ARTICLE

BIOTECHNOLOGY and BIOENGINEERING

Multi-Model Statistical Process Monitoring and Diagnosis of a Sequencing Batch Reactor

Chang Kyoo Yoo,[1,2,3] Kris Villez,[3] In-Beum Lee,[2] Christian Rosén,[4] Peter A. Vanrolleghem[3,5]

[1]Department of Environmental Science and Engineering, Center for Environmental Research, College of Environmental and Applied Chemistry, Kyung Hee University, Seocheon-dong, Giheung-gu, Yongin-Si, Gyeonggi-Do, 446-701; Korea; telephone: +82-31-201-3824; fax: +82-31-204-8114
e-mail: ckyoo@khu.ac.kr; ChangKyoo.Yoo@biomath.ugent.be

[2]School of Environmental Engineering and Science / Department of Chemical Engineering, Pohang University of Science and Technology, Pohang, Korea

[3]BIOMATH: Department of Applied Mathematics, Biometrics and Process Control, Ghent University, Ghent, Flanders, Belgium

[4]IEA: Department of Industrial Electrical Engineering and Automation, Lund University, Lund, Sweden

[5]modelEAU, Département de Génie Civil, Pavillon Pouliot, Université Laval, Québec, Canada

Received 31 March 2006; accepted 11 September 2006
Published online 10 October 2006 in Wiley InterScience (www.interscience.wiley.com). DOI 10.1002/bit.21220

ABSTRACT: Biological processes exhibit different behavior depending on the influent loads, temperature, microorganism activity, and so on. It has been shown that a combination of several models can provide a suitable approach to model such processes. In the present study, we developed a multiple statistical model approach for the monitoring of biological batch processes. The proposed method consists of four main components: (1) multiway principal component analysis (MPCA) to reduce the dimensionality of data and to remove collinearity; (2) multiple models with a posterior probability for modeling different operating regions; (3) local batch monitoring by the T^2- and Q-statistics of the specific local model; and (4) a new discrimination measure (DM) to identify when the system has shifted to a new operating condition. Under this approach, local monitoring by multiple models divides the entire historical data set into separate regions, which are then modeled separately. Then, these local regions can be supervised separately, leading to more effective batch monitoring. The proposed method is applied to a pilot-scale 80-L sequencing batch reactor (SBR) for biological wastewater treatment. This SBR is characterized by nonstationary, batchwise, and multiple operation modes. The results obtained for the pilot-scale SBR indicate that the proposed method has the ability to model multiple operating conditions, to identify various operating regions, and also to determine whether the biosystem has shifted to a new operating condition. Our findings show that the local monitoring approach can give more reliable and higher resolution monitoring results than the global model.

Biotechnol. Bioeng. 2007;96: 687–701.

KEYWORDS: batch monitoring and supervision; biological system; multiple operational modes; probabilistic modeling; sequencing batch reactor (SBR); wastewater treatment

Introduction

Biological processes will only achieve near-optimal performance if they are modeled and monitored using reliable techniques. An adequate model of a biological process enhances the understanding of the process and can form the basis for better process design, control, and operation. In addition, efficient process monitoring and early fault detection methods allow corrective action to be taken well before a dangerous situation occurs. In biological wastewater treatment plants, most changes are slow when the process is recovering from a 'bad' state to a 'normal' state.

Correspondence to: C.K. Yoo
Contract grant sponsor: Brain Korea 21 project
Contract grant sponsor: Visiting Postdoctoral Fellowship of the Fund for Scientific Research-Flanders (FWO)
Contract grant sponsor: Institute for Encouragement of Innovation by means of Science and Technology

When a problem occurs in such a process, it may give rise to subtle changes that gradually grow until they become a serious operational problem. In recently developed industrial biological plants, many variables are measured in various operating units and an abundance of data is recorded. However, such data sets are highly correlated and are subject to considerable noise. In the absence of an appropriate processing method, only limited information can be extracted. In such situations, the operator may not understand the process sufficiently to maintain stable operation. If properly treated, however, this process data can provide a wealth of information that can assist plant operators in understanding the process status and enable them to take appropriate action when an abnormality is detected (Carrasco et al., 2004; Lee and Vanrolleghem, 2003, Rosen and Lennox, 2001; Saarinen et al., 2003; Xiao and Luong, 2003; Yoo et al., 2003).

Sequencing batch reactor (SBR) processes have demonstrated their efficiency and flexibility in the treatment of wastewaters with high concentrations of nutrients (nitrogen, phosphorous) and toxic compounds from domestic and industrial sources. A SBR has a unique cyclic batch operation, usually with five well-defined phases: fill, react, settle, draw, and idle. Most of the advantages of SBR processes can be attributed to their single-tank design and their ability to adjust the durations of the different phases, which endows these processes with a flexibility that allows them to meet many different treatment objectives. However, SBR processes are highly nonlinear, time-varying, and subject to significant disturbances such as hydraulic changes, composition variations, and equipment failures. Small changes in concentrations or flows may have large effects on the biological reaction kinetics, leading to batch-to-batch variations in effluent quality and microorganism growth. Such influent variations cause SBR processes to evolve over time as the microorganisms adapt to the changing operating conditions. These factors lead to changes in the microbial community and multiple operation modes within a bioreactor. Moreover, compared to data from continuous wastewater treatment processes, SBR operation data have the added dimension of batch number, which, when combined with the measured variables and sample times, gives a three-way matrix (batches $\times$ variables $\times$ time). Batch processes generally exhibit some batch-to-batch variation in the trajectories of the process variables. However, treatment performance, the key indicator of process performance, is often only examined off-line in a laboratory. This lack of real-time on-line monitoring of treatment performance means that situations can arise where operators are aware of problems in the treatment performance, but cannot determine the underlying causes of the problems or predict when they will occur. Therefore, the monitoring and supervision of SBR processes are crucial to the detection and timely correction of faults. Prompt fault identification and correction is particularly necessary in biological processes because such processes may take days, weeks, or even months to recover from an abnormal state (Lennox et al., 2001; Nomikos and MacGregor, 1994).

Generally, the fault detection and diagnosis system of an SBR is based on statistical methods and analyzes on-line data by comparing the data of a new batch to the historical data of normal and abnormal batches. If a new batch is close to the identified normal batches, it can be classified as normal, otherwise if it is close to an identified abnormal batch, it could be classified to the nearest known abnormal batch. On the other hand, if it is not close to either normal or abnormal batches, it could be classified as a new fault batch.

Multiway principal component analysis (MPCA) has been successfully applied to batch processes (Chen and Liu, 1999; Nomikos and MacGregor, 1994; Tipping and Bishop, 1997). Although the data are obtained from a process that is running stable, the biological adaptation characteristics of biological processes mean that the data may not be found in a single operating region. Biological treatment plants show different behavior patterns depending on the influent loads, temperature, and the activity of the microorganisms. Thus the models used for such plants should vary depending on the operating conditions. One approach is to represent the biological process using a suite of several models, where each model is valid only in a specific operating domain. Several studies have sought to develop methods for continuous monitoring of bioprocesses with multiple operation conditions. Chen and Liu (1999) proposed a mixture PCA model that takes advantage of PCA and heuristic smoothing clustering techniques. Xiao and Luong (2003) developed a method for the on-line monitoring of cell growth and cytotoxicity using electric cell-substrate impedance sensing to measure the concentration and time response function. Tipping and Bishop (1997) suggested a mixture of a probabilistic principal component (PC) analyzer for image compression and handwritten digit recognition, and Choi et al. (2003) applied a Gaussian mixture model to continuous process monitoring. Yoo et al. (2003) suggested nonlinear modeling and an adaptive monitoring technique for a continuous biological treatment plant with various operating regions.

In this article, we propose a local batch monitoring method with multiple probabilistic models in order to monitor a batch biological process with several operating conditions and to identify when a process has shifted to a new operating condition. First, MPCA is used to reduce the dimensionality of the data and to remove collinearity. Second, the transformed data are classified into several operation regions based on the posterior probabilities of the mixture models. Third, a discriminant measure (DM) is used to find a new operating condition that does not belong to known operating regions. Finally, the corresponding local monitoring system driven by the probabilistic knowledge is used for SBR supervision.

Materials and Methods

Multiway Principal Component Analysis

MPCA is used for the analysis and monitoring of batch process data. Batch data are typically reported in terms of

DOI 10.1002/bit

batch numbers, variables, and times. Data are arranged into a three-dimensional matrix $\mathbf{X}$ $(I \times J \times K)$, where I is the number of batches, J is the number of variables, and K is the number of times each batch is sampled. This matrix can be decomposed using various three-way techniques, one of which is MPCA. MPCA is equivalent to performing ordinary PCA on a large two-dimensional matrix $\mathbf{X}$ constructed by unfolding the three-way data in the manner shown schematically in Figure 1 (Nomikos and MacGregor, 1994).

MPCA decomposes the three-way array $\underline{\mathbf{X}}$ into a summation of the product of a score t_r and a loading matrix $\mathbf{P_r}$ plus a residual array $\mathbf{E}$ that is minimized in the least squares sense as follows:

$$\underline{\mathbf{X}} = \sum_{r=1}^{R} t_r \otimes \mathbf{P}_r + E = \sum_{r=1}^{R} t_r p_r^T + \mathbf{E} = \hat{X} + \mathbf{E} \qquad (1)$$

where $\otimes$ denotes the Kronecker product ($\underline{\mathbf{X}} = \mathbf{t} \otimes P$) is $\underline{\mathbf{X}}(i,j,k) = t(i)P(j,k)$, R denotes the number of PCs retained, t_r expresses the relationship among batches, p_r is related to variables and their time variation, and $\mathbf{E}$ is the residual matrix. The first expression in Equation 1 gives the 3-D decomposition while the second expression corresponds to the more common 2-D decomposition.

The statistics used for monitoring multivariable batch processes are Hotelling's T^2-statistic and the Q-statistic (Nomikos and MacGregor, 1994; Wise and Gallagher, 1996). The T^2-statistic is a Mahalanobis distance between new data and the center of the normal operating condition data in a reduced dimension. The pattern of the residuals is monitored using the Q-statistic, also referred to as the squared prediction error (SPE). The T^2-statistic monitors systematic variations in the PC subspace, while the Q-statistic represents variations not explained by the retained PCs. That is, faults in the process that violate the normal correlation of variables are detected in the PC subspace by the T^2-statistic, whereas faults that violate the PCA models are detected in the residual space by the Q-statistic. At end-of-batch, the T^2- and Q-statistics for batch i are calculated as follows:

$$T_i^2 = t_r^T \mathbf{S}^{-1} t_r \sim \frac{R(I^2-1)}{I(I-R)} F_{R,I-R} \qquad (2)$$

$$Q_i = e_i e_i^T = \sum_{c=1}^{KJ} \mathbf{E}(i,c)^2 \qquad (3)$$

where e_i is the ith row of $\mathbf{E}$, I is the number of batches in the reference set, t_r is a vector of R scores, $\mathbf{S}$ is the $(R \times R)$ covariance matrix of the t-scores calculated during the model development, which is diagonal due to the orthogonality of the t-score values, R is the number of PCs retained in the model, and $F_{R,I-R}$ is the F-distribution value with R and I–R–1 degrees of freedom. The statistical limits on the T^2- and Q-statistics are computed by assuming that the data conform to a multivariate normal distribution. The confidence limits of the T^2-statistic are calculated from the F-distribution. The distribution of the Q-statistic is calculated from the chi-squared distribution, $SPE_{k,\alpha} = (\nu_k/2m_k)\chi^2_{2m_k^2/\nu_k,\alpha}$, where m_k and ν_k are the mean and variance of the SPE and $\chi^2_{2m_k^2/\nu,\alpha_k}$ is the critical value of the χ^2 variable with $2m_k^2/\nu_k$ degrees of freedom at significance level α (Wise and Gallagher, 1996). For a new sample x_{new}, if $T^2_{new} < T^2_{\lim}$ and $Q_{new} < Q^2_{\lim}$, we consider a current batch to be in-control with $100(1-\alpha)\%$ confidence. Otherwise, a batch is designated as 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.

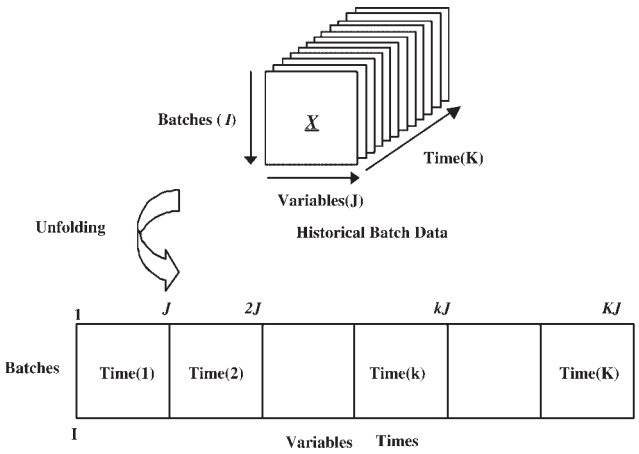

Figure 1. Multiway unfolding of a three-way batch data set.

Monitoring of Incomplete Batches

In on-line batch monitoring using MPCA, we know only the values from the start of batch to the current time; however, on-line monitoring can be achieved by augmenting the incomplete set of real batch process data with predicted data to create a data set spanning the entire batch. Nomikos and MacGregor (1995) suggested three ways to estimate variable trajectories, that is, to complete the remaining batch data: (1) zero deviation; (2) current deviation; and (3) PCA projection method. The choice of the most suitable approach depends on the characteristics of the batch process; however, the second and the third filling methods are mainly used. For on-line monitoring, the distribution of T_k^2 is approximated by an F distribution and that of SPE_k can be approximated by a weighted χ^2 distribution of $SPE_k \sim (\nu_k/2m_k)\chi^2_{2m_k^2/\nu_k}$ where m_k and ν_k are the mean and variance of SPE_k at time k. If $T_k^2 \le T^2_{\lim}$ and $\mathrm{SPE}_k \le \mathrm{SPE}_{\lim}$, the current batch at time k is in-control. Otherwise, the current batch at time instant k is deemed out of control.

Biotechnology and Bioengineering. DOI 10.1002/bit

Batch Monitoring Using Multiple MPCA Models

As noted in the Introduction section, biological processes like the SBR process under study have several operation patterns depending on the influent loads, temperature, and microorganism activity (see Fig. 4a). For biological processes with multiple operation modes, it would be appropriate to use multiple models to capture the local variations of each operating region. If data corresponding to different operating modes exhibit dissimilarities due to different batch trajectories or environmental changes, each local model can capture its operating region better than a global model, at the cost of poor characterization of the other modes. The local MPCA model, however, provides an effective method to deal with nonlinear structures in multivariate data, and thus overcomes some of the limitations of global models. Here, to enhance the monitoring performance and reduce missed alarms in the SBR process, we use multiple MPCA models within a probabilistic framework. When multiple PCA models are used, the most important issue is how to discriminate multiple operation modes.

Tipping and Bishop (1997) showed how local PCA learning can be incorporated into a Gaussian mixture modeling framework with a probabilistic learning rule. Their approach is based on a linear Gaussian model that decomposes the input space into subspaces of multivariate data. An arbitrary probability density of a sample vector $\boldsymbol{x}$, $p(\boldsymbol{x}|\boldsymbol{\theta})$, can be approximated by a mixture of basis density functions within a Bayesian framework (Bishop, 1995; Tipping and Bishop, 1997) as follows:

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = \sum_{i=1}^{k} p(\boldsymbol{x}|i)P_i \tag{4}$$

where $\boldsymbol{\theta}$ is a parameter vector whose entries are model parameters, $p(\boldsymbol{x}|k)$ is the sample density, and P_i is the prior probability that a data point is generated from component i of the mixture. The prior probabilities satisfy $\sum_k P_i = 1$ and $0 \le P_i \le 1$. Here, the model parameters of the local models with an isotropic variance are determined by probabilistic maximum likelihood estimation, that is, the so-called expectation and maximization (EM) algorithm (Tipping and Bishop, 1997). In this article, Bayesian inference modeling with a mixture of MPCA models is used to capture local subspace behaviors within the SBR process.

The diagnosis of local models is achieved by the following procedure. With the loading vectors of the local MPCA model, **P**, we can obtain the principal scores. Then, the chi-squared distance of a batch from the center of each local model is computed by

$$\chi_j^2(k) = (t_k - \overline{t}_j)\mathbf{D}_j^{-1}(t_k - \overline{t}_j)^T \tag{5}$$

where t_k is the score vector of sample x_k, $\mathbf{D}_j$ is the covariance matrix of the scores t for group j, and $\overline{t}_j$ is the group center of the principal scores. The posterior probability that sample x was sampled from the jth group can be obtained using Bayes' theorem, as follows:

$$p(\omega_j|x) = \frac{p(x|\omega_j)P_j}{\sum_{i=1}^{K} p(x|\omega_j)P_i} = \frac{P_j|\mathbf{D}_j|^{-1/2}\exp(-\chi_j^2/2)}{\sum_{j=1}^{c} P_j|\mathbf{D}_j|^{-1/2}\exp(-\chi_j^2/2)} \tag{6}$$

where $p(\omega_j|x)$ is the probability of a sample x to be class j, ω_j is the category of the jth class and P_j is the prior probability for group j. The Bayes decision rule to minimize risk calls for selecting the action that minimizes the conditional risk. Thus, to minimize the average probability of error, we should select the i that maximizes the posterior probability $p(\omega_j|x)$ (Duda et al., 2001). Therefore, in our probabilistic framework a new batch is classified into the group for which $p(\omega_j|x)$ is highest. A posterior probability in a probabilistic framework is exploited to find the most likely estimates of the true batch states by maximizing the probability of the multiple models. Probabilistic monitoring gives information on the shift in the process away from the normal steady state condition. The probabilities of the current batch for each local model provide a measure of the degree to which the current batch conditions conform to a particular local model. After selecting the corresponding jth specific model, local batch monitoring is performed using the T^2- and Q-statistics of the jth MPCA model. For this batch, if ${T_{new}}^2 < T_{j,\mathrm{lim}}^2$ and $Q_{new} < Q_{j,\mathrm{lim}}^2$, we consider the current batch to be in-control with $100(1-\alpha)\%$ confidence. Otherwise, the batch is designated as out of control.

Find a New Operating Condition

Another important issue is how to detect and diagnose a new operation condition that does not belong to the historical batches. Batch processes such as SBR processes should be checked to find a new operating condition since typical biological processes are sensitive to environmental changes. If a new batch is different from the historical known batches, this kind of a batch is not diagnosed. Fortunately, monitoring of the evolution of the posterior probabilities of the multiple models gives information that can be used to overcome this problem. As noted above, a posterior probability provides a measure of the membership of the current batch to a specific local model. In the case of an abnormal batch, the posterior probability of one particular model will suddenly increase, and the posterior probabilities of the other models will be close to zero. For a new operating condition, by contrast, the posterior probabilities of all models will gradually decrease together. Thus, the posterior probability data provide valuable information for determining whether the system has shifted to a new operating condition.

DOI 10.1002/bit

Motivated by the approach of Yoo et al. (2003), here we propose a new DM: the maximum value of the entries in the posterior probability matrix

$$p_k p_k^T,\ p_k = [p_{\omega_1|x_k}\ p_{\omega_2|x_k} \cdots p_{\omega_c|x_k}]^t.$$

$$DM_k := \arg \max_{1 \le k \le c} p_k p_k^T \tag{7}$$

For the two-model case, this DM corresponds to the inner product matrix for two posterior probabilities, p_{ik} and p_{jk}, where each element becomes each inner product (p_{ik}, p_{jk}) between the posterior probabilities (p_{ik} and p_{jk}) for models i and j, respectively. It has a mathematical form similar to that of a covariance matrix. There can be many advanced statistical rules to determine the value of the confidence limit of the DM. For example, the monitoring statistics of DM with the kernel density estimation (KDE) can be used to approximate its distribution, whereafter confidence limits can be computed. Alternatively, the confidence limit may be obtained using another assumed distribution function. For simplicity, we used a heuristic rule to determine the confidence limit in this paper, for which the specific limit depends on the number of classes, $(1 - 1/c)$. Using this definition, a value in the DM(k) at batch k below the specific limit $(1 - 1/c)$ means that the current batch cannot be allocated to the previous mixture models and has moved into a new operating condition region. That is, the system is said to have taken on a new, unknown operating condition when the posterior probabilities from all of the local models have relatively low values and DM starts to decrease.

Proposed Local Batch Monitoring Framework Using Probabilistic Multiple MPCA Models

Figure 2 shows the proposed framework of local batch monitoring using multiple models. First, the global and the multiple MPCA models with a priori operation knowledge are constructed using historical batch data. Second, transformed data in the global model are classified into several clusters using a posterior probability. Third, based on the posterior probability of each local model, the local model that best represents a current operating condition is selected and used for batch monitoring. In addition, the DM is monitored to determine whether the system has shifted to a new operating condition. If the DM indicates a new operating mode, a new model is built using recent batches. Finally, the batch at a particular time instant is on-line monitored using the jth local model and its statistical models. If the on-line batch monitoring detects a nonconforming batch using the jth local model, contribution plots of the jth local model can be used to diagnose the event so as to assign a cause.

Results and Discussion

Sequencing Batch Reactor

The data used in this research were collected from a pilot-scale SBR system with the configuration shown in Figure 3. This fill-and-draw SBR with an 80-L working volume was operated in a 6-h cycle mode where each cycle consisted of fill/anaerobic (1 h), aerobic (2h 30 min), anoxic (1 h),

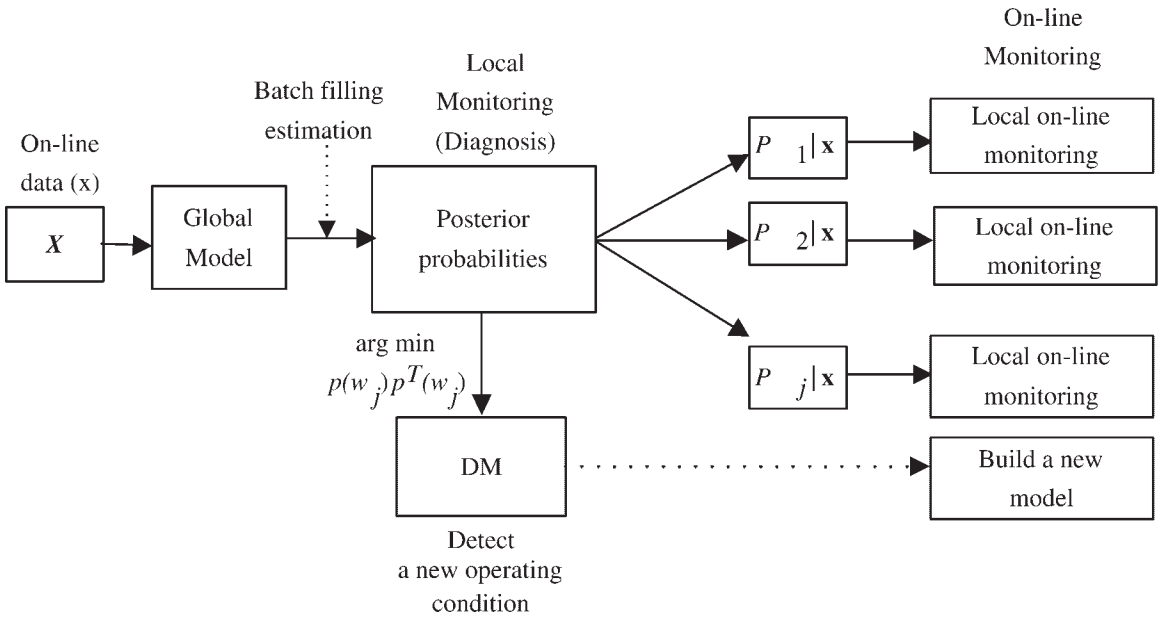

Figure 2. Framework of the multiple batch diagnosis and local monitoring method.

Biotechnology and Bioengineering. DOI 10.1002/bit

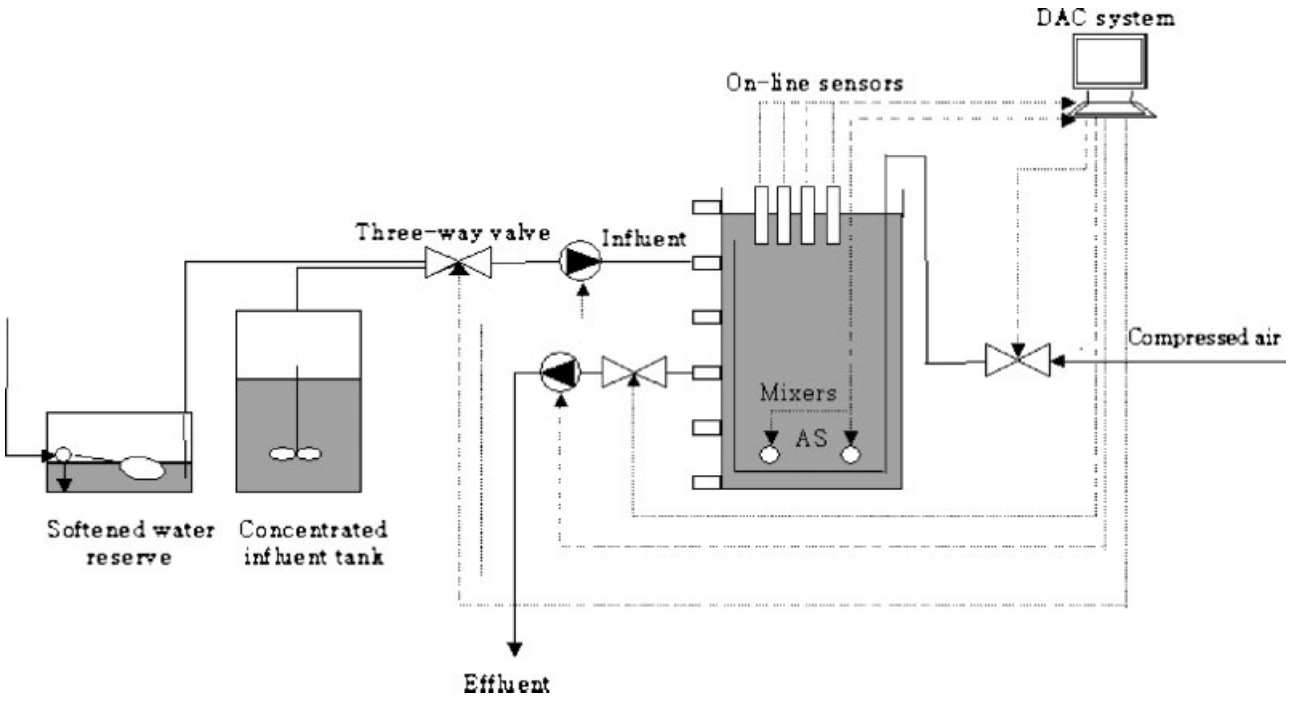

Figure 3. Schematic diagram of a pilot-scale sequencing batch reactor (SBR).

re-aerobic (30 min), and settling/draw (1 h) phases. The hydraulic retention time (HRT) and the solid retention time (SRT) were maintained at 12 h and 10 days, respectively. The loading amounts of chemical oxygen demand (COD) as synthetic municipal-like sewage, NH_4^+-N, and PO_4^{3-} per cycle in standard conditions were 440, 60, and 95 mg/L, respectively.

The duration/sequence of phases and on/off status of peristaltic pumps, mixer, and air supply were automatically controlled by a Labview data acquisition and control (DAC) system. The DAC system consisted of a computer, analog/digital interface cards, sensors, transmitters, and solid-state relays. Electrodes for measuring pH, oxidation-reduction potential, dissolved oxygen (DO), temperature, weight, and conductivity were installed and connected to the individual sensors (see Table I for a list of variables measured). It has been reported that on-line sensor values collected in SBRs are related to the dynamic characteristics of the nutrient concentrations (COD, NH_4^+-N, PO_4^{3-}, and NO_3^-) in the SBRs, where dynamic variations in the pH, ORP, and DO profiles can detect the ends of phosphate release, ammonia conversion, and phosphate uptake (Chang and Hao, 1996). In this work, a multivariate monitoring system using simple on-line sensors is used to monitor the status of the SBR process.

Table I. On-line measured variables of the SBR under study.

No.	Variables
1	Temperature (°C)
2	pH
3	Dissolved oxygen concentration (mg/L)
4	Oxidation reduction potential (mV)
5	Conductivity (mS)
6	Weight (g)

Every 1 min, a set of on-line measurements was collected, including pH, ORP, DO, conductivity, temperature, and weight of the SBR reactor which do not require advanced or expensive measurement devices. In total, 521 batches of the SBR reactor were stored in a database of historical information that represented the full dynamics of the biological process. During these 521 batches, three representative operating condition changes occurred in the SBR process. Batches 1–280 are operated with the conventional SBR strategy (operational period 1, OP1), that is, anaerobic–aerobic–anoxic–re-aerobic–settling phases. During batches 281–362, the DO control strategy of the SBR process changed with a new precise control algorithm (operational period 2, OP2). During batches 363–521, a model-based optimization scenario with a step-feed control scheme which exhibits 4 intermittent aeration sub-phases has been evaluated in a SBR (operational period 3, OP3, Sin et al., 2004). These operation changes took the form of a change in the SBR process dynamics, followed by an alteration in the microbial community, and finally a change in the sludge type (Sin et al., 2005). Only the first 300

DOI 10.1002/bit

Table II. Percent variance captured by the global MPCA model and the local MPCA models.

	Global MPCA model		Local MPCA model 1		Local MPCA model 2		Local MPCA model 3	
PC	% Variance of this PC	% Total variance	% Variance of this PC	% Total variance	% Variance of this PC	% Total variance	% Total variance	% Variance of this PC
1	75.20	75.20	26.09	26.09	38.60	38.60	29.98	29.98
2	13.02	88.22	19.05	45.14	22.66	61.26	19.13	49.11
3	6.27	94.49	17.00	62.14	17.02	78.28	13.90	63.01
4	3.68	98.17	13.22	75.36	5.89	84.18	7.45	70.46

sampling time instants of each 360 min batch were used to develop the model during normal operating conditions (NOC), since biological reactions in the settling and drawing phases (corresponding to those of the last 60 time instants) were assumed to be negligible (Lee and Vanrolleghem, 2003; Yoo et al., 2004).

SBR Operation Analysis Using Global and Multiple MPCA Models

Batch diagnosis and monitoring of all historical batches was performed using the global MPCA model and three local MPCA models. When designing an MPCA model, it is important to determine the number of PCs of the model. In determining this number, both the curse of dimensionality and loss of information are taken into account. In the present work, four PC's were used for both the global and local MPCA models to capture the batch dynamics of a SBR using a cross-validation method (Tipping and Bishop, 1997; Yoo et al., 2004). The results of the global and local MPCA models are summarized in Table II. The explained variances cannot be compared directly because they are based on their own means and variances. Figure 4a shows the score plot of all 521 batches of the SBR in the PC_1–PC_2 plane using the global and local MPCA models with 99% confidence limits of the global MPCA model and three ellipsoids of 99% confidence limits for each local operating condition. As shown in Figure 4a, since batches with similar operation conditions tend to cluster together in distinct regions in the reduced space of global MPCA model, it is possible to analyze the batch operation history in the reduced dimensional space. Figure 4b shows the T^2 and SPE plots obtained using the global MPCA model.

From Figure 4a it can be inferred that the first group in the score plot represents the batches of the first operational period (OP1). The second group (group 2) represents the operating region during the second operational period (OP2) and the third group (group 3) comprises the observations during the third operational period (OP3).Given these observations, a single global MPCA model may not be valid for a time-varying SBR process of this type. Because the SBR process exhibits several operation modes, the confidence limits of the single global MPCA model would need to be widened, leading to less sensitive monitoring performance.

Local Probabilistic Batch Monitoring of SBR Using Multiple MPCA Models

In order to represent the distinct groups that characterize the process dynamics of a process undergoing multiple variations, the proposed local monitoring method with multiple MPCA models is used. Figure 5a–c shows the posterior probabilities of three local MPCA models, where the posterior probabilities of the local models indicate the extent to which the current batch belongs to the respective local models. Figure 5d shows the DM, which is used to discriminate a change to a new process from the previous local models.

During batches 1–280, the posterior probability values for the first local model (Fig. 5a) are large compared to those of the other models, which are almost zero, indicating that the batches during this period can be assigned to the first local model. Moreover, with the exception of two troughs, the DM value remains above the set limit during this period, confirming that the first local model can describe the data well. The batches whose DM values fall below the confidence limit cannot be represented by any of the three local models; thus the periods in which these batches occur correspond to an unknown operating condition. Consistent with this, the posterior probability values of the first local model are relatively low during these periods. The three troughs in the DM values coincide with external disturbances. Detailed inspection of the process data revealed that the first, at batch 80, was caused by a drain problem of the effluent pump, the second, spanning batches 90–92, was caused by an abnormal feeding by the influent pump and the third, spanning batches 163–164, was caused by an abrupt increase in the air flow rate to the SBR. During the periods affected by these disturbances, the performance of the SBR process was unstable and the effluent quality was quite poor. Because the microorganisms of the SBR process could control these external disturbances, the SBR process returned to the normal operation region in each instance.

During batches 281–362, the posterior probability values of the second local model are large compared to those of the other models, which are close to zero. This shift coincides with the batch at which the DO control strategy of the SBR

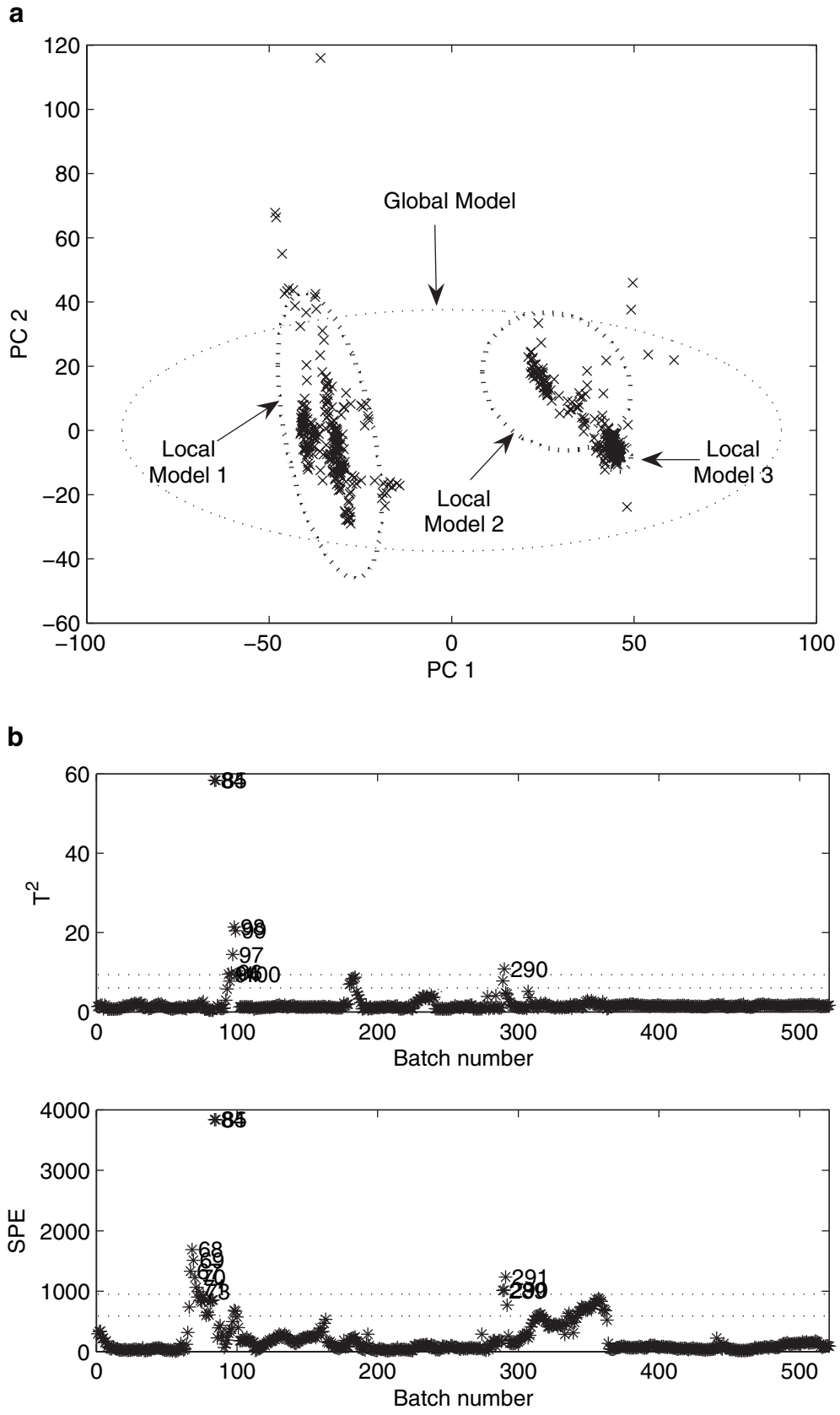

Figure 4. Batch diagnosis and monitoring of all historical batches using a global model: (**a**) score plot in the PC_1-PC_2 plane, and (**b**) T^2 and SPE plots.

DOI 10.1002/bit

process was changed (batch 281). In the period immediately after the introduction of the new DO control strategy (batches 281–285), the DM values fall below the confidence limit, presumably because the microorganisms of the SBR were adapting to the new operation strategy during those batches. These findings therefore confirm both the DM's ability to immediately detect a change in the DO control strategy as well as the microorganism's ability to adapt and bring the SBR to a new steady state.

After batch 360, the posterior probability values of the third local model (Fig. 5c) are almost unity, whereas those of the other two local models are almost zero. Thus, these batches can be assigned to the third local model. The DM values are below the confidence limit for batch 363–370, indicating that a change to a new operation regime occurred in this interval. This observation coincides perfectly with the implementation of a new operational step-feed strategy with 4 intermittent aeration sub-phases at batch 363, resulting from a model-based optimization. This operation change improved the nitrogen and phosphorus treatment performances of the SBR, and affected all process conditions after this batch, leading to substantial changes to the process. At around batch 445 the DM value show one dip representing an abnormal batch caused by an abnormal continuous feeding of influent pumps, where the process seems to shift briefly from the third local model to the second one. These results confirm that the proposed method can efficiently detect local process changes using the mixture models. It is therefore an effective technique for extracting information related to changes in unknown process operating conditions as well as for localizing multiple process disturbances.

Figure 6 shows the T^2 and SPE charts obtained from the global and local MPCA models during the third operation period (batches 363–521). The T^2 and SPE data from the global MPCA model do not indicate the presence of any abnormal batches during this operation period, whereas the T^2 and SPE charts of the corresponding local model detect the eight non-conforming batches: 363, 364, 367, 368, 369, 442, 457, and 516 (batch numbers 1,2,4,5,6,79,84,153 in Fig. 6). Although the single global model could be used to model the data in the subspace by MPCA, it would ignore

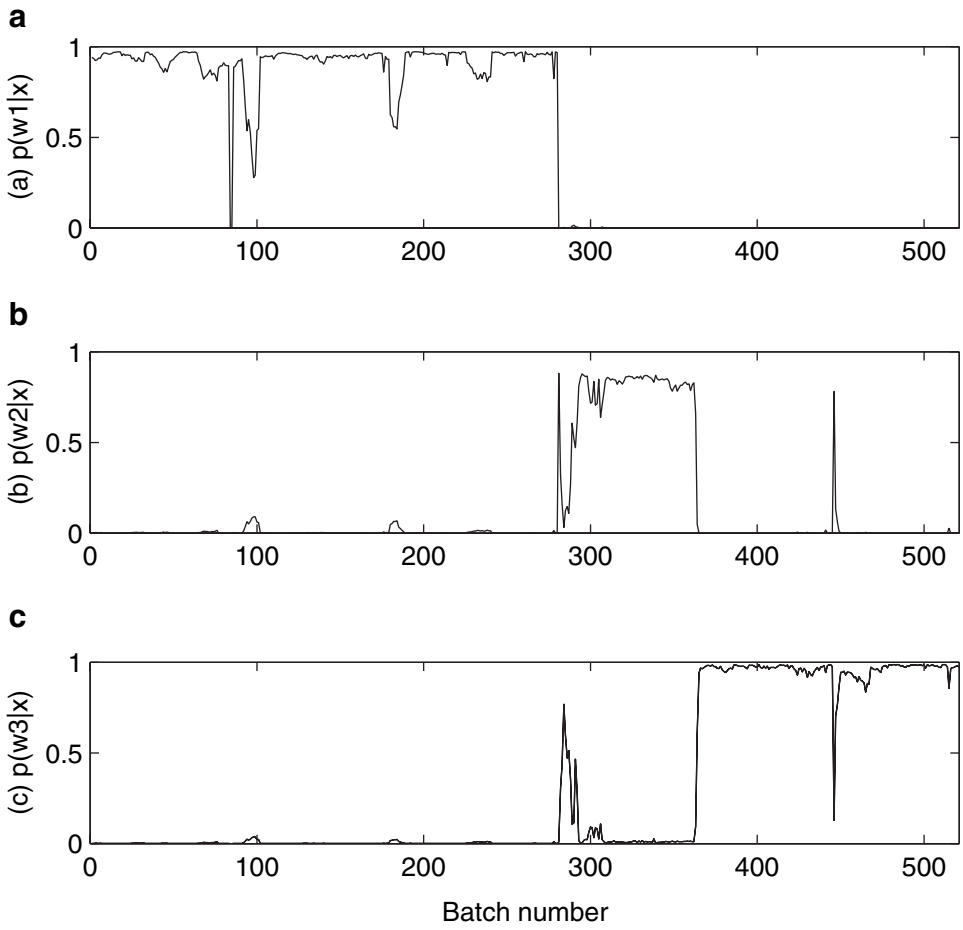

Figure 5. Local probabilistic monitoring with multiple MPCA models. **a–c**: Local-model monitoring of the SBR using three posterior probabilities; and (**d**) discrimination measure (DM) for the detection of operation mode changes.

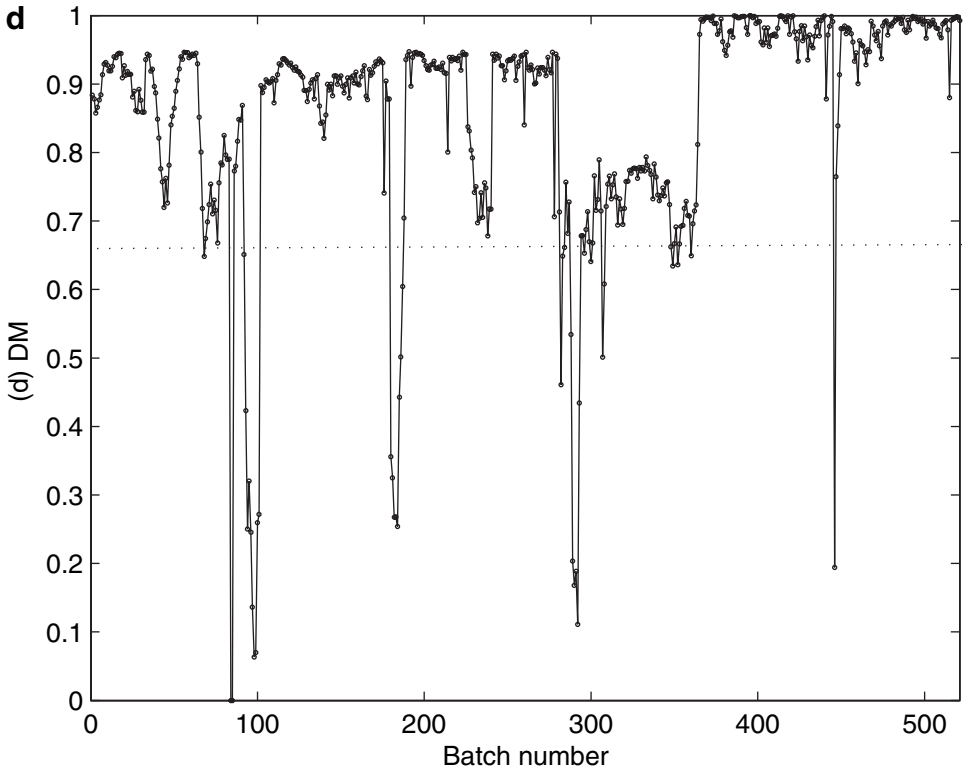

Figure 5. (*Continued*)

the local distributed characteristics of each local operational region. In the single global MPCA model, the symptoms of abnormal batches are buried in a confidence region that is too broad to catch such batches. On the other hand, local batch monitoring using the appropriate local model shows better monitoring performance and increases the reliability of the monitoring system because it is capable of automatically extracting the key components that represent the kinetics of the biological reaction leading to batch-to-batch variations in the treatment performance and the microorganism community.

On-Line Monitoring of an Abnormal Batch Using the Local MPCA

Figure 7 shows the on-line batch monitoring results (T^2 and SPE charts) of the global model and the third local model for the *abnormal* batch 84 identified in Figure 6b. For on-line batch monitoring, the current deviation is used as a filling method; in this approach, all future measurements are filled-in with their current deviation from the average batch. As shown in Figure 7a, the monitoring charts of the global MPCA model do not detect this abnormal batch. By contrast, the T^2 chart of the corresponding local model do detect the abnormal situation (Fig. 7b). Therefore, during on-line batch monitoring, the local probabilistic model can detect abnormalities that the global model fails to detect. It is interesting to note that the confidence limit of the SPE chart of the global MPCA moves up and down two times, at the around the 60th and 250th sampling times. These times correspond to the end of the filling phase and the start of the second aerobic phase of the old operation strategy (anaerobic, 1st aerobic, anoxic, and 2nd aerobic phases). On the other hand, the confidence limit of the SPE chart of the local MPCA model moves up and down five times because the operating conditions of this batch have a unique cyclic batch operation (anaerobic, 4 alternating phases of aerobic and anoxic and final aerobic phases), as shown in Figure 4c. These findings illustrate that the local model approach can better extract the dynamic characteristics, the biological phenomena and their relationships during the current operation of the SBR.

On-line monitoring charts only detect non-conforming batches; contribution plots must then be used to diagnose the detected events so as to assign a cause. Contribution plots indicate which variables are predominantly responsible for the deviations from the normal batch behavior. Figure 8 shows the contribution plots for the T^2 chart of the third local model at 230 h of batch number 84. The deviation of T^2 comes from the pH (variable 2), conductivity (variable 5), and weight (variable 6). To examine this in greater detail, Figure 9 shows univariate plots of the temperature, pH, DO, ORP, conductivity, and weight for normal batches and for

DOI 10.1002/bit

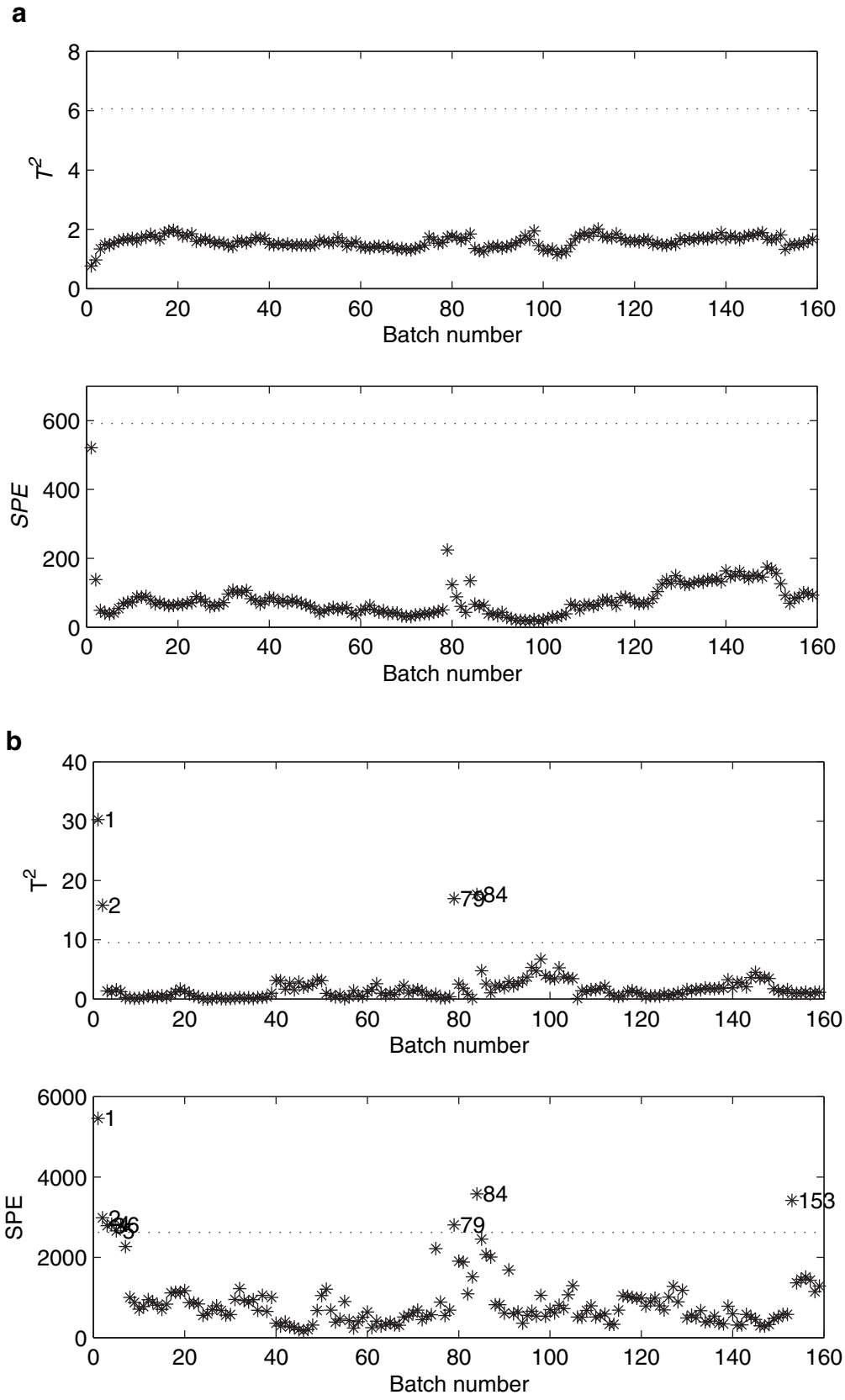

Figure 6. Batch monitoring charts of the third operation mode using (**a**) the global model (**b**) the corresponding local model. The dotted lines correspond to the 99% confidence limit.

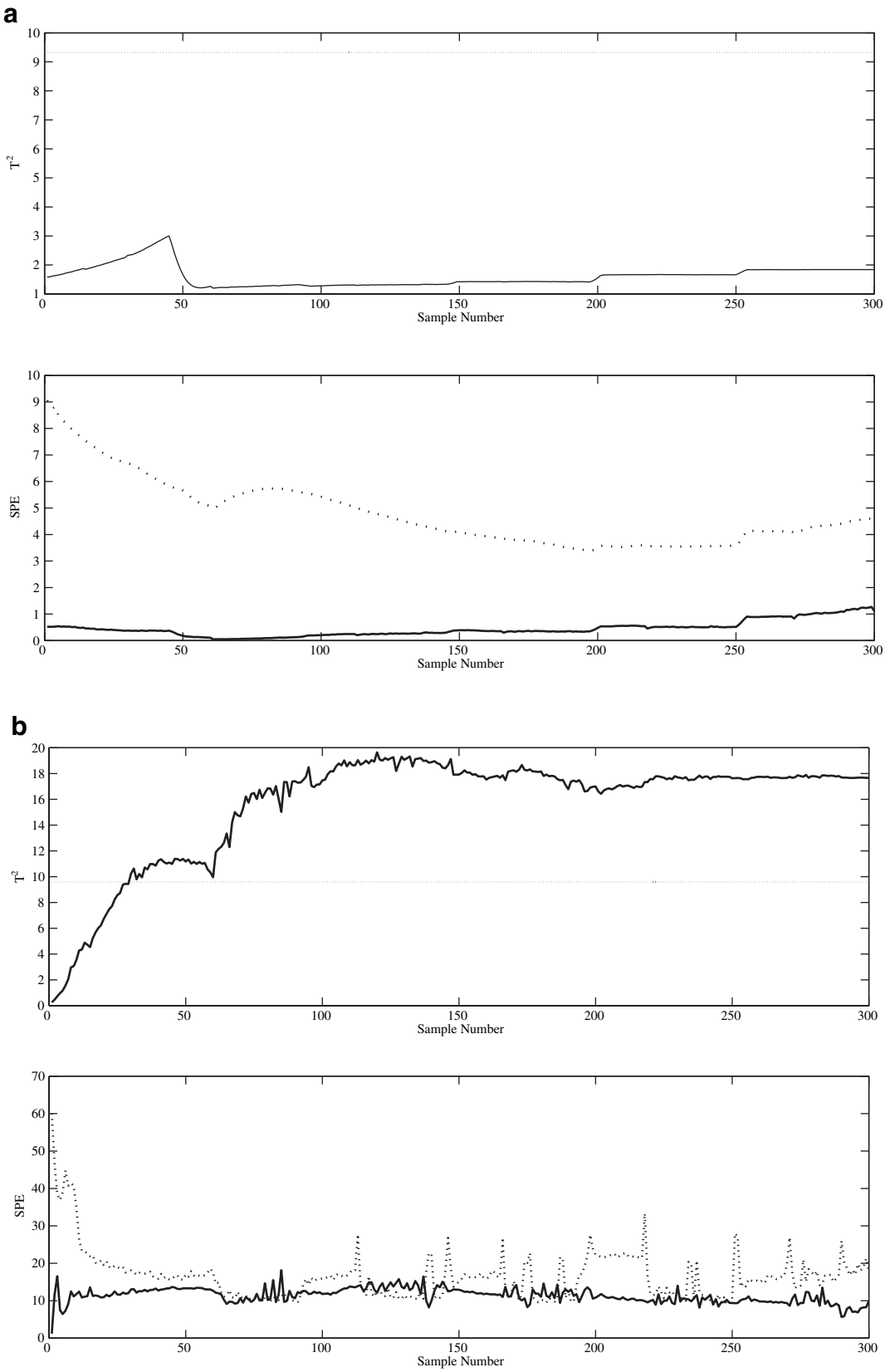

Figure 7. On-line monitoring of charts for *abnormal* batch number 84 among the third operation mode batches using (**a**) the global model (**b**) the corresponding local model. The dotted lines correspond to the 99% confidence limit.

DOI 10.1002/bit

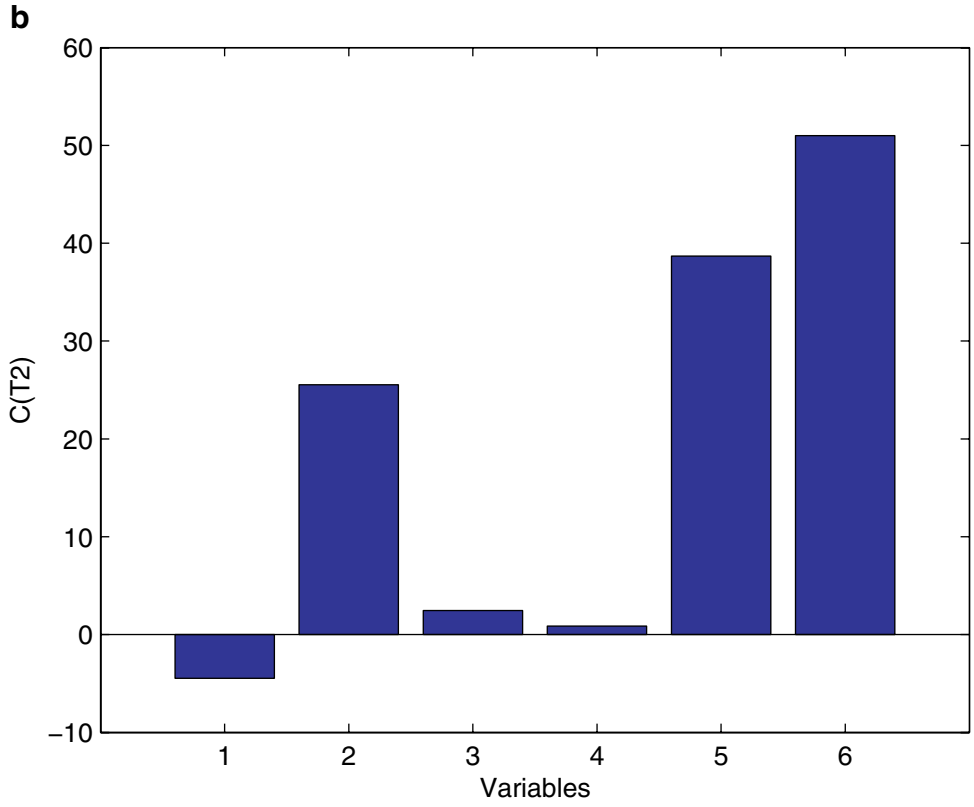

Figure 8. Contribution plots for the T^2-statistics at sample 230 of batch number 84 among the third operation mode batches. [Color figure can be seen in the online version of this article, available at www.interscience.wiley.com.]

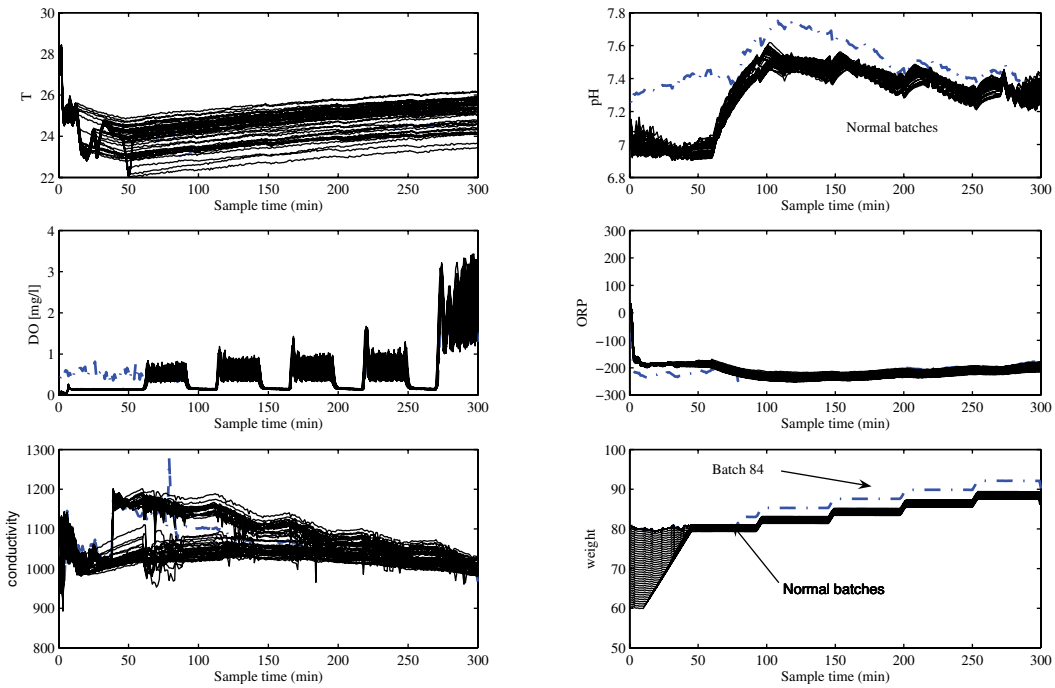

Figure 9. Univariate plots of six on-line SBR variables for normal batches and batch number 84 among the third operation mode batches.

batch number 87. The univariate plots confirm the findings indicated by the contribution plots. Specifically, the weight is the main variable contributing to this abnormal batch, consistent with the nature of the malfunction: an accidental influent feeding to the SBR process occurred due to an influent pump malfunction. Thus, it can be inferred that too much feeding of the carbon source (external disturbance) affects other variables, especially conductivity and pH. The T^2 chart of the local model in Figure 7b detects the abnormal carbon source feeding as soon as it occurs. Thus the local-model based method can give more reliable and higher resolution monitoring results. We note here that in this paper our research was constrained to the application of the multi-model approach to a data set that was also used for calibration of this specific model. Future research is therefore warranted to check this approach against new data stemming from operational regions that are presented in the historical data set but not used in the calibration step, that is, a true validation of the method. Also, on-line use of our approach warrants the search for strategies to detect and identify new operations in case they are yet not present in the historical data, that is, methods to tackle contingency problems.

Conclusions

In this work, a new local batch monitoring method based on multiple probabilistic models has been proposed and applied to an 80-L SBR. Under this multiple modeling approach, the entire operation data of the SBR is divided into distinct regions, each of which is modeled separately. Then, these local regions can be supervised separately, leading to more effective batch monitoring. The method developed in this study was shown to be a valuable tool for supervising a biological batch process with multiple operating conditions. Specifically, our local model-based method was able to detect process disturbances within a SBR that were not detected using a global model. Moreover, it was able to localize the cause of an abnormal batch and gave a much more sensitive and clear indication of the process disturbances within the SBR than did the global model. The superiority of the local model-based approach derives from its ability to identify multiple operating conditions and capture the biological relationship among batch process variables. The present results indicate that the proposed local batch monitoring approach may be a good alternative to a global modeling approach for processes with multiple operation conditions such as biological processes. Since model-based optimization and control may lead to changes in the microbial community or shift the system dynamics, future studies should endeavor to incorporate microbial community information into the process monitoring and control method.

Nomenclature

E	residual matrices
F	F distribution
$p(\mathrm{x}\|\boldsymbol{\theta})$	a probability density of a sample vector x
$p(\mathrm{x}\|k)$	a component density generated from component k of the mixture
$p(\omega_j\|\mathrm{x})$	the posterior probability of sample x from group j
P_i	the prior probability of the samples
$\mathbf{P}$	loading matrix
Q	Q-statistic
Q_{lim}	confidence limit for Q-statistic
$\mathbf{T}$	score matrix
T^2	Hotelling's T^2-statistic
T^2_{lim}	confidence limit for Hotelling's T^2-statistic
$\underline{\mathbf{X}}$	three-way batch data array
$\mathbf{X}$	input data matrix
ω_j	the category of the jth class

Greek Letters

$\chi^2_j(\mathrm{k})$	the chi-squared distance of the jth sample from the center of each local model

This work was supported by the Second-Phase of BK(Brain Korea) 21 project, the Visiting Postdoctoral Fellowship of the Fund for Scientific Research-Flanders (FWO) and the Institute for Encouragement of Innovation by means of Science and Technology in Flanders (IWT). Peter Vanrolleghem is Canada Research Chair in Water Quality Modelling. The authors thank Gürkan Sin, Ingmar Nopens, Ruxandra Govereanu, Dae Sung Lee, and Griet Schelstraete who greatly contributed to the BIOMATH SBR research.

References

Bishop CM. 1995. Neural networks for pattern recognition. New York: Clarendon Press.

Carrasco EF, Omil F, Garrido JM, Arrojo B, Mendez R. 2004. Advanced monitoring and supervision of biological treatment of complex dairy effluents in a full-scale plant. Biotechnol Prog 20:992–997.

Chang CH, Hao OJ. 1996. Sequencing batch reactor system for nutrient removal: ORP and pH profiles. J Chem Tech Biotechnol 67:27–38.

Chen J, Liu J. 1999. Mixture principal component analysis models for process monitoring. Ind Eng Chem Res 38:1478–1488.

Choi SW, Park JH, Lee IB. 2003. Process monitoring using a Gaussian mixture model via principal component analysis and discriminant analysis. Comp Chem Eng 28:1377–1387.

Duda RO, Hart PE, Stork DG. 2001. Pattern classification, 2nd edn. New York: John Wiley & Sons.

Lee DS, Vanrolleghem PA. 2003. Monitoring of a sequencing batch reactor using adaptive multiblock principal component analysis. Biotechnol Bioeng 82:489–497.

Lennox B, Montague GA, Hiden HG, Kornfeld G, Goulding PR. 2001. Process monitoring of an industrial fed-batch fermentation. Biotechnol Bioeng 74(2):125–135.

Nomikos P, MacGregor JF. 1994. Monitoring batch processes using multiway principal component analysis. AIChE J 40(8):1361–1375.

Nomikos P, MacGregor JF. 1995. Multivariate SPC charts for monitoring batch processes. Technometrics 37:41–59.

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

Saarinen MA, Reece JS, Arnold MA, Murhammer DW. 2003. Monitoring and controlling the dissolved oxygen (DO) concentration within the high aspect ratio vessel (HARV). Biotechnol Prog 19(4):1335–1341.

Sin G, Insel G, Lee DS, Vanrolleghem PA. 2004. Optimal but robust N and P removal in SBR's: A model-based systematic study of operation scenarios. Wat Sci Tech 50(10):97–105.

DOI 10.1002/bit

Sin G, Govoreanu R, Boon N, Schelstraete G, Vanrolleghem PA. 2005. Evaluation of the impacts of model-based operation of SBRs on activated sludge microbial community. Wat Sci Tech 54(1):157–166.

Tipping ME, Bishop CM. 1997. Mixtures of probabilistic principal component analysers. Technical report, Neural Computing Research Group. Aston University, UK.

Wise BM, Gallagher NB. 1996. The Process chemometrics approach to process monitoring and fault detection. J Process Control 6:329–348.

Xiao C, Luong JHT. 2003. On-line monitoring of cell growth and cytotoxicity using electric cell-substrate impedance sensing (ECIS). Biotechnol Prog 19(3):1000–1005.

Yoo CK, Vanrolleghem PA, Lee I. 2003. Nonlinear modeling and adaptive monitoring with fuzzy and multivariate statistical method in a biological wastewater treatment plant. J Biotech 105(1–2):135–163.

Yoo CK, Lee DS, Vanrolleghem PA. 2004. Application of multiway ICA for on-line monitoring of a sequencing batch reactor. Wat Res 38(7):1715–1732.