Environ. Eng. Res. Vol. 10, No. 2, pp. 88~103, 2005
Korean Society of Environmental Engineers

# MODELING AND MULTIRESOLUTION ANALYSIS IN A FULL-SCALE INDUSTRIAL PLANT

**Chang Kyoo Yoo[1,2,†], Hong-Rok Son[1] and In-Beum Lee[1]**

[1]School of Environmental Science and Engineering/Department of Chemical Engineering, Pohang University of Science and Technology, San 31 Hyoja Dong, Pohang, 790-784, Korea
[2]BIOMATH, Department of Applied Mathematics, Biometrics and Process Control, Ghent University, Coupure Links 653, B-9000 Gent, Belgium

(received November 2004, accepted April 2004)

**Abstract** : In this paper, data-driven modeling and multiresolution analysis (MRA) are applied for a full-scale wastewater treatment plant (WWTP). The proposed method is based on modeling by partial least squares (PLS) and multiscale monitoring by a generic dissimilarity measure (GDM), which is suitable for nonstationary and non-normal process monitoring such as a biological process. Case study in an industrial plant showed that the PLS model could give good modeling performance and analyze the dynamics of a complex plant and MRA was useful to detect and isolate various faults due to its multiscale nature. The proposed method enables us to show the underlying phenomena as well as to filter out unwanted and disturbing phenomena.

**Key Words** : Disturbance detection and diagnosis, multiscale modeling, multiresolution analysis (MRA), partial least squares (PLS), process monitoring, 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 plant (WWTP) are necessary to maintain the system performance as close as possible to optimal conditions. Specially, monitoring of the biological treatment process is very important because the recovery from failures is time-consuming and expensive, where some of changes are not very obvious and may grow gradually until they produce a serious operational problem. That is, most of the changes in biological treatment process are very sluggish when the process is recovered back from a 'bad' state to a 'normal' state or back from a 'bad' state to a 'good' state. Therefore, early fault detection and isolation in the biological process are very efficient to execute corrective action well before a dangerous situation happens. At the same time the discrimination between serious and minor abnormalities is of primary concern. To accomplish this task, a reliable detection procedure is needed. Several approaches have been available to utilize large on-line and off-line data sets despite of the increasing popularity and the decreasing price of on-line measurement systems in the field of WWTP.[1-11)]

The underlying point is that improving process monitoring and control necessarily means

† Corresponding author
Tel: +82-54-279-5966, Fax: +82-54-279-8299
E-mail address: ckyoo@postech.edu or
ChangKyoo.Yoo@biomath.ugent.be

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 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.[7,8]

To date, the most successful model and the industrial standard in biological WWTP has been the deterministic mechanistic model, activated sludge model (ASM) no. 1, 2, 2d, 3.[9-13] It has proven to be an effective model for carbonaceous, nitrogenous and phosphorous 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 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 have been limited.

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 process behavior and of the properties of the product or system output. Machine learning algorithms such as 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.[1,2,6,7]

A wastewater treatment plant is a quite complex system including lots of equipment and complex processes. The operators are under increasing regulatory pressure to reduce pollutant levels in their effluent. One response to this has been the installation of extensive on-line sampling capable of measuring flow rates, concentrations and other variables frequently. Data acquisition systems may collect a large amount of data, normally tens of process and control variables, but there are relatively few significant events. Therefore, the data from all the measurements should be mapped into significant description of current process. The obtained data will give much process information, if only the important and relevant information can be extracted and interpreted. Not only are there a lot of variables to be considered, but also they are often highly cross-correlated (*i.e.* the measured variables are not independent of one another) and auto-correlated. So, redundancy that variables carry the same information at least to some extent is observed. It is desirable to develop the schemes for providing reliable on-line information on the status of the plant so that early corrective actions may be taken.[1,4]

Multivariate statistical process control (MSPC) is a possible solution to the dimensionality and collinearity problems. Contrary to univariate techniques, multivariate techniques are more successful solutions to monitor the process data having severe collinearity and noise. They contain such methods as principal components analysis (PCA) or partial least squares (PLS) combined with standard sorts of control charts. These methods are the basis of the field of *chemometrics*, which has traditionally been concerned with multivariate analyses in chemistry, particularly those of spectroscopy. PCA and PLS aim to represent a multivariate set of measurements with a smaller number of the transformed variables.[1,14,15] Process monitoring system consists largely of three sequential parts: data rectification, fault detection and diagnosis. Figure 1 illustrates a process monitoring scheme for an industrial plant. Data rectification means a screening of available data to remove redundant information. Fault detection is defined as a combination of process observations and measurements, data analysis and interpretation to detect abnormal features or effects and the isolation of faults. Fault diagnosis involves the analysis of effects to identify aberrant variables and rank likely causes. Advice includes a synthesizing strategy to eliminate the causes and return the process to normal operating conditions.[1]

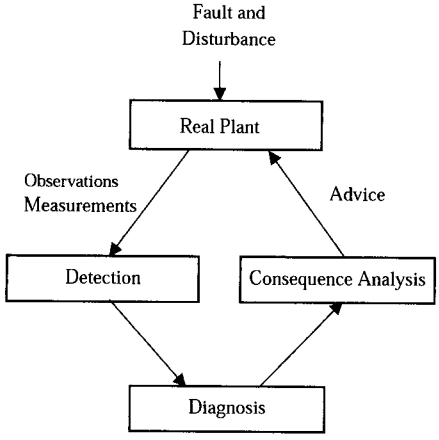


Figure 1. Fault detection and diagnosis scheme for a process monitoring.

In order to extract useful information from process data and utilize it for the monitoring of WWTP, applications of MSPC in the biological process have recently drawn a great interest by a few researchers.[2-8,16,17] Krofta *et al.*[16] applied the analysis techniques for dissolved air flotation. Rosen[2] adapted multivariate statistics based methods to the wastewater treatment monitoring system using simulated and real process data. Van Dongen and Geuens[17] illustrated that multivariate time series analysis can be a valid alternative of the dynamic modeling in WWTP. However, multivariate statistical analysis method has fundamental weak points in the biological WWTP. The biological treatment process has several peculiar features unlike chemical or industrial engineering. First of all, it is *nonstationary*, which means that the process itself changes gradually over time. WWTPs are hardly ever normally operated for long periods and what normality means also changes because of the nonstationarity. For example, seasonable variations show a dynamic pattern, for example, the process normal condition evolves according to the seasonal variations. So, conventional static PCA is not suited for nonstationary process monitoring as it assumes that data are independently identically distributed (iid) and they are obtained from a normal operating condition for a particular process. This is a problem for developing statistical control charts as they should be developed from a set of *stationary* and *normal* operating data. Second, many underlying phenomena of WWTP takes place simultaneously and it may be difficult to separate specific phenomenon among them. Namely, it has *multiscale* characteristics that have multiple simultaneous phenomena affecting the data at different time or frequency scales.[18,19] If these synchronous characteristics interfere or mask other time or frequency variations, called the masking effects, the situation turns troublesome because the multiscale variations are enlarging the confidence limits. This is unfavorable because the actual events can stay undetected by the monitoring algorithm while the plant is being under way of

the events.

Shortly, this paper applies two methods, one is for data-driven prediction modeling and the other is for multiresolution analysis technique. In this way, it is possible to take into account the multivariate, nonstationary and multiscale natures of WWTP. These approaches are organized by putting the PLS model and the multiresolution analysis together. An outline of this paper is as follows. The first subsection introduces the basic PLS principle and the conventional monitoring method briefly. In the second subsection, multiresolution analysis combine with the PLS regression is suggested for the multiscale monitoring. Then experimental results are illustrated and followed by discussion. Finally, the conclusions of this article are addressed.

# MATERIALS AND METHODS

## Partial Least Squares

Partial least squares (PLS) 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 that maximize the covariance between input and output variables. This projection decomposes variables of high collinearity into low dimensional variables (input score vector $\boldsymbol{t}$ and output score vector $\boldsymbol{u}$). The decomposition of **X** and **Y** by score vectors is formulated as follows[3]:

$$\mathbf{X} = \sum_{h=1}^{m} t_h p_h^T + \mathbf{F} = \mathbf{T}\mathbf{P}^T + \mathbf{E} \tag{1}$$

$$\mathbf{Y} = \sum_{h=1}^{m} u_h q_h^T + \mathbf{F} = \mathbf{U}\mathbf{Q}^T + \mathbf{F} \tag{2}$$

where $\boldsymbol{p}$ and $\boldsymbol{q}$ are the loading vectors which contain information about the relationship of variables, $m$ is the number of latent variables, **T** and **U** are the score matrices and **E** and **F** are residuals. A score vector is orthogonal and a loading vector is orthonormal. Although PLS is a regression technique, it is a visualizing technique whose ability enables us to interpret and search data sets more minutely.[2,5,7,14]

PLS projects **X** and **Y** variables simultaneously onto the same subspace, **T**, in such a manner that there is a good relation between the position of one observation on the **X**-plane and its corresponding position on the **Y**-plane. Once the PLS model has been derived, it is important to grasp its meaning. For this, the scores $\boldsymbol{t}$ and $\boldsymbol{u}$ are considered. They contain information about the observations and their similarities/dissimilarities in **X** and **Y** space with respect to the given problem and model. **X** and **Y** weights provide the way how the variables combine to form $\boldsymbol{t}$ and $\boldsymbol{u}$, which in turn express the quantitative relation between **X** and **Y**. Hence, these weights are essential for the understanding which **X** variables are important for modeling **Y**, which **X** variables provide common information, and also for the interpretation of the scores $\boldsymbol{t}$.[2,3]

On the other hand, once the PCA and PLS 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 (or $SPE_X$) are usually used for this purpose. After decomposing the observed data, the score value in the model space at time $k$,

$$\boldsymbol{t}_k = [\boldsymbol{p}_1\ \boldsymbol{p}_2 \cdots \boldsymbol{p}_m]^T \boldsymbol{x}_k = \hat{\boldsymbol{P}}^T \boldsymbol{x}_k \in \Re^m \tag{3}$$

is distributed as $N(\boldsymbol{0}, \boldsymbol{\Lambda}_s)$, where $\boldsymbol{\Lambda}_s$ is the diagonal part of $\boldsymbol{\Lambda} = \boldsymbol{P}^T\boldsymbol{R}\boldsymbol{P}$ and $\boldsymbol{R}$ is the sample covariance matrix. $\boldsymbol{t}_k$ is thus an $m$-dimensional reduced representation of the observed vector $\boldsymbol{x}_k$. On the other hand, the residual at time $k$

$$\mathbf{e}_k = \mathbf{x}_k - \hat{\mathbf{x}}_k = \left(\mathbf{I} - \hat{\mathbf{P}}\hat{\mathbf{P}}^T\right)\mathbf{x}_k \in \Re^n \tag{4}$$

is the part not explained by the PCA and PLS models. Generally, the squared weighted score ($T_k^2 = \boldsymbol{t}_k{}^t \boldsymbol{\Lambda}^{-1} \boldsymbol{t}_k$) and the squared residual ($Q_k = \mathbf{e}_k \mathbf{e}_k^t$) are used as monitoring indices for process monitoring or fault detection. Generally, the approximated 100(1-$\alpha$)% control limit for $T^2$ can be calculated by means of a F-distribution as

$$T_{\lim}^2 = \frac{m(n-1)}{n-m} F(m, n-1; \alpha) \tag{5}$$

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

$$Q_{\lim} = \theta_1 \left[ \frac{c_\alpha \sqrt{2\theta_2 h_0^2}}{\theta_1} + 1 + \frac{\theta_2 h_0 (h_0 - 1)}{\theta_1^2} \right]^{1/h_0} \tag{6}$$

where $\theta_j = \sum_{s+1}^{p} (\Sigma_{ii})^j$ for j = 1,2,3, $h_0 = 1 - 2\theta_1\theta_3 / 3\theta_2^2$ and $c_\alpha$ is the normal deviate cutting off an area $\alpha$ of the upper tail of the distribution if $h_0$ is positive and under the lower tail if $h_0$ is negative.[2,3,15]

For a new on-line sample $x_{new}$, if $T_{new}^2 < T_{\lim}^2$ and $Q_{new} < Q_{\lim}^2$, 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. However, the conventional MSPC method, such as $T^2$ and $Q$ statistic, does not always function well, because it cannot detect the changes of correlation among process variables if $T^2$ and $Q$ statistic are inside the confidence limits. Also, the autocorrelated observations form a time series in MSPC. Using control limits only, one observation is considered at a time and therefore, the presence of proceeding and succeeding points is masked out by a 'window' consisting of only on observation. For this autocorrelation problem, window must be increased to the size required to asses sequences of plotted points for the relevant pattern or disturbance.

## Generic Dissimilarity Measure

Recently, several dissimilarity indices with the distribution between two data sets have emerged.[4,19,20] They are based on the idea of that a change of process operation can be detected by comparing the distribution of data sets with reference data set because the data distribution reflects the corresponding process operating condition. In this paper, we applied a generic dissimilarity measure (GDM) algorithm which considers the importance of each transformed variable and compare successive data set for nonstationarity.[4] It compares covariance structures of two successive data sets with time-window concept and represents the degree of dissimilarity between them by considering the importance of each transformed variable.

The GDM algorithm is divided into two major steps, which are the training phase of historical data sets under normal conditions and the monitoring phase of a new data set in various events. As a training phase of normal data, the intervals and limits of characteristic values are defined. On the basis of the fact that the covariance matrix of the pooled matrix of two data sets can be decomposed using singular value decomposition (SVD). Suggested by Yoo *et al.*,[4,20] GDM is as follows.

First, start from building two successive data sets with a moving window and normalize them with sample mean and sample variance ($\mathbf{X}_1$ and $\boldsymbol{X}_2$). Then, find the sample covariance matrix and apply SVD to it.

$$\boldsymbol{S}_i = \frac{1}{N_i - 1} \boldsymbol{X}_i^T \boldsymbol{X}_i, \quad i = 1,2 \tag{7}$$

$$\mathbf{S} = \frac{1}{N-1} \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix}^T \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix} = \frac{N_1 - 1}{N-1} \mathbf{S}_1 + \frac{N_2 - 1}{N-1} \mathbf{S}_2 \tag{8}$$

where $S_i$ is the sample covariance of data set $i$, $S$ is the pooled sample covariance, $N_1$ and $N_2$ is the sample number of data set 1 and 2, respectively, and $N$ is the total sample number, that is, $N_1 + N_2$. Figure 2 represents the concept of window and step size using a moving window. Window size means the sample number in each data set and step size means the monitoring interval. It should be noticed that window size of suitable duration is chosen to obtain enough redundant information about the system dynamics and step size be selected to preserve enough monitoring performance.

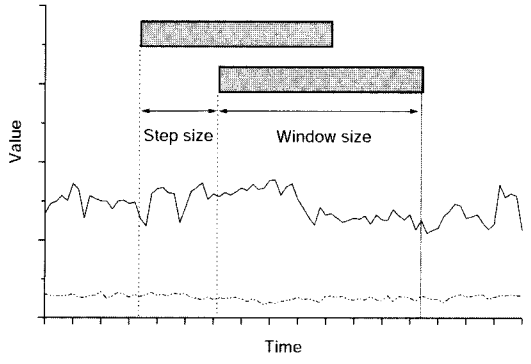


Figure 2. Moving windows between successive two data set.

Second, apply SVD to $S$ and transform the data matrix ($X_i$) to orthogonal variables ($Y_i$).

$$SP = P\Lambda \tag{9}$$

$$Y_i = \sqrt{\frac{N_i - 1}{N - 1}} T_i \tag{10}$$

where $P$ is the loading matrix, $\Lambda$ is the diagonal matrix, and $T_i$ is $X_iP$.

Third, find the sample covariance matrix ($R$) of two transformed data sets ($Y_1$ and $Y_2$) and apply SVD to $R$.

$$R_1 + R_2 = \Lambda \tag{11}$$

$$R_i q_j^i = \lambda_j^i q_j^i, \quad i = 1,2 \ \ and \ \ j = 1,\ldots,r \tag{12}$$

where $q_j^i$ is the loading vector, $\lambda_j^i$ is the eigenvalue, and r is the PC numbers. By combining Eqs. (11) and (12),

$$R_1 q_j^1 = \lambda_j^1 q_j^1 \ \ and \ \ R_2 q_j^1 = (\Lambda_{jj} - \lambda_j^1) q_j^1 \tag{13}$$

After these linear transformations, two sample covariance structures of the transformed matrices share the eigenvectors, then the eigenvalues satisfy the following equation.

$$\lambda_j^1 + \lambda_j^2 = \Lambda_{jj} \tag{14}$$

where $\lambda_j^i$ is the $j^{th}$ eigenvalue in the $i^{th}$ data set and $\Lambda_{jj}$ is the eigenvalue in the total data set. As two data sets are more similar, their eigenvalues is closer to $0.5\Lambda_{jj}$. As $j$ increases, $\Lambda_{jj}$ sharply decreases. In general, the first few principal components ($r$) explain most variation of data sets.

The following generic dissimilarity measure (GDM, $D$) is defined for measuring the dissimilarity of two data sets.

$$D = \frac{4\sum_{j=1}^{r}\left(\lambda_j - \frac{\Lambda_{jj}}{2}\right)^2}{\sum_{j=1}^{r}\Lambda_{jj}^2} \tag{15}$$

$D$ has a value between 0 and 1. The more similar two data are, the closer $D$ is to 0. The more dissimilar two data are, the closer $D$ is to 1.

Finally, find the $(1-\alpha)100\%$ confidence interval of each eigenvalue. For many samples, it is reasonable that we assume that each eigenvalue is a normal random variable by the central limit theorem. With the samples obtained from a normal operation, the interval that there exist $(1-\alpha)100$ percents of the eigenvalues calculated above is obtained by

$$\begin{aligned} &-t(1-\alpha/2; N-2)s\{\lambda_j^i\} + \overline{\lambda_j^i} \\ &\le \lambda_j^i \le t(1-\alpha/2; N-2)s\{\lambda_j^i\} + \overline{\lambda_j^i} \end{aligned} \tag{16}$$

where $\bar{\lambda}_j^i$ is the sample mean and $s\{\lambda_j^i\}$ is the sample variance. That is, (1-$\alpha$)100% of $\lambda_j^i$ are below the limit value and the remaining are above it.[7] Typically $\alpha$ takes the value of 0.05 and 0.01 for the warning and action limit. In this paper we used that $\alpha$ is 0.05, that is, 95% confidence limit. The control limits of each index, $\alpha$, are determined so that the number of samples outside the control limit is 5% of the total samples.[19]

For the monitoring phase, the confidence limits of the normal data set are calculated through the previous step. The sample representing the current operating condition is scaled by the sample mean and the sample variance obtained in the previous steps. Then the GDM and the corresponding eigenvalues are calculated using the explained method. The GDM evaluating the difference between two successive data sets quantitatively can detect a change of process operations and monitor a distribution of time series data. If the GDM is outside the control limit or deviates from zero value, the operating condition is changed and the existence of disturbances is detected. Then, we focus on the individual variation of each eigenvalue at several scales. Most of variation is captured by the first several eigenvectors and so only several eigenvalues are considered as a monitoring index. The remaining variations which are not captured by the principal eigenvectors are relatively very small and they are not critically identified whether they are caused by process change or noises. If any eigenvalue exceeds its corresponding confidence limit, the process operation at that scale may be changing and a certain event or disturbance occurs.[4,20]

Important consideration in monitoring the process changes or the operating condition is the determination of appropriate window and step sizes. These quantities should be carefully selected taking into consideration the process characteristics. A step size of one means there is very slim change of pattern recognition because patterns require more than one data point (i.e. data vector in MSPC) in order to be detected. Therefore, a step size of 1 can provide only a simplistic measure for showing whether the process is a state of statistical control or not. As with the univariate statistical control, larger step size (*i.e.* >1) can potentially provide the required data for identification of abnormal patterns. The latter are also known as runs or a succession of items of the same class or repetitive patterns within a sequence. On the other hand, window size should be determined according to the theoretical probability of committing a Type I error. While a process is in statistical control, that is, for any given number of points there can be found a theoretical probability that they will fall within a certain distance from the process mean and in a pre-specified order (e.g. distances from the mean and order of occurrence can resemble some possible forms of pre-specified unnatural patterns or disturbances. Therefore, assuming that a process is behaving randomly, the larger the number of windowed points the less likely the chance that a sequence of random values will match a pre-specified order such as a trend or stratification. This is, the larger the window size, the smaller the probability of wrongly detecting a non-random patter and disturbance. But too excessive window size makes the detection of pattern change slow, delayed detection. In this respect, process knowledge which indicates true behavior of the process may provide more reliable source and less computational load than theoretical approach. We suggest that the window size should be large in comparison to the time constant of the process, and the step size should be small in comparison to the sampling time.[4]

## Multiresolution Analysis and Monitoring

Figure 3 shows the proposed method of the PLS modeling and multiresolution analysis (MRA) which can treat the peculiar characteristics of the biological treatment process and isolate and diagnose their fault sources with a multiscale approach. In the first place, the PLS model is constructed with normal historical data

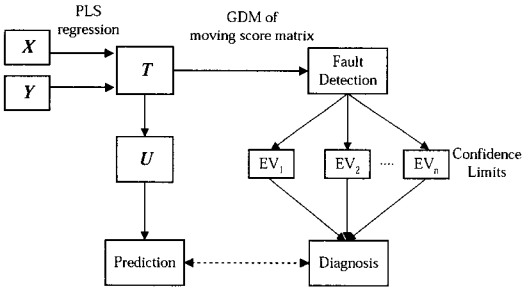


Figure 3. Multiresolution analysis for the PLS modelling and monitoring.

in order to solve the multivariate and collinear problems in a plant. It is used to represent the process behavior and the common-cause variations of a plant and excludes noise, measurement errors, and those variations that are uncorrelated to **Y** variables. Then, MRA for the score values of the PLS model is executed by the GDM to detect the process change and to diagnose different kinds of faults and disturbances. In order to tackle normality problem[19], each successive data set in the GDM consists of the PLS score values with a moving window because the PLS score values are normally distributed than the original variables themselves. This is a consequence of the central limit theorem, which can be stated as follows: If the sample size is large, the theoretical sampling distribution of the mean can be approximated closely with a normal distribution. Thus, we would expect the PLS scores, which are a weighed sum like a mean, to be distributed approximately normally.[14,15] Figure 4 demonstrates the normality comparison between the original values and the PLS scores. Therefore, as the abnormality will manifest itself as a shift or time series distribution change in the score value than the original variables. As the abnormality will manifest itself as a shift in the score plane like $T^2$ statistic of the PCA and PLS monitoring, it will be shown as a dissimilarity value between successive two data sets, that is, GDM. A moving window of the PLS score values with a GDM concept may be a remedy of nonstationary problem of the PLS monitoring. On the other hand, if the relationships between

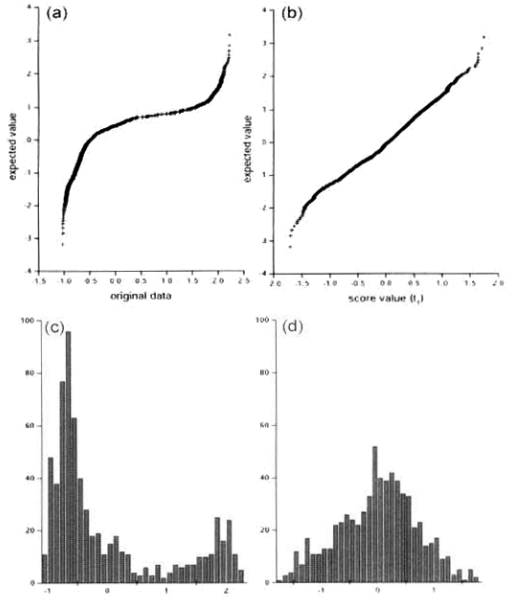


Figure 4. Normal probability plot and histogram of original data and the PLS score values (a) normal probability plot of original data, (b) probability plot of the score values, (c) histogram of original data, (d) histogram of the score values.

the process variables are rapidly changed and the correlation structure has a breakdown, $SPE_X$ of the PLS model can be included in two data sets of the proposed MRA algorithm. Since the inner relationship between input and output variables in WWTP is slowly changed, only score values of the PLS model are sufficient for process monitoring.

In this work, the confidence limit of individual eigenvalue is used to multiscale fault detection and isolation. If each of eigenvalues exceeds to its corresponding confidence limit, it means that the current process at that scale is changing and a certain event is occurring. By monitoring at each scale, we can diagnose diverse process variations and events, i.e., diagnosis of slow variations (seasonal fluctuations or other long-term dynamics), middle scale variation (internal disturbance, process operation change), and instantaneous variations (input disturbances, faults or sensor noises).[3,20] Because it represents the corresponding characteristics at each scale, this

multiresolutional analysis can discover information on the scale where process changes, faults and events occur and analyze the physical/biological reasons. The proposed MRA method can give us the diagnosis and interpretation capability of events and fault sources. Note that it can get rid of nonstationary problem systematically by comparing successive data sets with a moving window concept. Moreover, it does not bring about the zero padding problems unlike the other MRA, such as wavelet.[18]

# EXPERIMENTAL RESULTS AND DISCUSSION

## Process Data

Process data were collected from a biological treatment plant treating coke wastewater from an iron and steel producing plant in Korea, so called biological effluent treatment (BET). Figure 5 shows the layout of the studied full-scale plant. This treatment plant uses an activated 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 treatment plant 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. In particular, cyanide (CN) concentration is a very important load among the influent loads.

## PLS Modeling

Table 1 describes the process variables of **X** and **Y** blocks. Eleven process and manipulated variables, the **X** block, are used to model three process output variables, the **Y** block. The **Y** block consists of the sludge volume index (SVI), the reduction of cyanide, and the reduction of COD. 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 are used for the calibration of the PLS model. The remaining 314 observations are used as a test set in order to verify the proposed method. For the determination of the latent variable number of PLS model, a cross-validation method is used and four LVs were selected in the PLS model. It manages to capture about 54% of the $\boldsymbol{X}$ block variance and 61% of the $\boldsymbol{Y}$ block variance by projecting the variables from dimension 14 to dimension 4, which is originated from the troublesome and difficult treatment of coke wastewater. The results of the PLS model are represented in Table 2.

An appealing feature of the PLS model is the

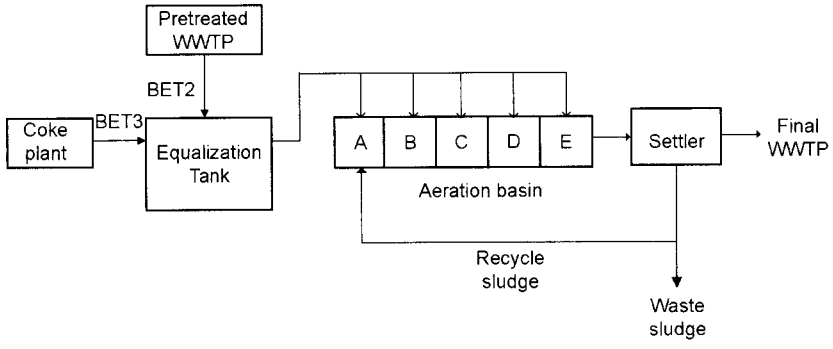


Figure 5. Plant layout of full-scale coke wastewater treatment process.

Table 1. Process input and output variables in an industrial biological wastewater treatment plant

| Var. No | Variable | Description | Unit | Mean | S.D. |
|---|---|---|---|---|---|
| $X_1$ | $Q_2$ | Flow rate from BET2 | $m^3h^{-1}$ | 178 | 15.3 |
| $X_2$ | $Q_3$ | Flow rate from BET2 | $m^3h^{-1}$ | 84.8 | 8.1 |
| $X_3$ | $CN_2$ | Cyanide from BET2 | $mgL^{-1}$ | 2. 5 | 0.35 |
| $X_4$ | $CN_3$ | Cyanide from BET3 | $mgL^{-1}$ | 14.9 | 1.768 |
| $X_5$ | $COD_2$ | COD from BET2 | $mgL^{-1}$ | 157.8 | 19.88 |
| $X_6$ | $COD_3$ | COD from BET3 | $mgL^{-1}$ | 2088 | 306 |
| $X_7$ | MLSS_%E | MLVSS at a final aeration basin | $mgL^{-1}$ | 1547 | 292.8 |
| $X_8$ | $MLSS_r$ | MLSS in recycle line | $mgL^{-1}$ | 2194 | 346 |
| $X_9$ | $DO_{aerator}$ | DO at a final aeration basin | $mgL^{-1}$ | 2.0 | 0.98 |
| $X_{10}$ | $T_{influent}$ | Influent temperature | °C | 37.6 | 2.513 |
| $X_{11}$ | $T_{aerator}$ | Temperature at a final aerator | °C | 30.74 | 2.28 |
| $Y_1$ | $SVI_{settler}$ | Sludge volume index at settler | $mgL^{-1}$ | 63.31 | 21.73 |
| $Y_2$ | $CN_{red}$ | Cyanide reduction | $mgL^{-1}$ | 19.31 | 2.2 |
| $Y_3$ | $COD_{red}$ | COD reduction | $mgL^{-1}$ | 605.4 | 97 |

Table 2. Variations explained by the PLS model

| LV | X Blocks (Cumulative) | Y Blocks (Cumulative) |
|---|---|---|
| LV 1 | 0.192 | 0.319 |
| LV 2 | 0.338 | 0.481 |
| LV 3 | 0.446 | 0.581 |
| LV 4 | 0.540 | 0.607 |

modeling ability, that is, predictive capability. Figure 6 shows the real and predicted value from PLS model and displays the residual of ***Y*** blocks. The prediction values of the reduction of COD and the reduction of CN are explained very well in the test periods and manifest the prediction power of the PLS model for the response **Y** variables. However, the prediction of SVI of secondary settler is not satisfied unlike other two quality variables. That may result from measurement inaccuracy and the operator's carelessness, which needs a precise measurement skill to the operator. The residual values of **Y** blocks show the sum of differences between the real and predicted values for three response variables, which is mainly caused by the residual error of SVI prediction.

## Interpretation of PLS Modeling

For the interpretation of the plant, the PLS loading weights are considered to see how **X** and **Y** variables are interrelated. The loading plot in Figure 7 confirms the underlying physical and biological phenomena as the PLS model distinguished chemical and biological variables. It represents that the specific **X** and **Y** variables load strongly in the first two LVs, where $COD_3$, $COD_2$, and $T_{aerator}$ for COD reduction are closely correlated as seen in the left middle side of Figure 7. The first Y variable, the COD removal rate of the plant is strongly influenced by the COD load from BET2 and BET3 and the temperature in aerators. This corresponds to the fact that heterotrophic biomass activity for the carbonaceous nutrients is influenced by the temperature in the biological treatment. These variables are uncontrolled o r partially controlled throughout the process and therefore exhibit large variations. The second group for CN reduction is related to $CN_2$, $CN_3$, $T_{influent}$, $Q_2$ and $Q_3$, and DO of aerator which are rate related components of the reaction rate, such as monod equation. It indicates that the DO concentration in the aeration tank should be controlled. On the other hand, it is usually known that cyanides are toxic to heterotrophic

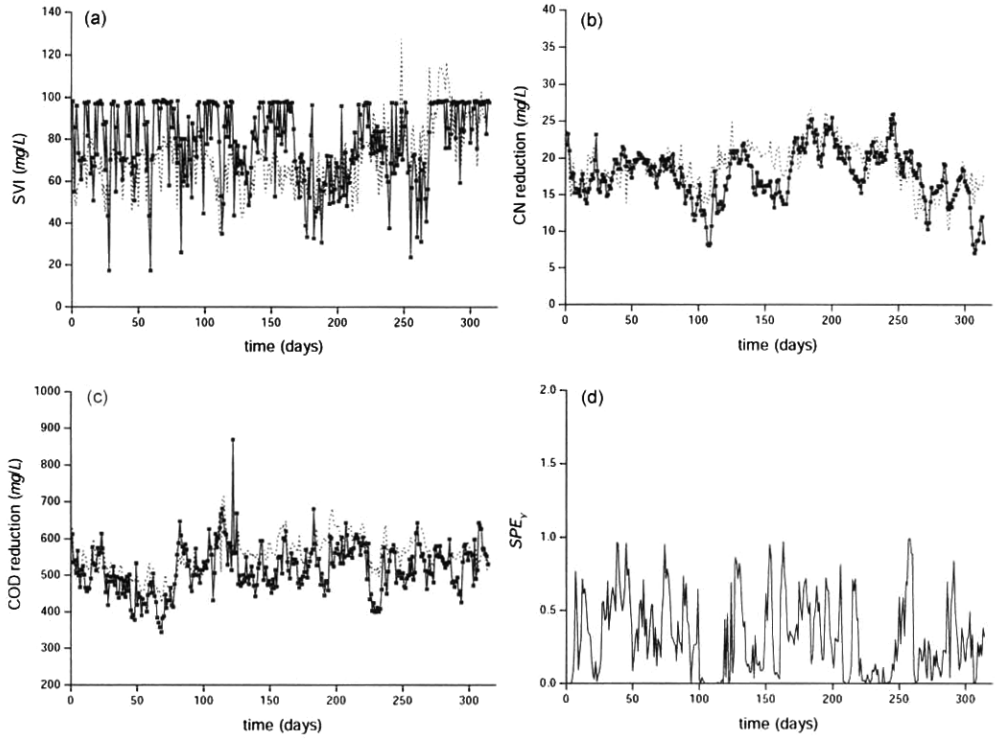


Figure 6. Prediction results of the PLS model with real Y value (solid line with squares) and predicted value (dotted line), (a) SVI, (b) reduction of CN, (c) reduction of COD, (d) squared residual error of Y variables ($SPE_Y$).

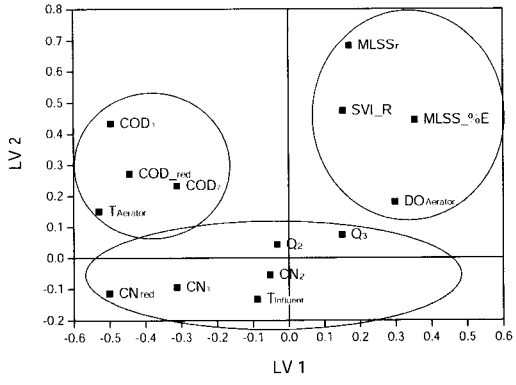


Figure 7. Loading plot of the PLS model.

bacteria and inhibitory to their reaction rate. In Figure 7, the cyanide load is counter-connected with the heterotrophic organism concentration (MLSS_%E) which is shown in the opposite direction of each other in the loading plots. 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.[20,21] The third group is made up of $MLSS_r$ and MLSS_%E with SVI of secondary settler in the right upper side region. It exemplifies that the settleability of biomass is related to the microorganism amount in the aerator ($MLSS_r$) and the settler (MLSS_%E).

Since a real wastewater plant has generally more than 3 LVs, it may be quite useful to all the PLS weight vectors together with the fraction that is explained by the latent variables. That is, the variable importance in the projection (VIP) is a good measure of the influence of all variables in the PLS model on the response variables.[22] VIP plot in Figure 8 reveals that $COD_3$ is the most important variable, followed the $T_{aerator}$, $MLSS_r$, MLSS_%E, and so on, where the higher the value is, the more influ-

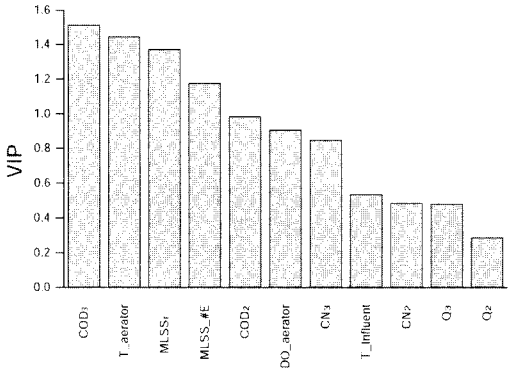


Figure 8. Variable influence on projection (VIP) for the predictor variables of the PLS model.

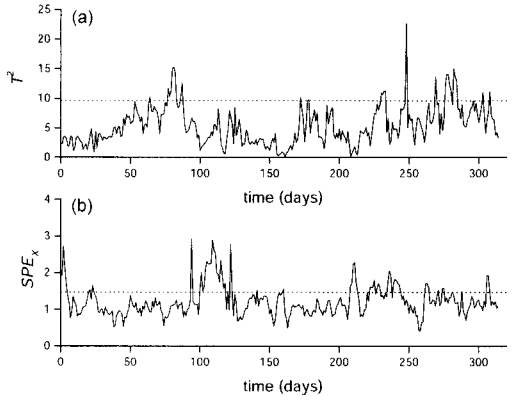


Figure 9. Monitoring performances based on $T^2$ and $SPE_X$ statistics with 95% confidence limits.

ential the variables are. This can be interpreted that COD influent from BET3 is most important to the plant treatment efficiency of the plant.

In Figure 9, the $T^2$ and $SPE_X$ monitoring charts are shown. The horizontal line corresponds to the 95% significance level of the training data. From this figure, we can see three deviations in the monitoring charts of $T^2$ and $SPE_X$ statistic. During samples 75 to 80, the $T^2$ charts deviated slightly, which indicates that the deviations are large within the internal model. However, the $SPE_X$ chart does not increase, which indicates that the internal mutual relations are not altered. During samples 100 and 120, the $SPE_X$ chart deviated. In this period, the process received influents with a high cyanide and COD 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. On the other hand, around sample 250, the $T^2$ chart has a peak value, while the $SPE_X$ chart is maintained in the vicinity of 95% confidence limit for a long time. We infer that the process has experienced the large transition in the operating condition at this time, but does not know its cause correctly. In order to identify more obvious cause for the deviation, the contributions from every measurement variable might be calculated. Also, it cannot diagnose and isolate their fault scales from the view point of the process dynamics.

## Multiresolution Analysis

After the construction of the PLS model, MRA was processed to the score matrix (**T**) of the PLS model. To monitor the process change or detect fault and event, window and step sizes are 15 samples considering the SRT and 3 samples considering the hydraulic retention time (HRT), respectively.

MRA to the PLS score values of the test data set are shown in Figure 10. As shown in Figure 10(a), the GDM started to change at sample 65 and deviated during around samples 65 - 120 (March 3, 2000 - April 27, 2000), where a large process change happened at this time. It shows more rapid detection ability than the PLS method. Four eigenvalues which indicate their own specific scale disturbance are depicted in Figure 10(b-e). The remaining eigenvalues have little information and gives only high frequency information such as measurement noises. From Figure 10, we can know that the first and second eigenvalues largely contribute to the increase of the GDM and are representative of middle scale disturbances. In detail, the process change is first detected in the GDM, which is caused by the peaks of the second eigenvalue and then has experienced the systematic variations of the first eigenvalue. It is easily

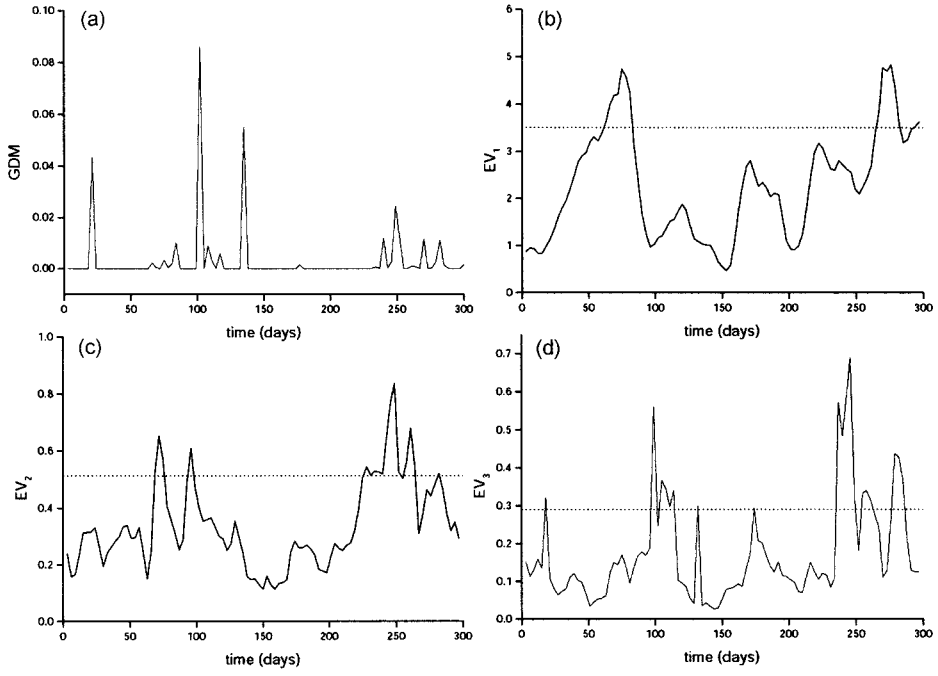


Figure 10. Monitoring performances of the proposed MRA method for the PLS score values with 95% confidence limits (a) GDM, (b) $EV_1$, (c) $EV_2$, (d) $EV_3$.

identified and visualized by monitoring each eigenvalue pattern at two scales. At this time, the plant received high input cyanide and COD load, while a small influent flow rate, that is, a highly concentrated load. It reduced the activity of the microorganisms and diminished the settling performance, then turned up the SVI increase in the secondary settler. From this result, it has been seen that sludge and floc formation changes due to high load and influent quality. Figure 11 shows the contribution plot at sample 70. It means that a large influent load broke out the external disturbance and were transformed into an internal disturbance, and then it changed the process operating region in the plant. Meanwhile, GDM deviated again from sample 230 to the last of test data set (August 16, 2000 - November 9, 2000). During the summer, WWTP was modified and a number of treatment equipments and facilities were appended. This made it feasible for operators to change the operation strategy which increased the MLSS concentration and maintained the high

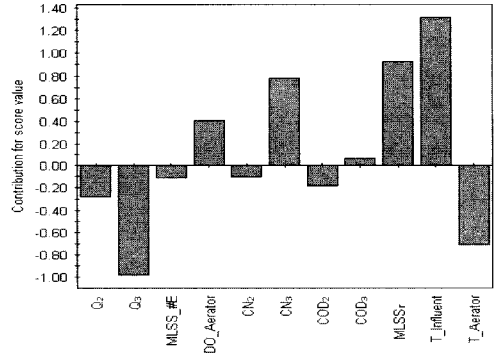


Figure 11. Contribution plot of the PLS score value at sample 70.

dissolved oxygen concentration.[4,20] It invokes the large process changes, which is shown as a gradual increase of the first eigenvalue in Figure 10(b). This result confirms that MRA is distinctly better than other conventional methods for a multiscale process change in a nonstationary signal of unknown characteristics since it can extract information resulting from the change in process operation which contributes the localization of different process faults and events.

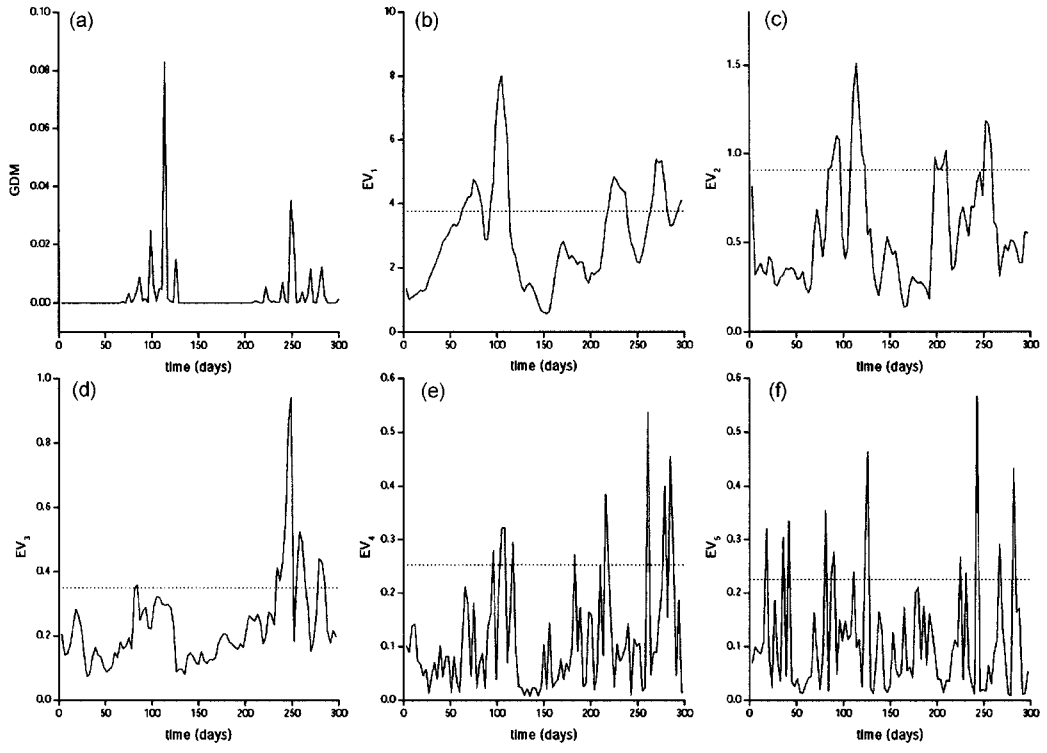


Figure 12. Monitoring performances of the proposed MRA method for the PLS scores and the $SPE_X$ value with 95% confidence limits (a) GDM, (b) $EV_1$, (c) $EV_2$, (d) $EV_3$, (e) $EV_4$, (f) $EV_5$.

For the case that the relationships between the process variables are rapidly changed and the correlation structure has a breakdown, the $SPE_X$ values of the PLS model is appended to the PLS score matrices for MRA. Figure 12 shows the monitoring results of the proposed method. Although it manifests similar monitoring performance of the previous method, it can detect more rapidly the events than the previous one. It indicates that WWTP had undergone a certain event, not explained by the PLS model, which is captured by adding the residual error of the PLS model. It can effectively unify both the $T^2$ and $SPE_X$ statistics together and put advantage of them into a single representing index.

## CONCLUSIONS

In this paper, a new approach of a data-driven modeling and multiresolution monitoring method is presented in order to solve collinear, multivariate, nonstationary, and multiscale problems in the biological treatment plant. It is achieved by combining the PLS regression and multiresolution analysis, where the PLS model is used for the prediction and MRA is utilized to detect and diagnose the fault and disturbance with a multiscale concept. It could give us the prediction, detection, and diagnosis power at a time and make the investigation about nonstationary and multiscale phenomena practicable. The case study showed that it not only gave good modeling performance and higher interpretability of a complex biological plant but also the suitable power of detection and isolation about various faults and events occurring in an industrial treatment plant.

## ACKNOWLEDGEMENTS

This work was supported by Korea Research Foundation Grant (KRF-2004-041-D00151) and

also by a grant No. (R01-2002-000-00007-0) from Korea Science & Engineering Foundation. And Peter Vanrolleghem is kindly thanked for his valuable discussion.

## NOMENCLATURES

$b_i$ = regression coefficient
$\boldsymbol{B}$ = diagonal matrix of regression coefficient $b_i$
$D$ = generic dissimilarity index
$\boldsymbol{E}$ = residual matrix of **X** variables
$\boldsymbol{F}$ = residual matrix of **Y** variables
$N_i$ = sample number of data set $i$
$\boldsymbol{N}$ = sample number of total data set
$\boldsymbol{p}_i$ = loading vector
$\boldsymbol{q}_i$ = loading vector
$r$ = principal component number
$\boldsymbol{R}_i$ = sample covariance matrix of the transformed variables of data set $i$
$\boldsymbol{S}$ = sample covariance matrix of total dataset
$s^2(i)$ = sample variance of $x_i$
$s\{\lambda_j^i\}$ = estimated standard deviation of $\lambda_j^i$
$\boldsymbol{S}_i$ = sample covariance matrix of data set $i$
$\boldsymbol{SPE}_X$ = sum of squared prediction error of **X** variables
$\boldsymbol{SPE}_Y$ = sum of squared prediction error of **Y** variables
$T_i$ = score matrix
$T_i$ = score vector
$U_i$ = score matrix
$u_i$ = score vector
$\boldsymbol{X}$ = total data matrix
$\boldsymbol{X}_i$ = sample data matrix $i$
$\bar{\boldsymbol{x}}_k$ = sample mean of $\boldsymbol{x}_k$
$\boldsymbol{X}$ = process variable
$\boldsymbol{Y}$ = output (response) variable
$\boldsymbol{Y}_i$ = transformed variable of data set $i$
$\boldsymbol{Z}_X$ = sample matrices of **X** variables
$\boldsymbol{Z}_Y$ = sample matrices of **Y** variables

### Greek Letters

$\lambda_j^i$ = $j^{th}$ eigenvalue of $i^{th}$ data set
$\bar{\lambda}$ = sample mean
$s\{\lambda\}$ = sample variance
$\Lambda_{jj}$ = eigenvalue of total data set

## REFERENCES

1. Olsson, G. and Newell, B., Wastewater Treatment Systems: Modelling, diagnosis and Control, IWA, UK (1999).
2. Rosen, C. and Lennox, J. A., "Multivariate and multiscle monitoring of wastewater treatment operation," *Water Res.*, **35**(14), 3402-3410 (2001).
3. Teppola, P., Multivariate process monitoring of sequential process data - A chemometric approach, Ph.D. Thesis, Lappeenranta University, Finland (1999).
4. Yoo, C. K., Choi, S. W. and Lee, I., "Disturbance detection and Isolation in the Activated Sludge Process," *Water Sci. Tech.*, **45**(4-5), 217-226 (2002).
5. 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. Biotechnol.*, **105**(1-2), 135-163 (2003).
6. Yoo, C. K., Lee, D. S. and Vanrolleghem, P. A., "Application of multiway ICA for on-line monitoring of a sequencing batch reactor," *Water Res.*, **38**(7), 1715-1732 (2004a).
7. Yoo, C. K., Yoon, H. B., Lee, I. B., Vanrolleghem, P. A. and Rosen, C., "Application of fuzzy partial least squares (FPLS) modeling nonlinear biological processes," *Korean J. Chem. Eng.*, **21**(6), 1087-1097 (2004b)
8. Ragot, J., Grapin, G., Chatellier, P. and Colin, F., "Modeling of a water treatment plant. A multi-model representation," *Environmetrics*, **12**, 599-623 (2001).
9. Henze, M., Gujer, W., Mino, T. And van Loosdrecht M. C. M., Activated sludge models ASM1, ASM2, ASM2d and ASM3 IAWPRC Scientific and Technical Report No. 9. International Water Association, UK (2000).
10. Yoon, S. P., "Comparison of biological nutrient removal processes by activated sludge model no.2," *J. of Korean Society of*

*Environmental Engineers*, **21**(4) 609-616 (1999).
11. Kim, I. S., Young, J. C., Kim, S. Y. and Kim, S. M., "Development of monitoring methodology to fingerprint the activated sludge processes using oxygen uptake rate," *Environ. Eng. Res.*, **6**(4), 251-259 (2001).
12. Kim, S. H., Lee, H. J., Kim, C. W., Ko, J. H. and Woo, H. J., "Application of the genetic algorithm for the parameter estimation of activated sludge models No. 1," *J. of Korean Society of Environmental Engineers*, **24**(10), 1723-1730 (2002).
13. Jeong, H. S., Choi, D. J. and Shin, H. S., "Estimation of model parameters in organic degradation and nitrification processes of ASM3 using respirometry and genetic algorithm," *J. of Korean Society of Environmental Engineers*, **26**(8), 904-910 (2004).
14. Johnson, R. A. and Wichern, D. W., Applied Multivariate Statistical Analysis, 3nd ed., Prentice Hall, Englewood Cliffs, USA (1992).
15. Wise, B. M. and Gallagher, N. B., "The process chemometrics approach to process monitoring and fault detection," *J. Process Control*, **6**, 329-348 (1996).
16. Krofta, M., Herath, B., Burgess, D. and Lampman, L., "An attempt to understand dissolved air flotation using multivariate analysis," *Water Sci. Tech.*, **31**(3-4), 191-201 (1995).
17. Van Dongen, G. and Geuens, L., "Multivariate tie series analysis for design and operation of a biological wastewater treatment plant," *Water Res.*, **32**, 691-700 (1998).
18. Bakshi, B. R., "Multiscale PCA with application to multivariate statistical process monitoring," *AIChE J.*, **44**, 1596-1610 (1998).
19. Kano, M., Nagao, K., Hasebe, S., Hashimoto, I., Ohno, H., Strauss, R. and Bakshi, B., "Comparison of statistical process monitoring methods: Application to the eastman challenge Problem," *Comput. Chem. Eng.*, **24**, pp. 175-181 (2000).
20. Yoo, C. K., Choi, S. W. and Lee, I., "Dynamic monitoring method for multiscale fault detection and diagnosis in MSPC," *Ind. Eng. Chem. Res.*, **41**, 4303-4317 (2003).
21. Lee, D. S., Neural Network Modeling of Biological Wastewater Treatment Processes, Ph.D. Dissertation, POSTECH, Korea (2000).
22. Eriksson, L., Hermens, J. L. M., Johansson, E., Verhaar, H. J. M. and Wold, S., "Multivariate analysis of aquatic toxicity data with PLS," *Aqu. Sci.*, **57**(3), 1015-1621 (1995).