupure Links 653, B-9000 Gent, Belgium*

[b] *Department of Chemical Engineering, Pohang University of Science and Technology, San 31 Hyoja Dong, Pohang 790-784, South Korea*

Received 25 February 2003; accepted 26 May 2003

Abstract

A new approach to nonlinear modeling and adaptive monitoring using fuzzy principal component regression (FPCR) is proposed and then applied to a real wastewater treatment plant (WWTP) data set. First, principal component analysis (PCA) is used to reduce the dimensionality of data and to remove collinearity. Second, the adaptive credibilistic fuzzy-c-means method is used to appropriately monitor diverse operating conditions based on the PCA score values. Then a new adaptive discrimination monitoring method is proposed to distinguish between a large process change and a simple fault. Third, a FPCR method is proposed, where the Takagi–Sugeno–Kang (TSK) fuzzy model is employed to model the relation between the PCA score values and the target output to avoid the over-fitting problem with original variables. Here, the rule bases, the centers and the widths of TSK fuzzy model are found by heuristic methods. The proposed FPCR method is applied to predict the output variable, the reduction of chemical oxygen demand in the full-scale WWTP. The result shows that it has the ability to model the nonlinear process and multiple operating conditions and is able to identify various operating regions and discriminate between a sustained fault and a simple fault (or abnormalities) occurring within the process data.

Keywords: Adaptive credibilistic fuzzy-c-means (ADFCM); Fuzzy-c-means (FCM) clustering; Fuzzy principal component regression (FPCR); Multivariate statistical analysis; Principal component analysis (PCA); Wastewater treatment process (WWTP)

Abbreviations: ACFCM, adaptive credibilistic fuzzy c-means; ADM, adaptive discrimination measure; AFCM, adaptive fuzzy c-means; ANN, artificial neural networks; CFCM, credibilistic fuzzy c-means; FCM, fuzzy c-means; FPCR, fuzzy principal component regression; PCA, principal component analysis; PCR, principal component regression; TSK, Takagi–Sugeno–Kang; WWTP, wastewater treatment plant.

* Corresponding author. Tel.: +32-9-264-6196; fax: +32-9-264-6220.
E-mail address: changkyoo.yoo@biomath.ugent.be (C.K. Yoo).
[1] Tel.: +82-54-279-2274; fax: +82-54-279-3499.

0168-1656/03/$ - see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/S0168-1656(03)00168-8

Nomenclature

C	number of clusters
d_{ik}	Euclidian distance between an observation and a cluster
$\mathbf{d}_k^i$	σ_k nearest neighbors of $\mathbf{x}_k$
$\mathbf{E}$	residual matrices
F	F distribution
m	fuzzifier
N	number of observations
p	number of independent variables
$\mathbf{P}$	loading matrix
$\boldsymbol{p}_i$	loading vector
$\widehat{\mathbf{P}}$	loading matrix of model subspace
Q	Q statistic
$Q_{\lim}$	confidence limit for Q statistic
$\mathbf{T}$	score matrix
$\mathbf{t}_i$	score vector
$\mathbf{t}_k$	score at time k
T^2	Hotelling's T^2 statistic
$T^2_{\lim}$	confidence limit for Hotelling's T^2 statistic
$\mathbf{U}$	membership function matrix
u_{ik}	membership value of an object (k) on a cluster (i)
$u_{j,lim}$	control limit for membership values in cluster j
$u_{j,max}$	maximum of membership values in cluster j
$\mathbf{V}$	cluster prototype matrix
$\mathbf{v}_i$	cluster prototype
$\mathbf{X}$	input data matrix
$\mathbf{x}_k$	object or observation
$\mathbf{x}_{\text{new}}$	new on-line sample
Greek letters	
β	recursive update parameter
κ_k	mean distance between $\mathbf{x}_k$ and the σ_k nearest neighbors
λ_k	Lagrange multiplier
σ	number of observations whose membership value is above the specific limit value in a cluster
ψ_k	credibility of an object

1. Introduction

Due to increasing environmental constraints and the necessity of reliable wastewater treatment, efficient modeling and monitoring methods are becoming more and more important. A clear necessity for reliable modeling and monitoring techniques of biological wastewater treatment plants (WWTP) exists to keep 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. On the other hand, process monitoring and early fault detection methods in the biological process are very efficient to execute corrective actions well before a dangerous situation occurs.

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 physical links and what degree of confidence can be attributed to the measurements...? All these questions are concerned with the characterization of processes, 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 the parameters which intervene in these models, the simplification of the models to make them compatible with real-time use and the validation of the 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 the establishment of a knowledge model, and the second uses only statistical processing of data to supply 'black-box' type models, which take no account whatever of the nature and intensity of the physical links between the variables. The 'truth' seems a trade-off between these two viewpoints, leading to a 'grey-box' model which uses simplified hypotheses on the system's operation and fundamental equations of physics, for example, in the form of matter balances and energy balances, statistics and data processing tools (Ragot et al., 2001).

To date, the most successful model and the industrial standard in WWTP has been the deterministic mechanistic model, called Activated Sludge Model no. 1 or ASM1 (Henze et al., 1987). It has proven to be the most 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 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.

It is essential to bear in mind that both physical and statistical models are applicable to systems under very precise operating conditions, but these conditions cannot always be controlled. These models are therefore applied by making restrictive hypotheses, and a single model can therefore not be claimed to constitute complete descriptions of the actual operation of a process. Furthermore, WWTPs have different behavior patterns depending on the influent loads, temperature and the activity of microorganisms. The models used for the various operating conditions should be different. The challenge is, however, to build a single model 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, is performed using weighting functions that reflect the domains of influence of each model (Wang et al., 1998; Yen et al., 1998; Tay and Zhang, 1999, 2000; Ragot et al., 2001).

In recent years, data-driven 'black-box' modeling approaches have been successfully applied to various WWTPs (Cote et al., 1995; Van Dongen and Geuens, 1998; Lee and Park, 1998; Lee et al., 2002; Ragot et al., 2001; Yoo et al., 2002a,b,c). They have a distinct ability to model nonlinear dynamic systems without requiring a structural knowledge of the process to be modeled. Owing to the complex interaction and various influent variations, it is difficult to establish the relation between input and output variables. Artificial neural networks (ANN) are able to learn complex nonlinear relationships in the biological process and capture their behavior, but are not able to help improve the heuristic understanding of the operating problems. Fuzzy modeling has been an efficient alternative to describe this nonlinear biological process (Fu and Poch, 1998; Huang and Wang, 1999; Tay and Zhang, 1999, 2000; Ragot et al., 2001; Yoo and Lee, 2001). Recently, Takagi–Sugeno–Kang (TSK) fuzzy models have received much attention because of their prediction ability and suitability to continuous process modeling. The TSK model allows to aggregate a set of linearized models into a global model to

approximate a complex nonlinear system with relatively small complexity. Many researchers have studied the TSK model extensively but only little research has been done in the application of TSK fuzzy modeling to WWTPs.

When there is a high dimensionality and correlation exists between measured data (collinearity) of a WWTP, it is significant to identify unimportant variables, remove them and reduce the dimension of the dataset. Principal component analysis (PCA) is a multivariate statistical data analysis which uses projection into latent variables (LVs) to reduce high-dimensional and strongly correlated data to a much smaller data set that can then be interpreted. This leads to reduced computational time, memory requirement and easy interpretation. This approach is important for problems with a large number of input variables and features in chemical and biological processes, and derives a compact set of data set, rule base, visual interpretability and partial solution to the determination of the number of rules. Kurtanjek (1998) presented modeling techniques for a principal component (PC) based ANN in Baker's yeast production, where the model was applied to adaptive control of the fermentation by internal model control. Teppola (1999) used the score vectors of partial least squares (PLS) as clustered data in fuzzy c-means clustering (FCM). They discovered that the compression of data before clustering caused FCM to become more stable and efficient when the original variables were highly correlated and high-dimensional. Joo et al. (2000) showed the necessity of data preprocessing of neural networks in the determination of the coagulant dosing rate, the preprocessing extracted the required information from the noisy data. Choi and Park (2001) proposed a hybrid ANN which combined PCA as a preprocessing step of an ANN and applied it to the prediction of an industrial wastewater treatment plant. Hybrid ANN showed enhancement of the prediction capability and reduced the over-fitting problem of a neural network from a multidimensional and collinear data set.

Process data from processes that work under several operating conditions can be classified into several clusters using an appropriate clustering algorithm, where each cluster or group of clusters covers a specific operating condition. The clustering procedure involves the creation of several clusters and the identification of the cluster centers. Most data samples belong to one or more clusters. However, samples that correspond to mixed and intermediate states between several operating conditions do not belong to any cluster, and it is easy to mistake such samples for modeling and monitoring. We therefore need some method for modeling and monitoring that allows to distinguish between various operating regions if we are to understand changes in operating states.

On the other hand, having a process monitoring system of the biological treatment process is very important because the recovery from failures is time-consuming and expensive. Moreover, some of the changes are not very obvious to detect and may grow gradually until they produce a serious operational problem. Therefore, early fault detection and isolation in the biological process is very efficient as it allows to execute corrective action well before a dangerous situation happens. A monitoring system for abnormalities is of primary concern. To accomplish this task, a reliable detection procedure is needed. A few statistical process monitoring techniques based on a data-based approach are available in the field of WWTPs (Huang and Wang, 1999; Teppola, 1999; Rosen and Lennox, 2001; Yoo et al., 2002a,b,c).

In this paper, we propose a nonlinear modeling and monitoring method which models not only the different normal operating regions of a process with various operating conditions but also the intermediate states between these normal states. First, we use a PCA method to reduce the dimensionality of the data and to remove collinearity. Second, the transformed data in the reduced dimension spanned by the transformation vectors of the PCA are classified into several clusters using an adaptive credibilistic fuzzy-c-means (ACFCM) clustering algorithm. Here, we propose an adaptive discriminant monitoring index to trace successive changes of states between several operating conditions. Finally, a fuzzy principal component regression (FPCR) method is proposed in which the TSK fuzzy model gives a powerful interpretation as well as a nonlinear

regression property to predict the key components of the WWTP. In summary, a nonlinear modeling and monitoring approaches using PCA, adaptive credibilistic fuzzy-c-means (CFCM), an adaptive discriminant monitoring index, and TSK fuzzy regression is presented in this paper. It is noted that this approach is the method of choice when a process has multi-operating conditions such as WWTP.

The outline of this paper is as follows. First, we briefly present the PCA and a simple statistical monitoring method. Second, fuzzy-c-means (FCM) and CFCM clustering is presented. Third, the adaptive discriminant monitoring method is proposed. Finally, we introduce the basic TSK fuzzy model and then propose the FPCR method. The proposed FPCR method is applied to predict the important output variables in a real full-scale wastewater treatment plant and the results are discussed. Finally, the conclusion of this article is given.

2. Materials and methods

2.1. Principal component analysis

Because the raw process data frequently contain severe correlations between measured variables and are of high dimensionality, it is useful to introduce multivariate statistical LV methods to provide independency and to reduce dimensionality. Two projection methods that are commonly used to explain the relationships among complex data sets are PCA and PLS, which are statistical methods that explain the variance–covariance matrix structure using a small number of PCs. These methods are usually used either to reduce the data dimension while retaining the important information or to display the data information in a form that can be easily interpreted.

PCA is an optimal dimensionality reduction technique in terms of capturing the variance of the data. It decomposes the data matrix $\mathbf{X}$ into the sum of the outer product of vector $\mathbf{t}_i$ and $\mathbf{p}_i$ and the residual matrix, $\mathbf{E}$.

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E} = \sum_{i=1}^{p} \mathbf{t}_i \mathbf{p}_i^{\mathrm{T}} + \mathbf{E} \tag{1}$$

where $\mathbf{t}_i$ is a score vector which contains information about the relation of samples, $\mathbf{p}_i$ is a loading vector which contains information about the relation of variables and p is the number of independent variables. A score vector is orthogonal and a loading vector is orthonormal. PCA can be obtained by several methods such as singular value decomposition (SVD) and nonlinear iterative partial least squares. Latent projection into the PC space reduces the original set of variables to l LVs or PCs. Usually, all the score and loading vectors are not required to explain important data information. In practice, only a few PCs are often sufficient to explain most of the variations in the data. Designing the PCA model, we have to determine the PC number. It should be determined considering both the curse of dimensionality and the loss of data information. Several techniques exist to determine the number of PCs, but there appears to be no dominant technique. Numerous methods have been developed for determining the number of PCs, including scree plot, parallel analysis, cross-validation and Akaike Information Criterion (AIC) (Jackson, 1991; Chiang et al., 2001).

On the other hand, once PCA models have been calculated, and those of interest retained, it is possible to calculate values to determine whether the process is in control or not, called '*process monitoring*'. In the monitoring phase, both the score values and the residuals are monitored in order to detect the occurrence of process faults and disturbances. For process monitoring, statistical control limits are needed to determine whether a process is in-control. Hotelling's T^2 and Q statistics are usually used for this purpose. After decomposing the observed data, the score value in the model space at time k,

$$\mathbf{t}_k = [\mathbf{p}_1 \quad \mathbf{p}_2 \quad \cdots \quad \mathbf{p}_m]^t \mathbf{x}_k = \hat{\mathbf{P}}^t \mathbf{x}_k \in \Re^m \tag{2}$$

is distributed as $N(0, \mathbf{\Lambda}_s)$, where $\mathbf{\Lambda}_s$ is the left-upper $m \times m$ diagonal part of $\mathbf{\Lambda} = \mathbf{P}^t\mathbf{R}\mathbf{P}$ and $\mathbf{R}$ is the sample covariance matrix. $\mathbf{t}_k$ is thus a m-dimensional reduced representation of the observed

vector $\mathbf{x}_k$. On the other hand, the residual at time k

$$\mathbf{e}_k = \mathbf{x}_k - \hat{\mathbf{x}}_k = (\mathbf{I} - \hat{\mathbf{P}}\hat{\mathbf{P}}^t)\mathbf{x}_k \in \Re^p \tag{3}$$

is the part not explained by the PCA model. Generally, the squared weighted score ($T_k^2 = \mathbf{t}_k^t \mathbf{\Lambda}^{-1} \mathbf{t}_k$) and the squared residual ($Q_k = \mathbf{e}_k \mathbf{e}_k^t$) are used as monitoring indices for process monitoring or fault detection. Given that we do not know the covariance matrix of the data exactly, the two quantities may not be distributed as the generalized χ^2. Fortunately, the approximated $100(1-\alpha)\%$ control limit for T^2 can be calculated by means of a F-distribution as

$$T_{\lim}^2 = \frac{m(n-1)}{n-m} F(m,\ n-1;\ \alpha) \tag{4}$$

where $F(m,n-1;\alpha)$ is a F-distribution with degree of freedom m and $n-1$ with level of significance α. On the other hand, the $100(1-\alpha)\%$ control limit for Q is

$$Q_{\lim} = \theta_1 \left[\frac{c_\alpha \sqrt{2\theta_2 h_0^2}}{\theta_1} + 1 + \frac{\theta_2 h_0 (h_0 - 1)}{\theta_1^2} \right]^{1/h_0} \tag{5}$$

where $\theta_j = \Sigma_{s+1}^{p} (\Sigma_{ii})^j$ for $j = 1,\ 2,\ 3,\ h_0 = 1 - 2\theta_1\theta_3/3\theta_2^2$ and c_α is the normal deviate cutting off an area α of the upper tail of the distribution if h_0 is positive and under the lower tail if h_0 is negative (Johnson and Wichern, 1992; Wise and Gallagher, 1996; Montgomery, 2001; Teppola, 1999; Yoo et al., 2002a).

For a new on-line sample $\mathbf{x}_{\text{new}}$, if $T_{\text{new}}^2 < T_{\lim}^2$ and $Q_{\text{new}} < Q_{\lim}^2$, we consider the process to be in-control with $100(1-\alpha)\%$ confidence. Otherwise, the process may be out of control. Here, the T^2 value is used to detect faults associated with abnormal variations within a model subspace, whereas the Q value is used to detect new events that are not taken into account in the model subspace. The Q value additionally tells us whether or not the current model subspace is valid. Thus, the evaluation of the monitoring statistics of the PCA model can be a '*reality check*' in order to decide if the prediction of a FPCR model is reliable and it can thus be used to create a monitoring concept.

2.2. Fuzzy c-means clustering

Clusters can be described as continuous regions of the measurement space containing a relatively high density of points, separated from other high density regions by regions containing relatively low densities of points. Clustering algorithms can be divided into two types: hard (crisp) clustering and fuzzy clustering. In the fuzzy schemes, each data point can simultaneously belong to more than one cluster, whereas in the crisp clustering based on a probabilistic scheme, each data point belongs exclusively to a single cluster. Naturally, fuzzy clustering algorithms provide more realistic partitions than crisp clustering algorithms.

The FCM clustering algorithm is a clustering method in which an object can be a member of different classes at the same time, i.e., it is possible to be between two or more classes (Bezdek, 1981; Teppola, 1999; Rosen, 2001). This method is an unsupervised classification algorithm which uses a certain objective function to iteratively determine the local minima. The objective function, which is minimized iteratively, is a weighted within-groups sum of distances $d_{k,i}$. The weighting is done by multiplying the squared distances by membership values $u_{k,i}$.

$$J_m(C,\ m) = \sum_{i=1}^{C} \sum_{k=1}^{N} (u_{k,i})^m d_{k,i}^2 \tag{6}$$

where C is the total number of clusters, N is the total number of objects in the calibration data, $d_{k,i}$ is the distance between an object k and a prototype (cluster) i, $u_{k,i}$ is the membership function. An important parameter in this algorithm is the exponent m. This exponent m determines the fuzziness of the classification. The closer the exponent is to unity, the closer the fuzzy partition is to a crisp partition. The higher the exponent, the more vague the boundary between different clusters is. This also serves as a way to diminish the effect of noise in identifying the cluster prototypes. Roughly estimating, a good value for the exponent (m) lies somewhere between 1.5 and 2.5. Different types of distance measures can be used to measure the distance between an object $k(x_k)$ and a cluster $i(v_i)$, for example, an Euclidean and a Mahalano-

bis distances (Teppola, 1999). Euclidean distance measure:

$$d_{k,i}^2 = (x_k - v_i)(x_k - v_i)^T, \quad \forall i,\ k \tag{7}$$

Mahalanobis distance measure:

$$d_{k,i}^2 = (x_k - v_i)A_i^{-1}(x_k - v_i)^T, \quad \forall i,\ k \tag{8}$$

A covariance matrix, $\mathbf{A}_i$, can be estimated by using the following sample covariance.

$$\mathbf{A}_i = \frac{\sum_{k=1}^{N}(x_k - v_i)^T(t_k - v_i)}{N-1}, \quad \forall i \tag{9}$$

where x_k is the measured value, v_i is the cluster centroid and N is the number of a training data set. Membership values for the individual objects are calculated using Eq. (10). The most important properties of FCM clustering are that the sum of an object's membership value over all clusters is equal to unity and the memberships get values only between zero and one. Membership $u_{i,k}$ to a certain cluster i of an instance at time k can be calculated by:

$$u_{k,i} = \frac{1}{\sum_{j=1}^{C}(d_{k,i}^2/d_{k,j}^2)^{2/(m-1)}}, \quad \forall i,\ k \tag{10}$$

$$\sum_{i=1}^{C} u_{k,i} = 1, \quad \forall k \tag{11}$$

After computing the membership values for all calibration objects, the cluster centers (v_i) are described by prototypes, Eq. (12), which are fuzzy weighed means.

$$v_i = \frac{\sum_{k=1}^{N}(u_{k,i})^m x_k}{\sum_{k=1}^{N}(u_{k,i})^m}, \quad \forall i \tag{12}$$

The minimization of Eq. (6) commences after giving initial values for the prototypes. Then Eqs. (8)–(12) are repeated successively in each iteration step. In the monitoring stage of a new test sample, a new value is computed using Eq. (13) along with the chosen distance measure (Teppola, 1999).

$$u_{N+1,i} = \frac{1}{\sum_{j=1}^{C}(d_{k,i}^2/d_{k,j}^2)^{2/(m-1)}} \tag{13}$$

2.3. Credibilistic fuzzy c-means clustering

The most basic fuzzy clustering scheme, FCM clustering, was developed by Bezdek (1981). Many modifications have been made to FCM to address its shortcomings. To identify the relative belonging of data points to each cluster, Krishnapuram and Keller (1993) introduced the possibilistic c-means (PCM) clustering and then showed intra-cluster information. PCM has a mode-seeking property and can reduce the effect of noise. Pal et al. (1997) proposed the fuzzy possibilistic c-means clustering method, which simultaneously generates inter- and intra-cluster information. In the search for a noise-resistant algorithm, many robust clustering algorithms such as the noise cluster algorithm and the least biased fuzzy clustering method (Beni and Liu, 1994) were developed. Moreover, FCM was modified in various ways to automatically determine the relevant number of clusters in the data set.

Because outliers in data can significantly affect the prototype clusters and may have large membership values, the issue of how to deal with noise in clustering has been of particular importance. Here we use the CFCM method for clustering. The CFCM algorithm was proposed by Chintalapudi and Kam (1998). The objective function of CFCM and its constraint are as follows.

$$J(\mathbf{V},\ \mathbf{U}) = \sum_{k=1}^{N}\sum_{i=1}^{C}(u_{ik})^m\|\mathbf{x}_k - \mathbf{v}_i\|^2 \quad \text{subject to} \quad \sum_{i=1}^{C} u_{ik} = \psi_k, \quad k = 1, \ldots,\ N \tag{14}$$

where C is the total number of clusters, N is the number of observations, $u_{ik} \in [0,1]$ represents a membership value of the kth object on the ith cluster, m is known as the fuzzifier, $\boldsymbol{v}_i$ is the ith cluster prototype, $\mathbf{x}_k$ is the kth object, and ψ_k is the credibility of object $\mathbf{x}_k$. In this research, the PCA score values of $\mathbf{t}_k$ are used in FCM instead of

the original data of $\mathbf{x}_k$. The difference between FCM and CFCM is the constraint term, i.e., CFCM becomes FCM when $\psi_k = 1$ for all k. Using the Lagrange multiplier method, the above optimization problem is converted to the following unconstrained optimization problem,

$$J(\mathbf{V},\ \mathbf{U}) = \sum_{k=1}^{N}\sum_{i=1}^{C}(u_{ik})^m \|\mathbf{x}_k - \mathbf{v}_i\|^2 - \sum_{k=1}^{N}\lambda_k\left(\sum_{i=1}^{C}u_{ik} - \psi_k\right) \tag{15}$$

where λ_k is a Lagrange multiplier.

The optimization problem is solved by setting the gradient of J with respect to u_{ik} and v_i equal to zero,

$$\frac{\partial J(\mathbf{V},\ \mathbf{U})}{\partial u_{ik}} = m u_{ik}^{m-1} d_{ik}^2 - \lambda_k\left(1 - \frac{\partial \psi_k}{\partial u_{ik}}\right) = 0 \ \text{ or}$$

$$u_{ik} = \left\{\lambda_k\left(1 - \frac{\partial \psi_k}{\partial u_{ik}}\right)(m d_{ik}^2)^{-1}\right\}^{\frac{1}{m-1}} \tag{16}$$

$$\frac{\partial J(\mathbf{V},\ \mathbf{U})}{\partial \mathbf{v}_i} = 2\sum_{k=1}^{N}u_{ik}^m d_{ik} + \sum_{k=1}^{N}\lambda_k\frac{\partial \psi_k}{\partial \mathbf{v}_i} = 0 \tag{17}$$

where $d_{ik} = \|\mathbf{x}_k - \mathbf{v}_i\|$ is the Euclidian distance between the kth observation and the ith cluster. Substituting Eq. (16) into the constraint in Eq. (14) yields,

$$\lambda_k = \psi_k^{m-1}\left[\sum_{i=1}^{C}\left\{\left(1 - \frac{\partial \psi_k}{\partial u_{ik}}\right)(m d_{ik}^2)^{-1}\right\}^{\frac{1}{m-1}}\right]^{-(m-1)} \tag{18}$$

and by combining Eqs. (18) and (16) and rearranging, we obtain

$$u_{ik} = \frac{\psi_k}{\sum_{j=1}^{C}\left(\frac{d_{ik}^2(1 - \partial\psi_k/\partial u_{jk})}{d_{jk}^2(1 - \partial\psi_k/\partial u_{ik})}\right)^{\frac{1}{m-1}}} \tag{19}$$

In most cases, the credibility is a function of all variables involved in the clustering scheme. The more similar an object is to another object, the larger the credibility value. Similarity can be represented in several ways; for example, the distance between two objects is a measure of their similarity. That is, credibility can be considered as the degree of isolation of a vector in the feature space. Let $\{\mathbf{d}_k^i \in \mathbf{X} | i = 1, \ldots, \sigma\}$ be the σ nearest neighbors of $\mathbf{x}_k$ in terms of Euclidian distance.

The credibility of a vector $\mathbf{x}_k$ is defined as

$$\psi_k = 1 - \frac{(\kappa_k - \min(\kappa_1, \ldots, \kappa_n))}{\max(\kappa_1, \ldots, \kappa_n) - \min(\kappa_1, \ldots, \kappa_n)} \tag{20}$$

$$\kappa_k = \frac{\sum_{i=1}^{\sigma}(\|\mathbf{d}_k^i - \mathbf{x}_k\|)}{\sigma} \quad \text{and} \quad \sigma = \gamma\frac{N}{C} \tag{21}$$

In Eq. (20) κ_k is the mean distance between $\mathbf{x}_k$ and its σ nearest neighbors and $\gamma \in [0,\ 1]$ is a constant. Under this scheme, outliers are dissimilar to most other objects because they are far away from other objects. Hence, outliers have low credibility values. On the other hand, non-outliers are similar to most other objects and therefore have high credibility values.

If we use the above credibility, which is a function only of $\mathbf{x}$, Eqs. (17) and (19) are simplified to

$$u_{ik} = \psi_k\left[\sum_{j=1}^{C}\left(\frac{d_{ik}^2}{d_{jk}^2}\right)^{\frac{1}{m-1}}\right]^{-1} \quad \text{and} \quad \mathbf{v}_i = \frac{\sum_{k=1}^{N}u_{ik}^m\mathbf{x}_k}{\sum_{k=1}^{N}u_{ik}^m} \tag{22}$$

The CFCM clustering method explained above has the following good characteristics. First, it provides inter and intra-cluster information simultaneously. That is, for an arbitrary sample, we can know its relative membership to each cluster. Furthermore, for each cluster, we can know the samples' relative membership to it. Second, the isolated samples, which are dissimilar to the rest of the samples in the data set, are automatically identified since they have low credibility values compared to non-outliers.

2.4. Adaptive discriminant monitoring method

Based on the CFCM clustering algorithm, the complete data set can be classified into several pattern groups. Each of these groups is then regarded as a normal operating region for a specific operating condition, and is monitored using the conventional monitoring method. However, this scheme is limited to time-invariant processes. Process data obtained under adaptive and multiple operating conditions which have a number of operating models and changes in operating conditions (such as biological WWTP) should be monitored carefully. In many cases it may be necessary to include some trend-following capability, thus making the clustering procedure adaptive. It is hard to discriminate changes due to adaptive and multiple conditions, because both may appear to be out of control when compared with one operating region. In this section, we will propose an adaptive discriminant monitoring method that uses the adaptive CFCM and a membership value monitoring system.

Most WWTPs are time-varying due to changes of influent characteristics, temperature, and microorganism activity. It may be useful to use this knowledge in partitioning a given set to classify more data during on-line operation. The data clusters will evolve with the process shifts. In general, the clusters could be chosen in such a way that they represent a set of preferred states among various operating conditions. Because of this continuous drift from one end of the operating condition to the other, the fuzzy boundaries and memberships also evolve. In order to achieve this objective, the FCM and CFCM approach can be made adaptive or flexible (Marsili-Libelli and Müller, 1996; Marsili-Libelli, 1998; Teppola, 1999).

A basic form for adaptive updating was proposed by Marsili-Libelli and Müller (1996). It is based on using the new membership values to change the prototype locations, assuming that the generic $(n+1)$-th point is considered and that all the previous n points have already been classified. However, this adaptation rule works well only in the case of a relatively small number of objects and for a small period of time because the effect of every new object will be added into the summation terms. The larger the summation terms, the less adaptive the rule becomes. This concept is very similar to the recursive PCA monitoring method which consists of recursive updating the mean, variance and covariance matrix, and especially updating of the mean (Li et al., 2000). Conversely, adaptive FCM and CFCM are updating the center of each cluster. In this paper, we used the modified flexible updating rule suggested by Teppola (1999) in order to preserve the adaptation efficiency. An effect of every new object (x_{n+1}) is multiplied by the number of added objects times the parameter β.

$$v_{i|n+1} = \frac{\sum_{k=1}^{n+1}(u_{k,i})^m \mathbf{x}_k}{\sum_{k=1}^{n+1}(u_{k,i})^m} = \frac{S(N) + \beta N (u_{n+1,i})^m x_{n+1}}{M(N) + (u_{n+1,i})^m} \tag{23}$$

where $S(N) = \Sigma_{k=1}^{N}(u_{k,i})^m \mathbf{x}_k$, $M(N) = \Sigma_{k=1}^{N}(u_{k,i})^m$, N is the number of calibration sets, and n is the parameter which continuously counts the number of added objects. This restriction works in a way that clusters are virtually fixed in the sense that a drifting of a cluster' location is not allowed. The effect of the modified flexible rules can be seen as a contrast enhancement in a membership plot, i.e., a plot in which the memberships of each cluster are shown against time. There is no rule how to determine to use its parameters. Teppola (1999) suggested the following 'quick and dirty' rule: the parameters are used to put more weight on the new objects in order to more quickly respond to changes in the process mean and the value for the parameter β can be set to be between 0.01 and 0.1. Other updating rules or criteria such as the partition entropy, moving window and conditional updating criteria can be used as well as an evaluation criterion for the sharpness increase of the updated partition.

The membership values of measurements for all clusters obtained through CFCM can give information on the current operating condition. That is, the membership value for the cluster responsible

for the current operating condition is relatively large compared to the membership values for other clusters. Therefore, the following adaptive monitoring guideline is proposed. If $u_{jk} := \max(u_{ik}) \geq u_{j,\lim}$, where u_{jk} is the new value of a membership function for the new object calculated with Eq. (23), and $u_{j,\lim}$ is a predefined control limit for the cluster j responsible for the current sample, then the process condition is considered to be in-control. The control limits for the membership values of every cluster are determined by simple statistics or heuristic methods, such as a confidence limit for the average u_j using the Student's t-test and the percent of maximum membership values in the calibration data set (Johnson and Wichern, 1992; Montgomery, 2001; Yoo et al., 2002b). In this paper, we use the simple statistics, i.e., Student's t-test. The control limits allow to find the $(1-\alpha)100\%$ confidence interval of each membership value. Based on the data obtained under normal operating conditions, the interval $(1-\alpha)100\%$ of the membership values is expressed as

$$\overline{u_j} - t_{1-\alpha/2,N-1}\frac{s\{u_i\}}{\sqrt{N}} \leq u_j \leq \overline{u_j} + t_{1-\alpha/2,N-1}\frac{s\{u_i\}}{\sqrt{N}} \quad (24)$$

where $\overline{u_j}$ is the sample mean of membership values of the jth cluster and $s\{u_j\}$ is the sample variance of membership values of the jth cluster. Typically the value of α is 0.05 for the warning limit and 0.01 for the action limit.

Various abnormal situations can be easily detected using the method described above. However, this method may on occasion mistakenly identify a process shift from one operating condition to another as a fault. In practice, it is hard to immediately distinguish between a process shift and a fault, because the detection scheme simply shows both situations to be out of control. Fortunately, the time traces of the membership values of an adaptive CFCM gives information that can be used to overcome this problem.

Teppola (1999) reported that the membership value for an extreme object such as an outlier approaches the value $1/C$, which is the reciprocal of the number of clusters, whereas a membership value close to zero states that an object belongs to another class. That is, membership values of different classes are connected to each other. Based on the connectivity of membership values of each class, we propose an adaptive discrimination measure (ADM) method to distinguish between a large disturbance change and a short disturbance, which also considers the adaptive characteristics existing in the process (Choi et al., 2003). The basic idea comes from the following simple observation of the variation of membership functions, which differs between a large process change and a momentary disturbance. For a momentary disturbance, a sudden decrease occurs in the membership value of the one cluster, whereas the membership value of the other clusters is hardly affected. However, for a large process change, both the membership value in the one cluster and the membership value in the other clusters gradually decrease together, and they cannot be classified to any specific cluster. Hence, the membership values are significantly correlated during a large process change.

On the basis of the above difference, the ADM is proposed as the minimum value of the entries in the adaptive membership function matrix $\mathbf{u}_k\mathbf{u}_k^t$, $\mathbf{u}_k = [u_{1k}u_{2k}\ldots u_{Ck}]^t$.

$$\mathrm{ADM}_k := \arg\ \min\ \mathbf{u}_k\mathbf{u}_k^{\mathrm{T}} \quad (25)$$

For two cluster cases, it means the inner product matrix for two membership values, u_{ik} and u_{jk}, where each element becomes each inner product (u_{ik}, u_{jk}) between the adaptive membership values (u_{ik} and u_{jk}) for clusters i and j, respectively. It has a similar mathematical form as the covariance matrix. With this definition, a gradual increase in the ADM means that the process has undergone a large operating condition change. Otherwise, the detected disturbance is due to a short disturbance.

2.5. TSK fuzzy modeling method

A fuzzy inference system is an effective means of creating models based on human expertise in a specific application by a selection of fuzzy IF-THEN rules, which form the key components of the system. Having selected the IF-THEN rules,

fuzzy set theory provides a systematic calculus to deal with information linguistically, and it performs numerical computation by using linguistic labels stipulated by membership functions. The fuzzy inference system has the properties of a structured knowledge representation in the form of fuzzy IF-THEN rules. Therefore, this system provides a good framework for applying human expertise in the construction of inference models.

The fuzzy inference system proposed by Takagi, Sugeno and Kang, known as the TSK model in fuzzy system literature (Jang et al., 1997; Bang et al., 2002; Yoo and Lee, 2001), provides a powerful tool for modeling complex nonlinear systems. It decomposes the input space into fuzzy regions and approximates the system in every region by a simple model. The overall fuzzy model is thus considered as a combination of interconnected subsystems with simpler models. Typically, a TSK model consists of IF-THEN rules that have the form

$$R_i\text{: if } x_1 \text{ is } A_{i1} \text{ and} \ldots \text{ and } x_r \text{ is } A_{ir} \text{ then } y_i = b_0 + b_1 x_1 + \cdots + b_r x_r$$

$$\text{for } i = 1, 2, \ldots, L \tag{26}$$

where L is the number of rules, $\mathbf{x}_i = [x_1 \quad x_2 \quad \cdots \quad x_r]^{\mathrm{T}}$ are input variables, y_i are local output variables, A_{ij} are fuzzy sets that are characterized by the membership function $A_{ij}(x_j)$, and $\mathbf{b}_i = [b_{i0} \quad b_{i1} \quad \cdots \quad b_{ir}]^{\mathrm{T}}$ are real-valued parameters. In general, Gaussian-type membership functions are used to build the model. They are defined by

$$A_{ir}(x_r) = \exp\left(-\frac{(x_r - c_{ir})^2}{2\sigma_i^2}\right), \quad i = 1, 2, \ldots, L \tag{27}$$

where c_{ir} is the center of the ith Gaussian membership function of the rth input variable x_r and σ_i is the width of the membership function. The overall output of the model is computed by aggregating the individual rules' contributions.

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

$$= \sum_{i=1}^{L} w_i (b_{i0} + b_{i1} x_1 + \cdots + b_{ir} x_r) \tag{28}$$

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

$$\tau_i = A_{i1}(x_1) \times A_{i2}(x_2) \times \cdots \times A_{ir}(x_r) \tag{29}$$

and $w_i = \tau_i / \Sigma_{i=1}^{L} \tau_i$ is the normalized firing strength of the ith rule. Fig. 1 shows a schematic block diagram of the TSK fuzzy model. Since each rule has a crisp output in the TSK model, the overall output is obtained via a weighted average.

The great advantage of the TSK fuzzy model is its descriptive 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 of Eq. (28) and the individual rules give insights into the local behavior of the model. The good interpretability of the fuzzy system may match the utility of the PCA method in intuitive data analysis.

2.6. Fuzzy principal component regression method

Since in practice data are usually nonlinear in most processes, it is desirable to have a nonlinear modeling approach which can represent any nonlinear relationship and still attain the robust prediction property of the PCA approach. For such a nonlinear modeling method, we propose the FPCR method which is basically an integration of PCA preprocessing and a TSK fuzzy model. First, we use PCA methods to reduce the dimensionality of the data and to remove the collinearity. The original variables are replaced by PCA score values that have better properties (orthogonality) and also span the multidimensional space of the original variables. The inverse of the PCA scores should give no problems because of the mutual orthogonality of the PCA scores. Then, the TSK fuzzy model is used to regress the PCA scores on the output variables. It captures the nonlinearity

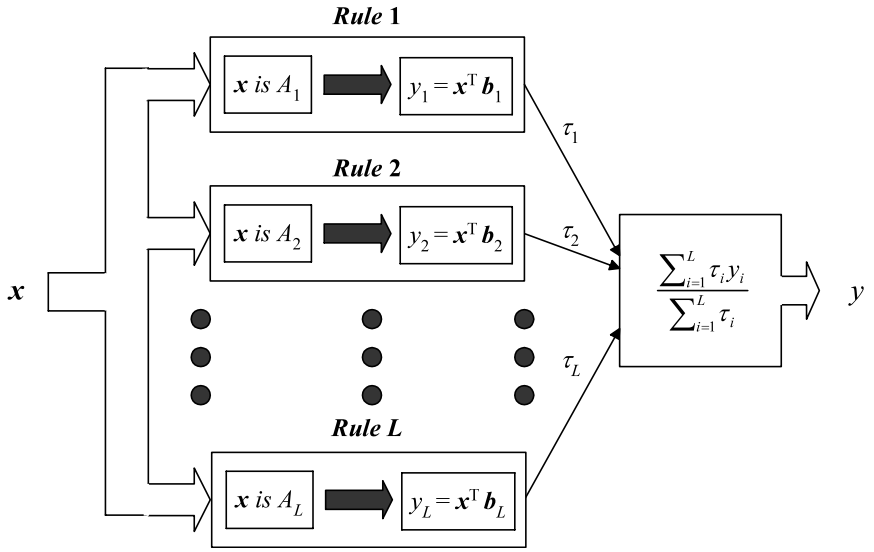

Fig. 1. Block diagram of the TSK fuzzy model.

and multi-model property of the biological treatment system.

Like the original TSK fuzzy model, the overall output of the FPCR model with score values (t_i) of the PCA model is computed by aggregating the individual rules' contributions:

$$y = \frac{\sum_{i=1}^{L} \tau_i y_i}{\sum_{i=1}^{L} \tau_i} = \frac{\sum_{i=1}^{L} \tau_i (b_{i0} + b_{i1} t_1 + \cdots + b_{ir} t_r)}{\sum_{i=1}^{L} \tau_i}$$

$$= \sum_{i=1}^{L} w_i (b_{i0} + b_{i1} t_1 + \cdots + b_{ir} t_r) \quad (30)$$

where y is the output variable, L is the number of rules, r is the number of PCs, and t_i are the score values of the PCA model. Since the TSK fuzzy model needs to construct linear regression models in different partitions, we should design the parameters of the membership function and the rule base of these partitions in the input space. The parameters of a fuzzy membership function can be determined by various heuristics and it can be a topic of research by itself. However, because it is difficult to manually design the fuzzy system, a clustering method can be utilized to automatically carry out this task. For simplicity, we use simple heuristic rules in this paper. Fuzzy rule bases are determined using the input space partition method with the CFCM clustering algorithm to identify natural grouping behavior of input data. The c_i, σ_i and $\mathbf{b}_i$ values are determined using the CFCM clustering algorithm, the nearest neighbor heuristic rule suggested by Moody and Darken's (1989) and a global learning procedure (see Appendix A).

The FPCR method differs from the direct TSK fuzzy modeling approach in that the data are not used directly to train the TSK model but rather the score values of the PCA are used as feature vector. This transformation decomposes the multivariate regression problem into a few univariate regression problems and simplifies the TSK model. The TSK method is a type of kernel regression method, where the input variables are transformed nonlinearly to feature space variables and then the transformed data set is regressed linearly. Well-designed nonlinear transformation procedures usually reduce the collinearity problem. In the kernel regression method, the method of nonlinear transformation is directly related to the regression performance. However, designing an optimal nonlinear transformation for high-dimensional and collinear data sets is very difficult, and the resulting models often suffer from over-fitting or local minima. However, the robust data reduction

characteristic of the PCA method can compensate for this problem in the TSK fuzzy modeling method. Moreover, the interpretation based on fuzzy rules can give a new way to monitor nonlinear systems. For example, each sample of a system modeled by FPCR can be classified according to the fuzzy rule that has the largest firing strength value on it.

Fig. 2 shows the framework of the proposed FPCR nonlinear modeling and monitoring method. First, a PCA model is constructed with normal historical data as a dimension reduction tool in order to remove the high-dimensional, collinear and over-fitting problem of the original TSK fuzzy model. Second, the transformed data in the reduced dimension spanned by the transformation vectors in PCA are classified into several clusters using an adaptive fuzzy c-means clustering to trace a successive change of states between several operating conditions, and are monitored using an ADM to distinguish between a long process change and short disturbances. Third, the TSK fuzzy model predicts the key output variable, which is able to capture the nonlinearity and multi-model characteristics of the process.

3. Results and discussion

Advanced modeling and monitoring strategies for WWTP have recently attracted much interest as a consequence of the increasing stringency of environmental regulations. However, most changes in biological processes are slow and recovery from failures can be time-consuming and expensive. For example, it can take several weeks or even months for the process to recover from an abnormal operation. Therefore, advanced modeling and monitoring methodologies are especially important for this process. The proposed method is applied to modeling and monitoring of the process data obtained from a full-scale wastewater treatment plant in Korea.

3.1. Process data

Process data were collected from a biological wastewater treatment plant which treats cokes wastewater from an iron and steel making factory. The data contained daily average values measured between 1 January 1998 and 9 November 2000, with a total of 1034 samples. This treatment plant uses an activated sludge process that has five aeration basins (each of size 900 m^3) and a secondary clarifier (1200 m^3). The plant layout is shown in Fig. 3. The treatment plant has two influents: wastewater arrives either directly from a cokes making plant (called BET3) or as pretreated wastewater from an upstream WWTP at another cokes making plant (called BET2). The cokes-oven plant wastewater is produced during the conversion of coal to cokes. 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

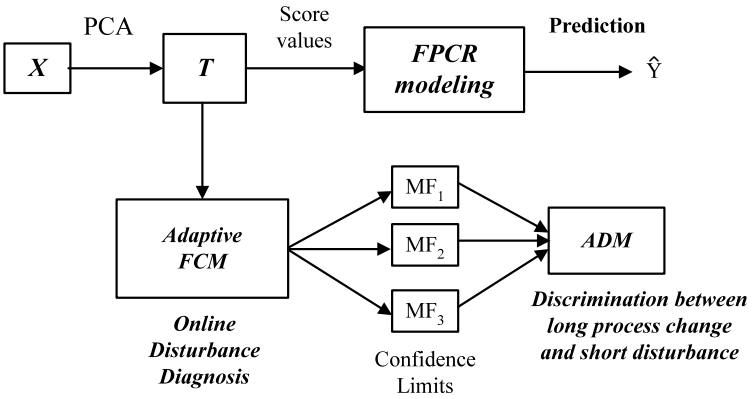

Fig. 2. Schematic diagram of the FPCR nonlinear modelling and monitoring method (for a detailed discussion, see text).

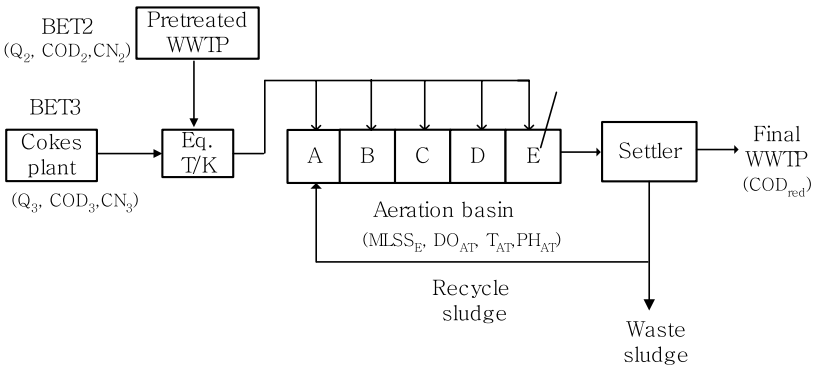

Fig. 3. Plant layout of a cokes wastewater treatment process in Korea.

that contain e.g., phenolics, thiocyanate, cyanides, poly-hydrocarbons and ammonium.

Eleven measured variables which are the **X** blocks in PCA, were used to model the process output variables, the removed COD, that is, COD_{in}–COD_{out}. Table 1 describes the process variables and presents the mean and standard deviation (S.D.) values of the input and output variables. The process data consisted of daily mean values with a total number of 948 observations, where 86 observations were eliminated due to measurement errors, based on discussion with the operators. The first 635 observations were used for the training of the PCA model with the mean-centered and auto-scaled data that were prefiltered with a median filter. The remaining 314 observations were used as a test data set in order to verify the proposed method.

3.2. Process analysis using the PCA method

When designing the PCA model, it is important to determine the number of PCs of the PCA model. It should be determined considering both the curse of dimensionality and the loss of information. Several techniques exist for determining the number of PCs, but there appears to be no dominant technique (Chiang et al., 2001). Four PCs were found adequate based on the cross-validation of the prediction residual sum of squares (PRESS). It managed to capture about 56% of the input variance by projecting the variables from dimension 12 to dimension 4. The results of the PCA model are given in Table 2.

For the interpretation of the WWTP data, we consider the PCA loading weights to see how **X** variables are interrelated. Fig. 4 shows that

Table 1
Process variables in coke WWTP

No	Variable	Description	Unit	Mean	S.D.
X_1	Q_2	Flow rate from BET2	$m^3\,h^{-1}$	178	15.3
X_2	Q_3	Flow rate from BET3	$m^3\,h^{-1}$	84.8	8.1
X_3	CN_2	Cyanide from BET2	$mg\,l^{-1}$	2.5	0.35
X_4	CN_3	Cyanide from BET3	$mg\,l^{-1}$	14.9	1.768
X_5	COD_2	COD from BET2	$mg\,l^{-1}$	157.8	19.88
X_6	COD_3	COD from BET3	$mg\,l^{-1}$	2088	306.9
X_7	MLSS_%E	MLVSS at final aeration basin	$mg\,l^{-1}$	1547	292.8
X_8	DO_{AT}	DO at final aeration basin	$mg\,l^{-1}$	1.99	0.98
X_9	$T_{influent}$	Influent temperature	°C	37.43	2.353
X_{10}	T_{AT}	Temperature at final aerator	°C	30.9	2.288
X_{11}	pH_{AT}	pH at final aeration basin	$mg\,l^{-1}$	7.225	0.23
Y	COD_{red}	COD reduction	$mg\,l^{-1}$	605.4	97

Table 2
Percent variance explained by PCA model with 4 PCs

	Eigenvalue	**X** blocks (cumulative)
PC 1	2.27	20.6
PC 2	1.50	34.3
PC 3	1.35	46.6
PC 4	1.21	57.6

specific **X** variables load strongly in the first two PCs dimension. It confirms that the PCA method distinguishes the chemical and biological variables, which occupy the different regions of the plot and exhibit a well-defined pattern. Eleven input variables fall into three groups. The first group is connected with the influent substrate and environmental conditions (COD_2, COD_3, $T_{Aerator}$, $pH_{Aerator}$). It demonstrates that the COD removal rate is strongly correlated with the COD load and the temperature in the aerators. This corresponds to the fact that heterotrophic biomass activity is influenced by the temperature in the biological treatment. These variables are uncontrolled or partially controlled throughout the process and therefore exhibit large variations. The second group is related to the microorganism concentration and reaction kinetic component (MLSS_#E and $DO_{Aerator}$) which are rate related component ($DO_{Aerator}$) in the biological reaction rate, such as in the Monod equation. This grouping indicates that the DO concentration in the aerator and temperature in the influent is strongly related to the amount of microorganisms in the total system (aerator and settler). This suggests that the DO concentration in the aeration tank should be controlled. The third group of CN_2, CN_3 and Q_2 in the upper region of the loading plot is associated with a reaction inhibition effect. Cyanides are toxic to heterotrophic bacteria and inhibitory to the reaction rate. Hence, the cyanide load is counter-correlated with the heterotrophic organisms concentration (MLSS_#E). This relationship manifests itself in opposing directions of the first and second clusters in the loading plot. Hence, shock loading of cyanides in the wastewater influent causes a deterioration of the biological treatment process. The adverse effects of cyanides have been well established in previous experimental studies (Lee et al., 2002; Yoo et al., 2002b).

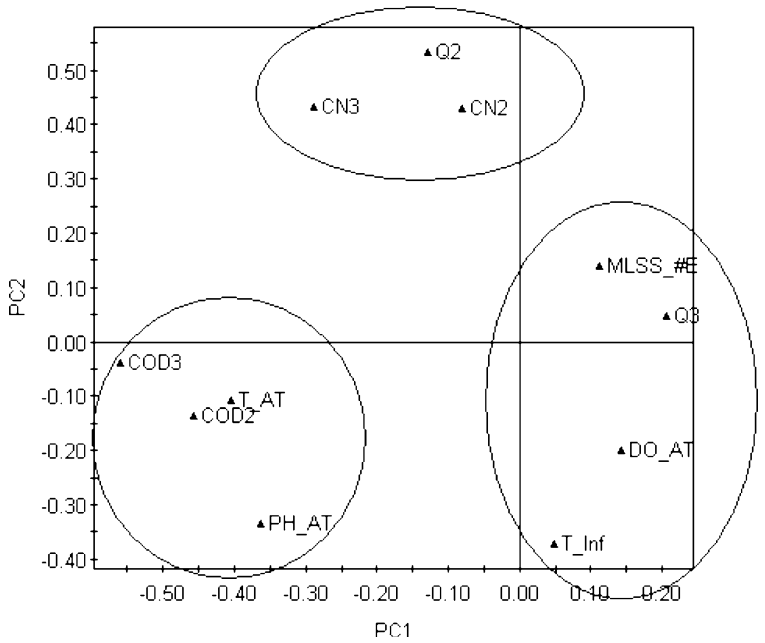

Fig. 4. Loading plot of variables of PCA model.

3.3. Statistical process monitoring

Fig. 5 describes the Hotelling's T^2 chart and *SPE* plot of the PCA analysis, where the 95% confidence limit is used. During samples 90–120, the T^2 and SPE values exceed the confidence limit, detecting a fault. We can infer a certain process change interval from Fig. 5. To identify the cause for the deviation, the contributions from every measurement variable can be calculated. Fig. 6 shows the contribution plot at sample 212, where the WWTP received a high input of cyanide, a high COD load and at high temperature in the aerator. It reduced the activity of the microorganisms. The second event, from sample 210 to the last sample of the test dataset is due to a change of the operating strategy. This results in a change of data structure in this period. After sample 200, the SPE chart fluctuates above and below the confidence limit. The operators confirmed that this fluctuation of the *SPE* chart originated from the new operating condition, in which the microorganisms had adapted to the disturbance and the WWTP is operating normally under the new conditions. That is, because of the microorganism's adaptation ability and the control actions applied in the WWTP that would bring the system to a new steady state after process changes or disturbances occurred, one could assume that a relation with similar variance but different mean could approximately hold. However, process monitoring based on a single PCA model is inadequate because not all of the measurements are static, which means that the assumptions of data normality and absence of time-correlation are not satisfied. As an alternative approach, the membership values of each cluster for the multiple operating conditions and an adaptive scheme to deal with the microorganism's adaptation will be tested.

In CFCM, appropriate values must be determined for several parameters. The number of partitions C was determined to be 4, the fuzzifier m was set to 1.5 and the adaptation constant β was set to 0.01 on the basis of the results of many simulations under various conditions. In addition, γ of Eq. (21) was set to a value of 0.9, although γ could vary over the range of 0.1–0.99 without significantly affecting the clustering results. A suitable initial cluster prototype matrix must be determined for fast convergence of the algorithm.

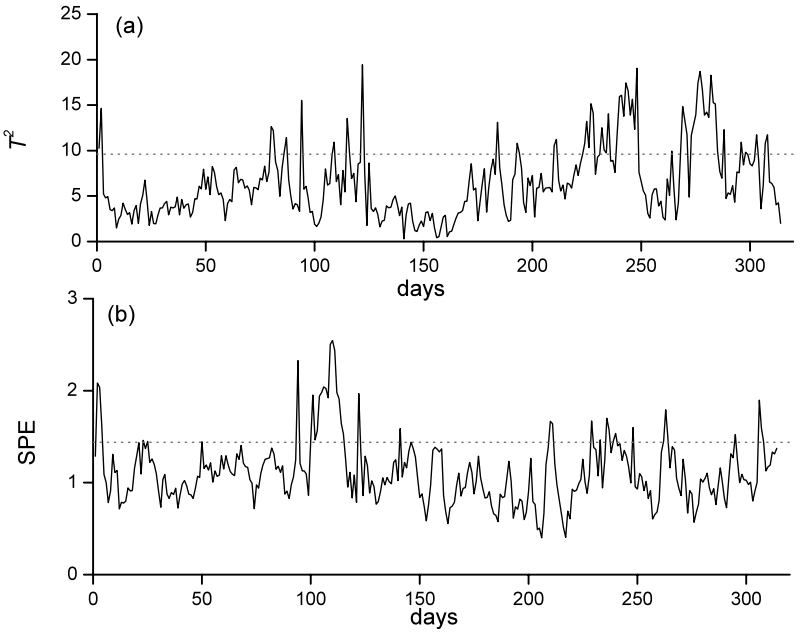

Fig. 5. Monitoring performance based on T^2 and *SPE* charts of the PCA model with 95% confidence limits.

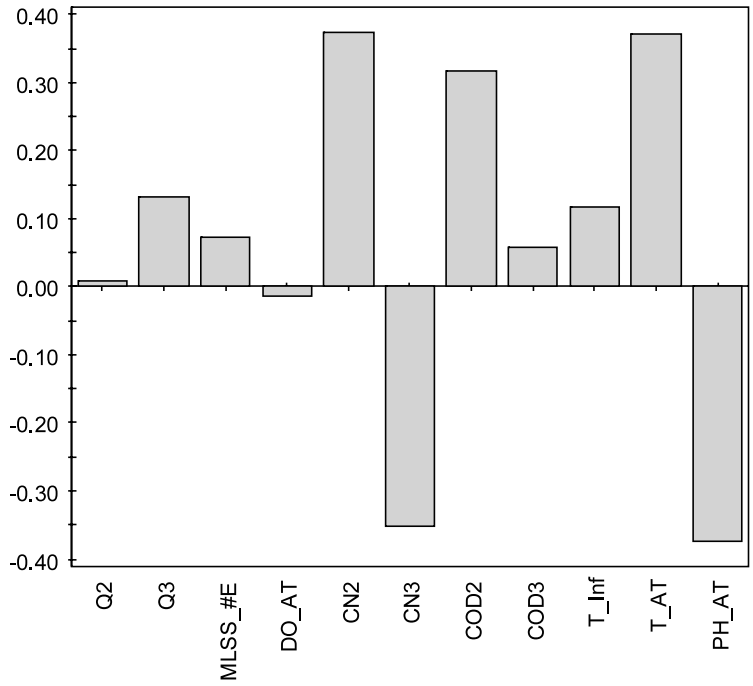

Fig. 6. Contribution plots of the PCA score values at sample 230.

Following the initialization scheme proposed by Chintalapudi and Kam (1998), we initialized the parameters of σ, Ψ_k and γ which represented their corresponding clusters mixed with some noise as the initial centroid vectors. For process monitoring, each membership value for every local cluster from ACFCM was monitored. Additionally, the ADM's were used to detect the process drift or shift and to discriminate such process changes from instantaneous disturbances.

The score values of the independent data set **X** are displayed in the reduced dimension spanned by the first four loading vectors, as shown in Fig. 7. Most observations fall into four small regions, where each region represents a distinct operating condition. The centers of the cross signs in Fig. 7 indicate cluster centroids (also called cluster prototypes). Four centroids represent the seasonal variations with successively the summer, spring, autumn and winter from the left side. This is, of course, due to the seasonal variations in the mean value. The process indeed has four local process operating conditions which change periodically according to the seasonal variations. The convergence toward the cluster prototypes is shown in Fig. 8. The circle symbols in this figure indicate the locations of cluster centers in the model subspace. All prototypes converge to their optimal points within 30 iterations. If the initial values of the cluster centers are selected reasonably, convergence is very fast and the prototypes converge.

Evaluating Fig. 9 in which the membership values of all samples are plotted against sample number can be more relevant than comparing several T^2 charts for the separate class models. Membership values of each cluster show periodic seasonal variations in a similar manner. The dotted lines represent the confidence limit using Students' t-test in each cluster. Here, the value of α, 0.05, is used to find the $(1-\alpha)100\%$ confidence interval of each membership value for the warning limit. According to this approach, a sample belongs to a cluster if its membership value exceeds the confidence limit for the cluster.

The recursive updating of the cluster prototypes by CFCM is shown in Fig. 10. Cluster prototypes of CFCM are recursively updated with the flexible updating rule, Eq. (23). Also, the projection of the

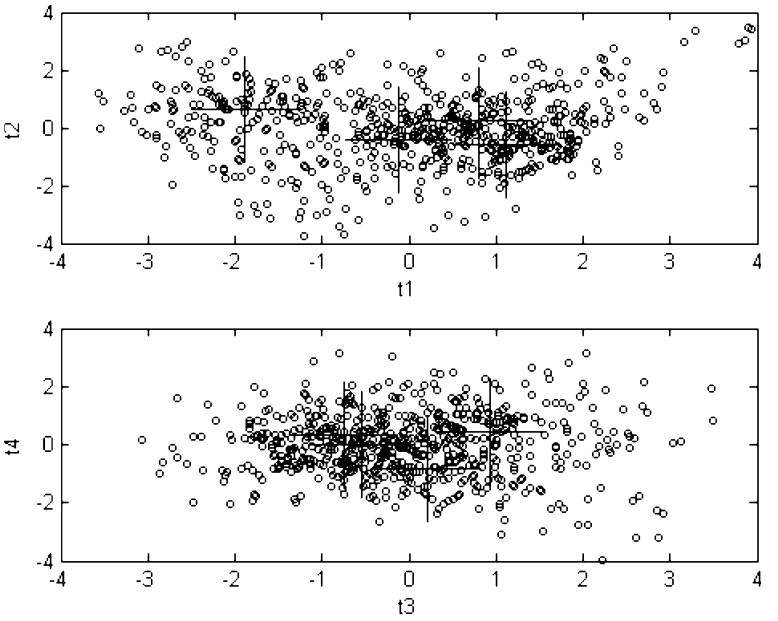

Fig. 7. CFCM clustering results in the reduced dimension spanned by the four loading vectors. The cross point of each + sign represents the corresponding cluster prototype.

test data into the model space reveals that large process changes and drifting occurred during the test. Fig. 11 shows the membership values of the test data using the ACFCM clustering. Inspection of Fig. 11 reveals that under the above classification scheme some samples do not belong to any cluster. These samples correspond to an operating mode change. However, each cluster prototype is affected very little by samples in such new condition because samples in a new operating region are

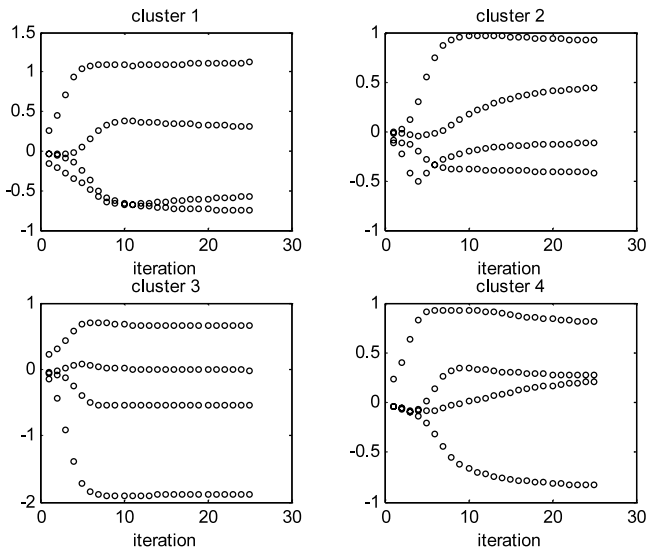

Fig. 8. Convergence of four cluster prototypes. (a) cluster 1, (b) cluster 2, (c) cluster 3, (d) cluster 4. Circle symbols represent the location of a cluster center in the reduced feature dimension.

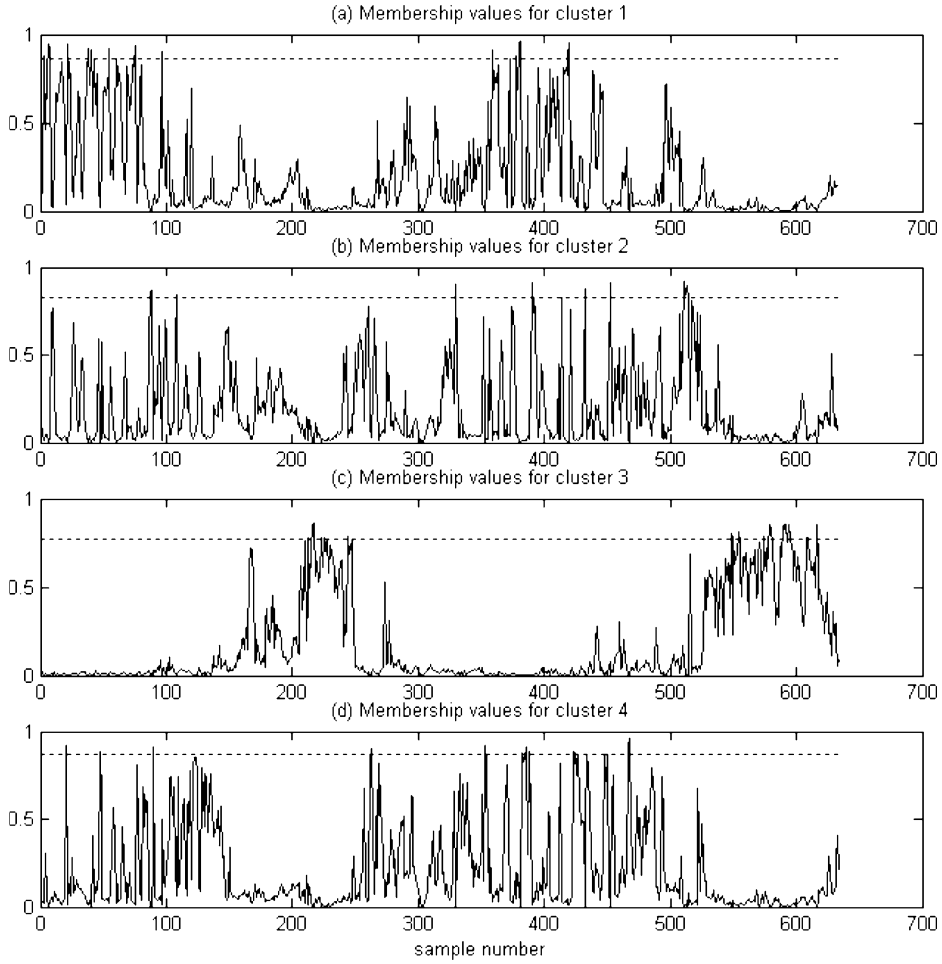

Fig. 9. Membership values of training data resulting from CFCM clustering. (a) cluster 1, (b) cluster 2, (c) cluster 3, (d) cluster 4.

assigned low credibility and, hence, have low membership values to both neighboring clusters. This confirms the ability of the CFCM clustering method. In general, FCM provides only inter-cluster information, whereas PCM provides only intra-cluster information. As compromising concept of these two different clustering algorithms, the membership values in the CFCM provide not only inter- but also intra-cluster information.

As mentioned above, the ADM can help to solve an important shortcoming, that is, the discrimination between an operating condition change and an instantaneous disturbance. The static discrimination measure (SDM) and the ADM of the membership values for the four clusters in the test data are shown in Fig. 12(a) and (b), where the SDM is composed of the membership functions computed using the non-adaptive CFCM clustering. As shown in Fig. 12(b), the ADM shows peaks at samples 100 and 120 and another peak at samples 220 and 250, indicating that process changes occurred at these times. At samples 100 and 120, the process received influents with a high cyanide and COD content and a small influent flow rate, i.e., a highly concentrated load. This influent reduced the activity of the microorganisms. These variations of the microorganism characteristics which were caused by the influent load, led biological process to a gradual operating change. In addition, we

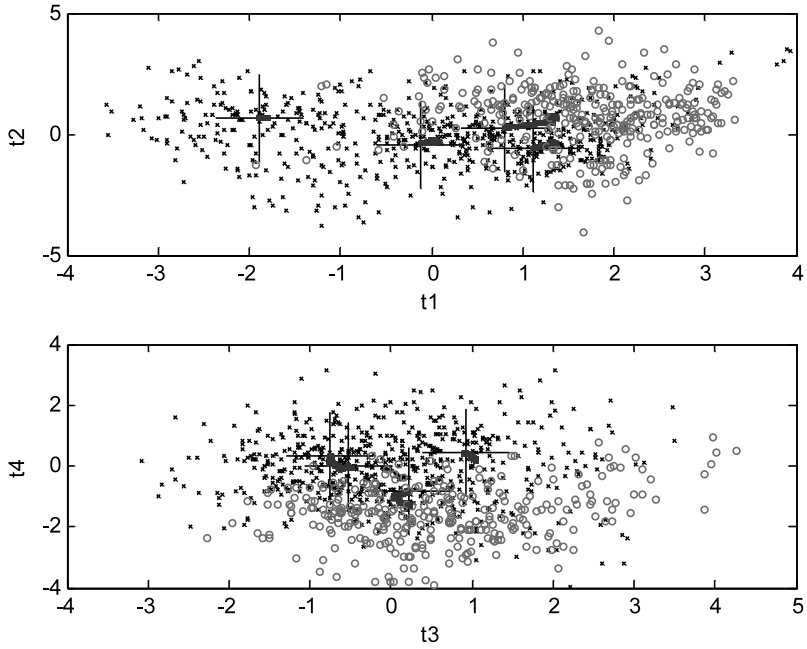

Fig. 10. Adaptation of cluster prototypes of CFCM with adaptive updating: Training data (×), test data (○), trajectory of cluster prototype (+).

found that an external disturbance, the large influent load, was transformed into an internal disturbance that changed the process operating region of the activated sludge process. Also, membership values of cluster 4 in Fig. 11 during this period show large changes indicating that a process change occurred in this interval. The ADM values show another deviation from sample 220 to 250. At this time, the wastewater treatment plant was modified, with the addition of facilities and treatment equipment. These changes made it feasible for operators to adjust the operation policy, i.e., they increased the MLSS concentration and maintained the DO concentration at a high value. The modifications to the plant caused substantial changes to the process and affected all process conditions after this time. This manifests itself as a variation of the ADM in Fig. 12(b). After sample 250, the ADM can describe the process change well by the recursive update of the clusters, whereas the SDM in Fig. 12(a) cannot catch these adaptive characteristics. This result confirms that the proposed method is efficient for adaptive process changes of unknown characteristics. Thus, the proposed method could easily distinguish normal process shifts from abnormal deviations or process faults. It is therefore an effective technique for extracting information related to changes in process operating conditions and can also be used to localize several process disturbances. On the other hand, other variations in the local clusters such as an instantaneous disturbance did not affect the ADM.

3.4. FPCR modeling

In this work, a FPCR predictor was designed which has four inputs and one output variable, where the input consists of the four score values of the PCA method and one output variable which is the reduction of COD in the wastewater treatment plant. Considering this parsimonious model and avoiding over-fitting, we aimed for a small rule base and selected $L = 4$ as the number of rules by clustering on the PCA score plane. This can also be changed by a visual check on the score plot by experts. It comes from the fact that the number of fuzzy partitions determines the number of fuzzy

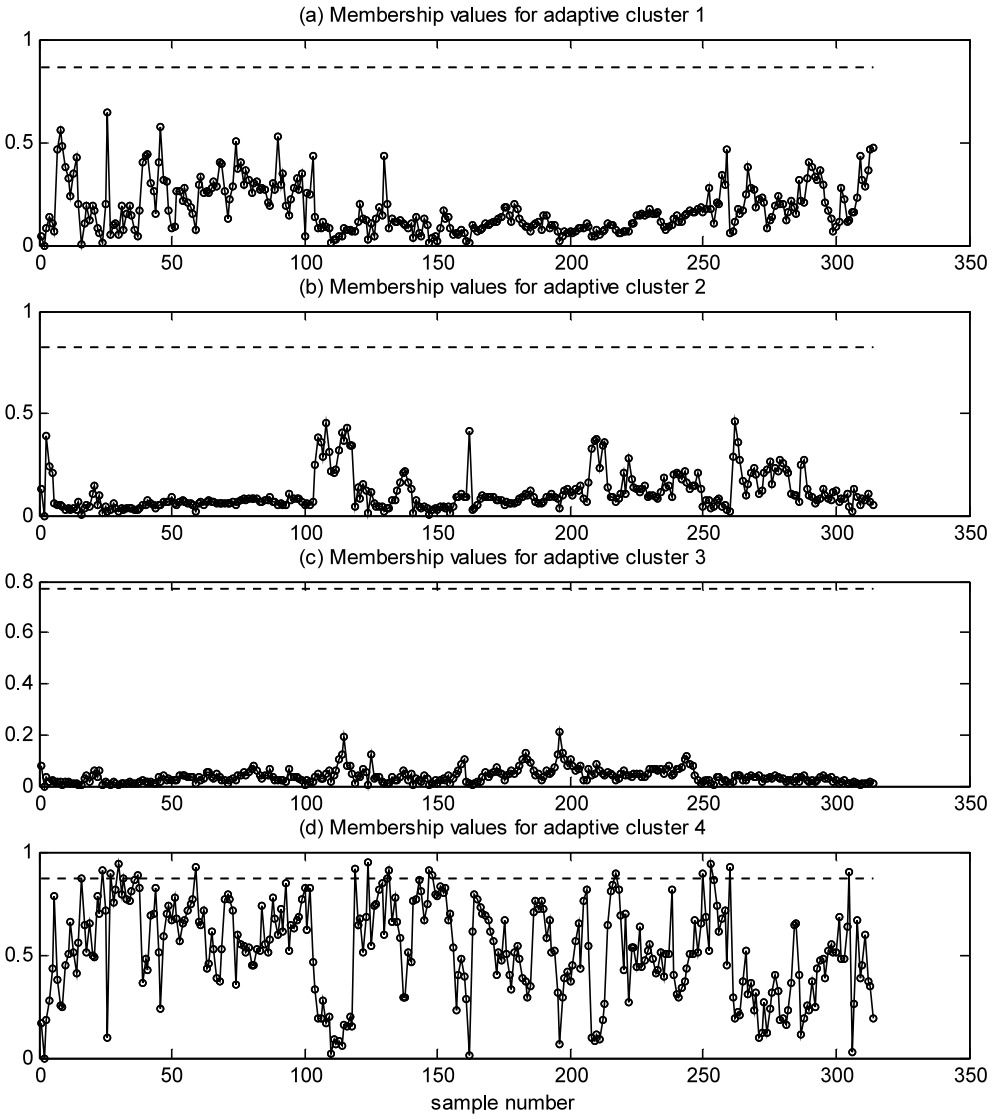

Fig. 11. Membership values of test data resulting with ACFCM clustering. (a) cluster 1, (b) cluster 2, (c) cluster 3, (d) cluster 4.

rules constituting the underlying TSK fuzzy model. Visual checking is possible because of the robust data reduction by the PCA method and the 2D presentation properties of the PCA method. Other PCR methods such as linear PCR and nonlinear PCR 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 which allows us to make a simple linear prediction of its behavior even in the extrapolation range.

Fig. 13 shows plots of the three membership functions of the TSK fuzzy model and Table 3 represents the result of the fuzzy rule sets using the CFCM clustering method, where both MF parameters and corresponding fuzzy rule bases are determined and linear parameters of the consequent part are determined by the generalized least square method. The membership function and the

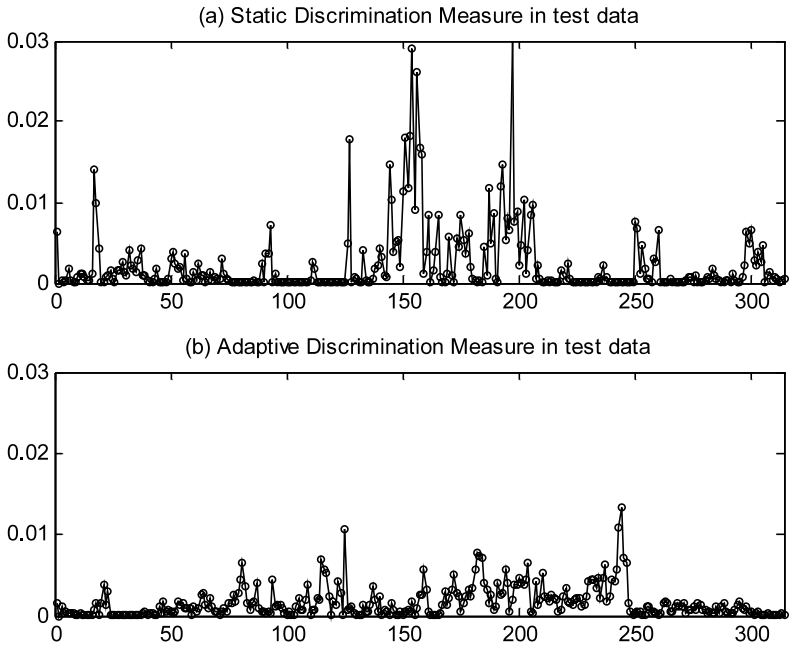

Fig. 12. Static and ADM of membership values for four clusters in the test data: (a) SDM, (b) ADM.

rule bases are determined using the input space partition method with the CFCM clustering algorithm identifying natural grouping behavior of the input data. Figs. 14 and 15 present the prediction results and the scatter plot of the PCR and FPCR model for the reduction of COD in the training data set and test data set, respectively. Fig. 16 shows the prediction results and scatter plot of the simple TSK fuzzy model using all eleven variables to predict the reduction of COD in the test data set. As expected, the FPCR shows a better prediction ability for the COD reduction than the simple TSK fuzzy model and the PCR method. We can infer from the scatter plot that the FPCR method is capable of reducing the variability of the scatter plot in the training data, that is, it is more normally distributed. The scatter plot shows some disparity between the predicted and the measured values in the test data. The reason of disparity in the scatter plot is due to the disturbance. As shown in the ADM, the WWTP had experienced a large change in operating conditions in the test periods. These process transitions altered the type of microorganisms and sludge, which changed the process dynamics. Because the FPCR model is designed to capture the adaptive and multiple operating conditions, the FPCR model showed superior prediction results in these disturbance events. In particular, the simple TSK fuzzy regression and the PCR model show almost the same prediction performance in the training and test data. This result confirms that the collinearity between the original eleven variables clearly exists. It can be concluded that a few PCs are enough to model the output variable in the process without using all variables and a considerable dimension reduction can be achieved by the PCA method.

To compare the prediction capability of the simple TSK fuzzy regression, the PCR and the FPCR, a root mean square error (RMSE) and nondimensional error index (NDEI) was used. The NDEI is defined as the RMSE divided by the S.D. of the target output. Table 4 lists the prediction capability results of different methods. The FPCR method shows a more accurate prediction capability and has a lower RMSE and NDEI than the simple TSK fuzzy regression and the PCR. On the other hand, it is possible to use the FCM method with more PCs or all eleven variables directly. Because there is no projected plane in the simple

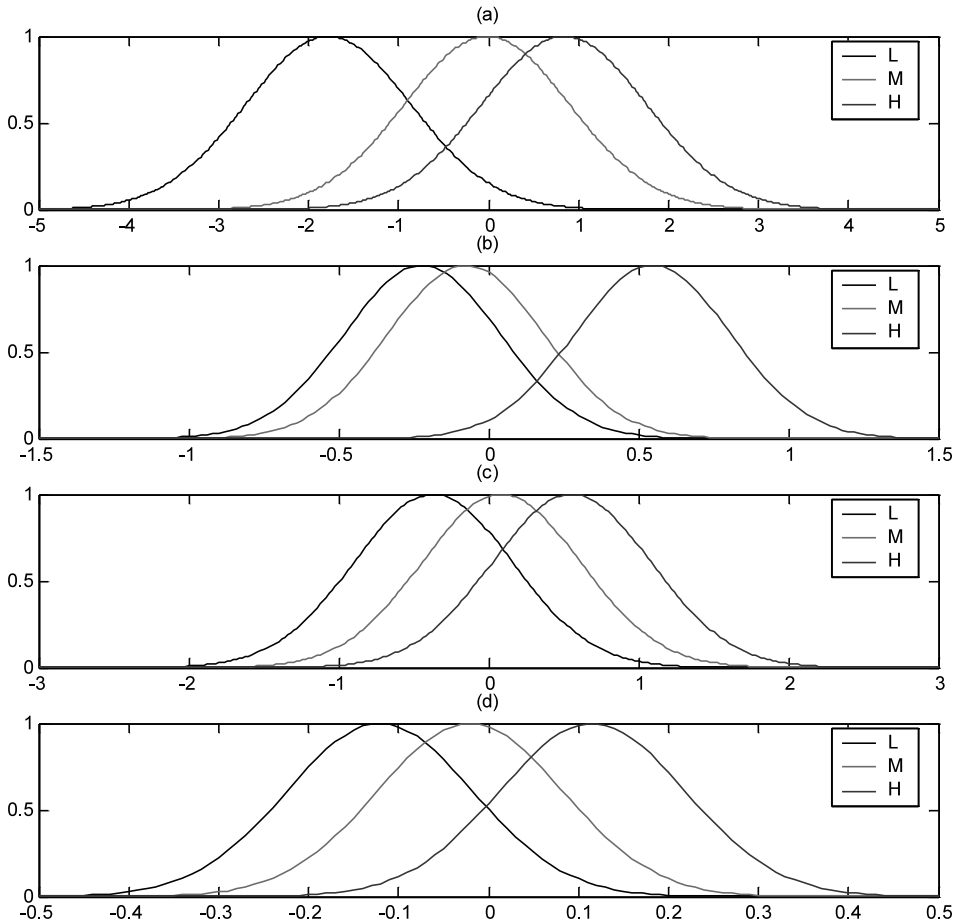

Fig. 13. Membership functions of the TSK fuzzy model.

TSK approaches using all eleven variables, it is not possible to visualize the data. Moreover, the collinearity problems between the original variables cannot be eliminated. That is, there are many variables and the collinearity between the different variables increases the inverse problem of the TSK fuzzy regression and the chance of spurious sub-optimal solutions of the FCM method (Teppola, 1999). On the other hand, with the PCA approach, the number of PCs is significantly smaller than the number of variables and these PCs are uncorrelated. Additionally, it is not necessary to include many PCs because a sufficient cluster separation is already obtained by using the first four PCs. Still it is true that an even better solution could be obtained by using more than the four PCs. Additional PCs are mainly needed to improve the prediction ability which is to be determined by using cross-validation. The use of PCA, as a data compression method, decreases the chance of getting spurious solutions.

Table 3
Fuzzy rule sets

Rule number	1	2	3	4
PC 1	L	H	M	H
PC 2	M	H	L	L
PC 3	H	H	L	L
PC 4	M	H	M	M

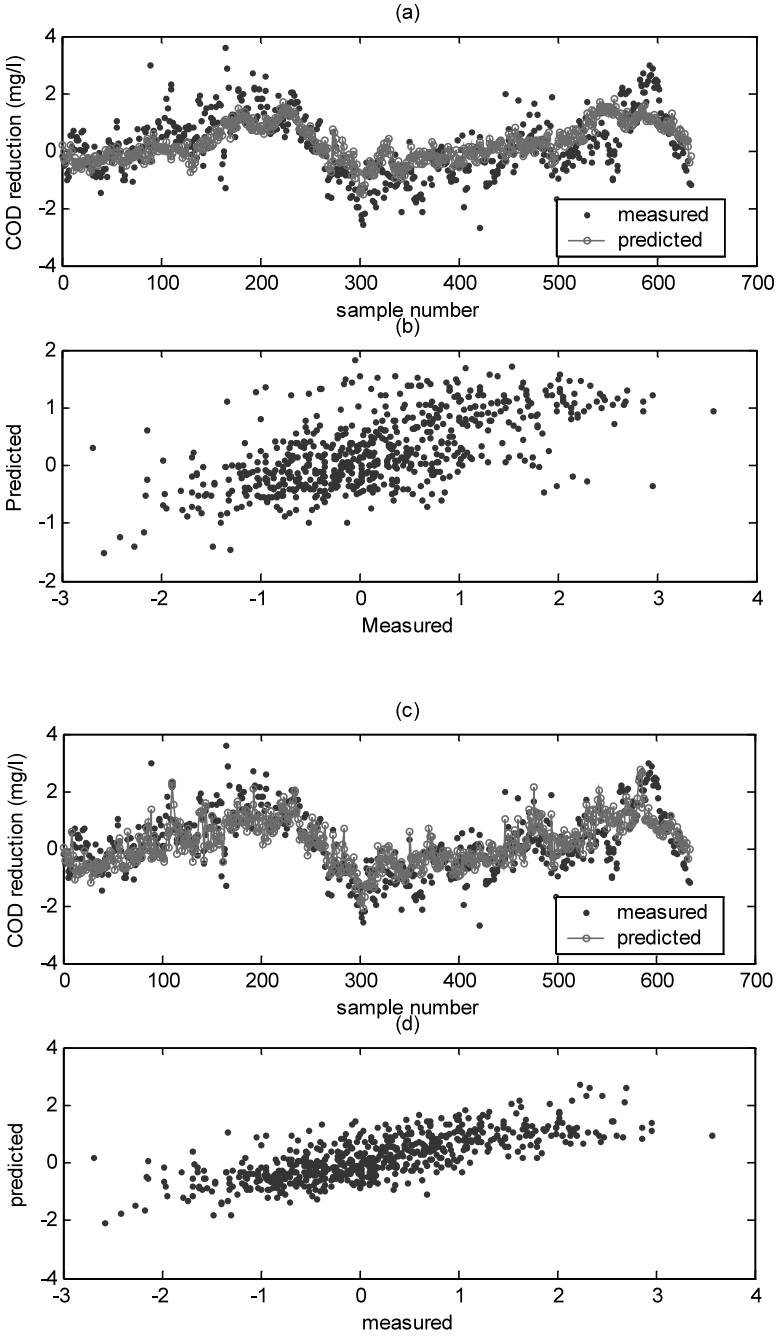

Fig. 14. Prediction results and scatter plot of the PCR and FPCR model for the reduction of COD in the training data set: (a) prediction of COD reduction by PCR; (b) scatter plot of PCR, (c) prediction of COD reduction by FPCR; (d) scatter plot of FPCR.

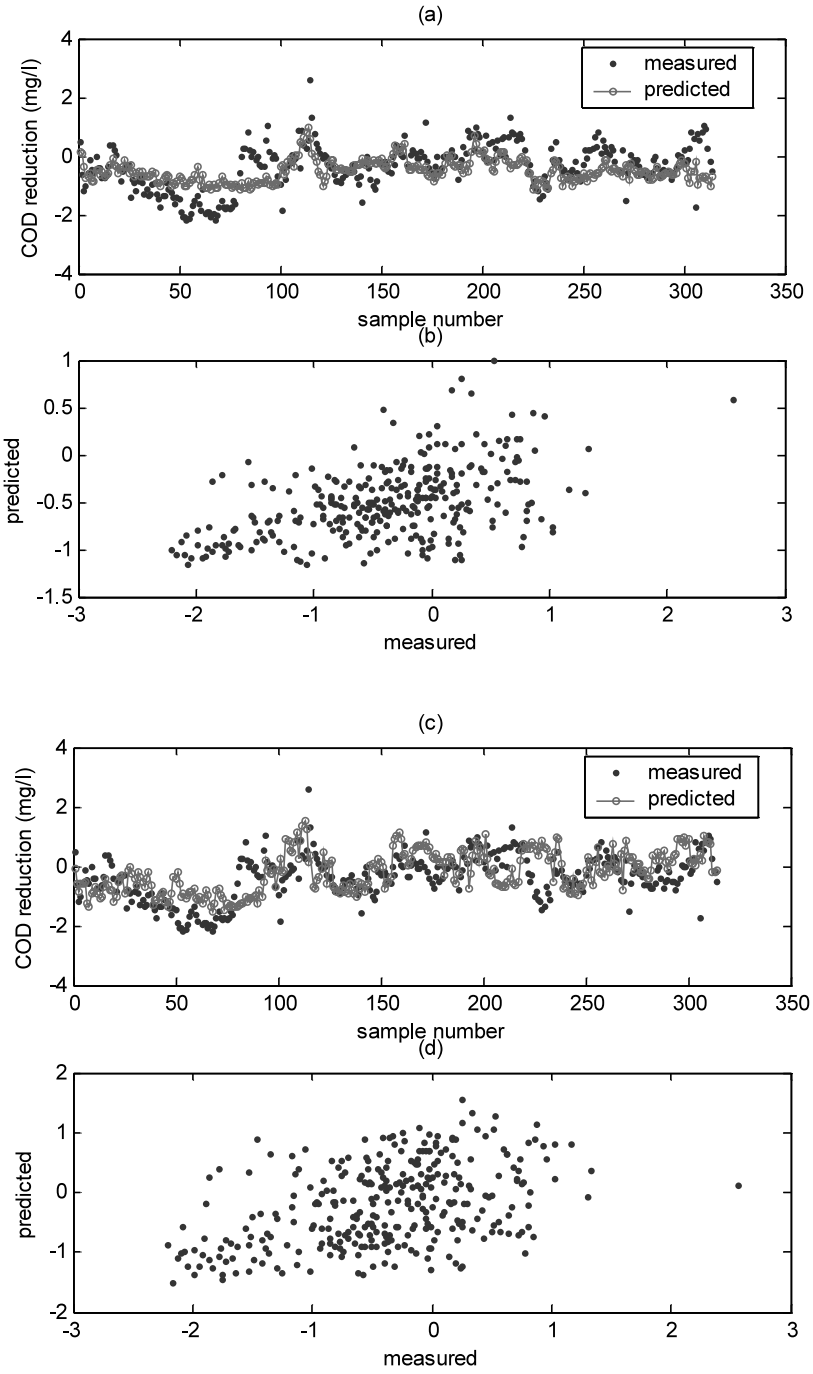

Fig. 15. Prediction results and scatter plot of the PCR and FPCR model for the reduction of COD in the test data set: (a) prediction of COD reduction by PCR; (b) scatter plot of PCR, (c) prediction of COD reduction by FPCR; (d) scatter plot of FPCR.

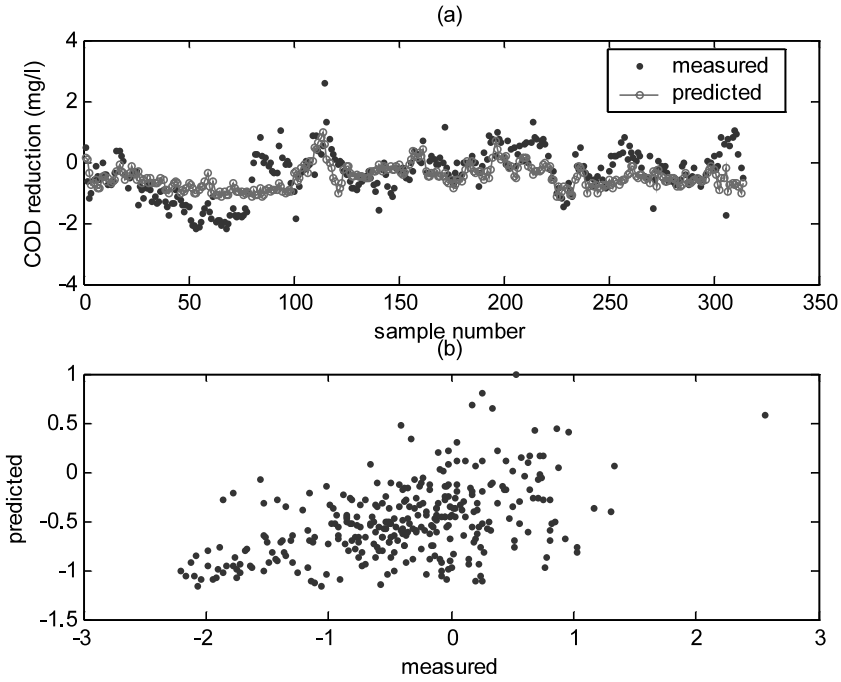

Fig. 16. Prediction results and scatter plot of the simple TSK fuzzy model for the prediction of the reduction of COD in the test data set: (a) prediction of COD reduction; (b) scatter plot.

Table 4
Comparisons of RMSE and NEDI between simple TSK fuzzy, PCR and FPCR

	Training data		Test data	
	RMSE	NEDI	RMSE	NEDI
Simple TSK fuzzy	0.8167	0.7784	0.6764	0.8855
PCR	0.8167	0.7784	0.6764	0.8855
FPCR	0.6933	0.6608	0.5895	0.7812

4. Conclusion

In this paper, nonlinear modeling using FPCR and a disturbance diagnosis method using an ADM and an ACFCM clustering are proposed. First, PCA is used to reduce the dimensionality of the data and to remove collinearity. Second, the adaptive credibilistic fuzzy-c-means method is used to model diverse kinds of operating conditions based on a recursive update of clusters. Subsequently, a new adaptive discrimination monitoring (ADM) method is proposed to distinguish between a large process change and a simple fault or a short disturbance. Third, FPCR which models the relation between the score value of PCA and the target output is suggested to avoid the overfitting problem with the original variables. The proposed FPCR model not only possesses a nonlinear modeling ability, but also the robustness and interpretability of the PCA and fuzzy methods. The case study clearly showed that the proposed method gave good modeling performance and reasonable monitoring results. It was able to identify various operating regions and discriminate between a sustained disturbance and a short disturbance.

Acknowledgements

This work was supported by the Brain Korea 21 project and the Post-doctoral Fellowship Program of Korea Science & Engineering Foundation (KOSEF). This research was also financially supported by the Fund for Scientific Research-

Flanders (F.W.O.) and the Ghent University Research Fund.

Appendix A: Identification of structure and parameters of FPCR model

(1) The center of a TSK fuzzy model (FCM)

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

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

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) Moody and Darken' rule (1989)

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

$$\sigma_i = \left[\frac{1}{p}\sum_{l=1}^{p}(c_i - c_l)^2\right]^{1/2}, \quad i = 1, 2, \ldots, L \tag{A2}$$

where c_l ($l = 1, 2, \ldots, p$) 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 Eq. (A2) over all L centers.

(3) Global learning algorithm

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.

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

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

where

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

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

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.

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

Or we can use a computationally efficient methods, such as SVD, to solve the singularity problem in computation of the inverse of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$. Applying SVD to $\mathbf{X}$ yields

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

where $\mathbf{U} = [\mathrm{u}_1 \quad \mathrm{u}_2 \quad \cdots \quad \mathrm{u}_N]^{\mathrm{T}} \in R^{N \times N}$ and $\mathbf{V} = [\mathrm{v}_1 \quad \mathrm{v}_2 \quad \cdots \quad \mathrm{v}_{2L}]^{T} \in R^{2L \times 2L}$ are orthogonal matrices, and $\mathbf{\Sigma} = \mathrm{diag}(\sigma_1, \sigma_2, \cdots, \sigma_{2L}) \in R^{N \times 2L}$ is a diagonal matrix with $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{2L} \geq 0$. Substituting Eq. (A8) into Eq. (A3) and after simple manipulations, the minimum Euclidean norm solution of the fuzzy rule parameters, b, is computed as

$$\mathbf{b} = \sum_{i=1}^{s} \frac{u_i^{\mathrm{T}} y}{\sigma_i} v_i \tag{A9}$$

where s is the number of nonzero singular values in Σ (Yen et al., 1998). There are only small differences between the LSE and SVD approaches because the FPCR method eliminates the singularity problem by PCA preprocessing.

References

Bang, Y.H., Yoo, C.K., Lee, I.-B., 2002. Nonlinear PLS modeling with fuzzy inference system. Chem. Int. Lab. Syst. 64 (2), 137–155.

Beni, G., Liu, X., 1994. A least biased fuzzy clustering method. IEEE Trans. Pat. Anal. Mach. Intel. 16 (9), 954–960.

Bezdek, J.C., 1981. Pattern Recognition with Fuzzy Objective Functions. Plenum Press, New York.

Chiang, L.H., Russell, E.L., Braatz, R.D., 2001. Fault Detection and Diagnosis in Industrial Systems. Springer, USA.

Chintalapudi, K.K., Kam, M., 1998. The credibilistic fuzzy c means clustering algorithms. IEEE Int. Conf. Syst., Man Cybern. 2, 2034–2039.

Choi, D.-J., Park, H.K., 2001. A hybrid artificial neural network as a software sensor for optimal control of a wastewater treatment process. Wat. Res. 35 (16), 3959–3967.

Choi, S.W., Yoo, C.K., Lee, I.-B., 2003. Overall statistical monitoring of static and dynamic patterns. I&EC 42 (1), 108–117.

Cote, M., Grandjean, B.P.A., Lessard, P., Thibault, J., 1995. Dynamic modeling of the activated sludge process: improving prediction using neural networks. Wat. Res. 29 (4), 995–1004.

Fu, C., Poch, M., 1998. Fuzzy model and decision of COD control for an activated sludge process. Fuzzy Sets Syst. 93, 281–292.

Johnson, R.A., Wichern, D.W., 1992. Applied Multivariate Statistical Analysis, third ed. Prentice-Hall, Englewood Cliffs, USA.

Henze, M., Grady, C.P., Gujer, W., Marais, G.v.R., Matsuo, T. 1987. Activated sludge model No. 1. (IAWPRC Scientific and Technical Report No. 1). IWA Publishing, London, United Kingdom.

Huang, Y.C., Wang, X.Z., 1999. Application of fuzzy causal network to wastewater treatment plants. Chem. Eng. Sci. 54, 2731–2738.

Jackson, J.E., 1991. A User's Guide to Principal Components. Wiley, New York, USA.

Jang, J.R., Sun, C., Mizutani, E., 1997. Neuro-Fuzzy and Soft Computing. Prentice-Hall, USA.

Joo, D.-S., Choi, D.-J., Park, H.K., 2000. The effect of data preprocessing in the determination of coagulant dosing rate. Wat. Res. 34 (13), 3295–3302.

Krishnapuram, R., Keller, J., 1993. A possibilistic approach to clustering. IEEE Trans. Fuzzy Syst. 1, 98–110.

Kurtanjek, Z., 1998. Principal component ANN for modeling and control of baker's yeast production. J. Biotechnol. 65, 23–35.

Lee, D.S., Park, J.M., 1998. Neural network modeling for on-line estimation of nutrient dynamics in sequentially-operated batch reactor. J. Biotechnol. 75, 229–238.

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. Biotech. Bioeng. 78, 125–135.

Li, W., Yue, H.H., Valle-Cervantes, S., Qin, S.J., 2000. Recursive PCA for adaptive process monitoring. J. Process Control 10, 471–486.

Marsili-Libelli, S., 1998. Adaptive fuzzy monitoring and fault detection. Int. J. Comadem. 1 (3), 31–37.

Marsili-Libelli, S., Müller, A., 1996. Adaptive fuzzy pattern in the anaerobic digestion process. Pattern Recog. 17, 651–659.

Montgomery, D.C., 2001. Introduction to Statistical Quality Control, fourth ed. Wiley, New York, USA.

Moody, J., Darken, C.J., 1989. Fast learning in networks of locally-tuned processing units. Neural Computat. 1, 281–294.

Pal, N.R., Pal, K., Bezdek, J.C., 1997. A mixed c means clustering model. IEEE Int. Conf. Fuzzy Syst., Spain, 11–21.

Ragot, J., Grapin, G., Chatellier, P., Colin, F., 2001. Modeling of a waterwater treatment plant. A multi-model representation. Environmetrics 12, 599–611.

Rosen, C. 2001. A chemometric approach to process monitoring and control with application to wastewater treatment operation. Ph.D. Thesis, Lund University, Sweden.

Rosen, C., Lennox, J.A., 2001. Multivariate and multiscale monitoring of wastewater treatment operation. Wat. Res. 35 (14), 3402–3410.

Tay, J., Zhang, X., 1999. Neural fuzzy modelling of anaerobic biological wastewater treatment systems. J. Env. Eng. 125 (12), 1149–1159.

Tay, J., Zhang, X., 2000. A fast predicting neural fuzzy model for high-rate anaerobic wastewater treatment systems. Wat. Res. 34 (11), 2849–2860.

Teppola, P., 1999. Multivariate process monitoring of sequential process data—a chemometric approach. Ph.D. Thesis, Lappeenranta University, Finland.

Van Dongen, G., Geuens, L., 1998. Multivariate time series analysis for design and operation of a biological wastewater treatment plant. Wat. Res. 32 (3), 691–700.

Wang, C.-H., Hong, T.-P., Tseng, S.-S., 1998. Integrating fuzzy knowledge by genetic algorithm. IEEE Trans. Evolut. 2 (4), 139–149.

Wise, B.M., Gallagher, N.B., 1996. The process chemometrics approach to process monitoring and fault detection. J. Process Control 6, 329–348.

Yen, J., Wang, L., Gillespie, W., 1998. Improving the interpretability of TSK fuzzy models by combining global learning and local learning. IEEE Trans. Fuzzy Syst. 6 (4), 530–537.

Yoo, C.K., Lee, I., 2001. Wastewater quality modeling by hybrid GA-Fuzzy model. Asian Waterqual 2001, 107–112, Fukuoka, Japan, September 12–15, 2001.

Yoo, C.K., Choi, S.W., Lee, I., 2002a. Disturbance detection and isolation in the activated sludge process. Wat. Sci. Tech. 45 (4–5), 217–226.

Yoo, C.K., Choi, S.W., Lee, I., 2002b. Dynamic monitoring method for multiscale fault detection and diagnosis in MSPC. I&EC 41, 4303–4317.

Yoo, C.K., Choi, S.W., Lee, I., 2002c. Adaptive modeling and classification of the secondary settling tank. Korean J. Chem. Eng. 19 (3), 377–382.