KV_C004
Ruiz, M., Villez, K., Sin, G., Colomer, J., Vanrolleghem, P. A. (2006). Influence of scaling and unfolding in PCA based monitoring of nutrient removing batch process. Proceedings 6th IFAC Symposium on Fault Detection, Supervision and Safety of Technical Processes (SafeProcess2006), Beijing, China, August 29-September 1, 2006, 114-119.

ELSEVIER IFAC PUBLICATIONS

INFLUENCE OF SCALING AND UNFOLDING IN PCA BASED MONITORING OF NUTRIENT REMOVING BATCH PROCESS

Magda Ruiz[1]*, Kris Villez[2], Gurkan Sin[2], Joan Colomer[1] and Peter Vanrrolleghem[2]

[1]Control Engineering and Intelligent Systems Group –eXiT,
Department of Electronics, Computer Science and Automatic Control,
University of Girona, Campus Montilivi CP 17071 Building PIV, Girona - Spain
[2]Department of Applied Mathematics, Biometrics and Process Control
Ghent University, Coupure Links 653, B-9000, Ghent – Belgium

Abstract: The data set of batch biological and biotechnological processes can be organized in a three-way data matrix. In this paper the usefulness of different PCA approaches for monitoring is analyzed. Different ways of unfolding and scaling of data have been applied to a pilot-scale SBR data. PCA is used to reduce the dimensionality and to remove the non-linearity dynamic of the data. Moreover, a new method to select the number of principal components is proposed. Loadings graphics are used to determinate the predominant variables for each one. The results show that whatever model can be applied depending on the goal of the monitoring, however the models implicate possible false alarms or faults omission. *Copyright © 2006 IFAC*

Keywords: Biological Process, Process monitoring, Multiway Principal Component Analysis (MPCA).

1. INTRODUCTION

Biological and biotechnological processes, including wastewater treatment plants (WWTP), are characterized by time-varying trajectories of all process measured variables. These process variables are physical and biological and they occur simultaneously. In addition to that, they contain valuable information that can be used to analyze the process behaviour establishing operation limits. With these limits, it is possible to detect faults or upsets within the process. Nowadays in the bibliography, the MSPC approach, based on Principal Component Analysis (PCA), is implemented in biological and biotechnological processes to monitor the performance of a process in order to detect faults that may occur and to identify or to diagnose the problem. PCA is the most widely used data-driven technique for monitoring in batch process which has as its objective the explanation of the variance-covariance structure of a multivariate dataset through a few linear combinations of the original variables with special properties in terms of variances (Nomikos, *et* al., 1994). In this work, the main objective is to compare and to discuss the results obtained using different PCA approaches for monitoring and diagnosis of a pilot-scale WWTP of type SBR. In Section 2, materials and methods implement to compare the PCA models is explained. In Section 3 the results will show. For each one model, the results comprise three parts: 1) Linearity study for each unfolding using normal probability plots. 2) Procedure to select the correspondent number of principal components for each one model. 3) Inspection of the models and validation. Finally, the conclusion are discussed

2. MATERIALS AND METHODS

2.1 Pilot-scale Sequencing Batch Reactor (SBR) WWTP

The data used in this paper were collected from a pilot-scale SBR system with a working volume of *80L* (Fig. 1) which is operated in a cycle of 6h (4 cycles per day). The fill phase start at t=0 – 60 min. The first aerobic phase starts at t=61 -210 min. The anoxic phase starts at t=210 – 270 min and the second aerobic phase starts at t=271 – 300 min. Settling phase occurs at t=301 – 345 min and finally draw phase takes the last 15 min. The excess sludge is waste from the end of the second aerobic phase for each cycle. The control of the duration/sequence of phases and on/off of peristaltic pumps, mixer and air

supply are automatically achieved by a Labview data acquisition and control (DAC) system.

The DAC system consists of computer, analog/digital interface cards, sensors, transmitters and solid state relays (SSR). Electrodes for pH, Oxidation-Reduction Potential (ORP), Dissolved Oxygen (DO), temperature, weight and conductivity are installed and connected to the individual sensors (6 process variables are measured). The status of the reactor is displayed on the computer and the time series of the electrode signals are stored in a data log-file (Sin, 2004). A set of measurements is obtained every 1 min (360 times points per cycle). These measurements were stored for 13 months which compose the database of historical information. Only 300 sampling are used to develop the models because biological reactions in settling and drawing phases are assumed as unimportant, these are the last 60 time instants.

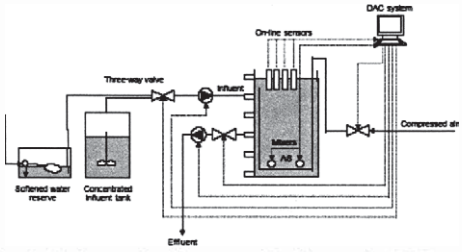

Fig. 1. Schematic diagram of pilot-scale SBR

2.2. Multiway Principal Component Analysis

Consider a typical batch run in which j=1,2,...,J variables are recorded at k=1,2,...,K time instants throughout the batch. Similar data is a generated for a number of such batch runs i=1,2,...,I. This results in an three-way array $\underline{\mathbf{X}}$ (IxJxK) array as illustrated in Fig 2, where the height gives the number of batches, I, the width gives the number of measurements (sensors), J, and the length gives the number of time instants, K. Each horizontal slice of this array is a (J x K) data matrix representing the time histories or trajectories for all variables of a single batch, i. Each vertical slice along the i,j-direction is an IxJ matrix representing the values of all the variables for all batches at a common time interval (k) (Nomikos and MacGregor, 1994). Similarly, each vertical slice along the i,k-direction represents the data of one sensor for all batches and all time instants.

The scaling batch data is select before using the multiway methods. Several authors recognize two options to make scaling; these are: autoscaling (in which the mean trajectories are removed and each column has equal variance) and variable scaling or group scaling (in which the mean trajectories are removed and each variable has equal variance). When the scaling data decision is ready MPCA can be performed to ordinary PCA on a large two-dimensional (2-D) matrix constructed by unfolding the three-way matrix. Six possible ways of unfolding the three-way data matrix $\underline{\mathbf{X}}$ are indicated in Table 1, as suggested by Westerhuis. *et* al. (1999). When aiming at PCA-based monitoring, unfolding types **B** and **D** will lead to models that are equivalent to the models constructed on the C-, respectively E-unfolded matrices. Matrix **F** is the transpose of **A**, and a PCA would just switch the scores and loadings of the two matrices if no centring or scaling is applied.

The unfolding used by Nomikos and MacGregor is of type D. This is straightforward for analysis of historical data and monitoring in batch process because subtracting the mean of each column of the matrix **X** removes the main nonlinear and dynamic components in the data. Nevertheless, batch-wise unfolding (type D and E) present a problem for monitoring in real time since the new batch is incomplete during the progress of the batch (Nomikos, *et* al., 1994). Nomikos and MacGregor suggest 3 ways to overcome the problem of incomplete batches in (Nomikos, *et* al., 1995), while not changing the unfolding type. Alternatively, Wold suggests a variable-wise unfolded PLS approach, which does not require complete batches in (Wold, *et* al., 1987). Applications of batch-wise unfolding (type D or E) in biological batch processes can be found in Sung Lee, *et* al. (2005), Sung Lee, *et* al. (2003), and Wold, *et* al., (1998).

Table 1 Types of unfolding a three way data matrix

Type	*Structure**	*Direction***
A	IK x J	variable
B	JI x K	time
C	IJ x K	time
D	I x KJ	batch
E	I x JK	batch
F	J x IK	variable

*Structure of the unfolding matrix

**Direction that remains unaltered

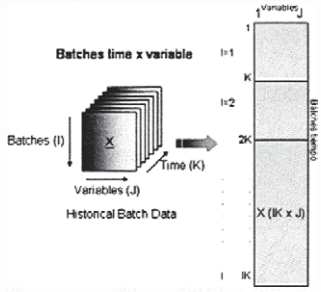

Fig. 2. Decomposition of X to 2-D (IK x J)

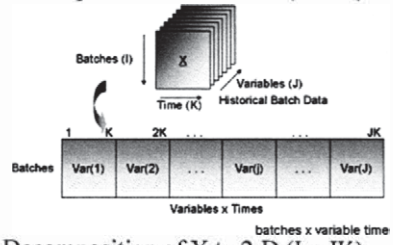

Fig. 3. Decomposition of X to 2-D (I x JK)

In this work the ways A (Fig. 2) and E (Fig. 3) are used. Type E was chosen instead ot the mathematically equivalent type D, for easiness of interpretation. The goal in MPCA is to decompose the three-way $\underline{\mathbf{X}}$, into a large two-dimensional matrix **X** separating the data in an optimal way into two parts: The noise or residual part (E), which is small in the sense of least squares, and the systematic part

($\sum_{r=1}^{R} t_r \otimes P_r$), which consists of a first factor (t) related to the batches and a second factor (P) related to the variables and their time variation (Nomikos, *et* al., 1994).

MPCA is performed by means of the NIPALS algorithm resulting in the matrix X. It is the product of the score vector tr and the loading matrices P_r, plus a residual matrix E, that is minimized in the sense of least-squares:

$$\underline{X} = \sum_{r=1}^{R} t_r \otimes P_r \qquad (1)$$

$$X = \sum_{r=1}^{R} t_r P_r^T + E = \hat{X} + E \qquad (2)$$

where $\otimes$ denotes the Kronecker product ($\underline{X} = t \otimes P$ is $\underline{X}(i,j,k) = t(i)P(j,k)$) and R denotes the number of principal components retained. The score matrices and loading matrices for type A and E have different sizes as show in 0. Equation (1) is the 3-D decomposition while Equation (2) displays the more common 2-D decomposition (Undey, *et* al., 2002).

Table 2 Size of matrices T and P for types A and E

Matrix	*Type A*	*Type E*
T	(*IK* x *R*)	(*I* x *R*)
P	(*J* x *R*)	(*JK* x *R*)

Abnormal behaviour of a process in batch direction is generally identified by means of the Q-statistic or the D-statistic, which are compared with control limits determining whether the process is in control or not. These methods are based on the assumption (generally motivated with the central limit theorem) that the underlying process follows approximately a multivariate normal distribution where the first moment vector is zero. The Q-statistic is a measure of the distance of the observation on the reduced space to the center of that model. For batch number i, Q_i is defined as

$$Q_i = \sum_{j=1}^{J} \sum_{k=1}^{K} \left(e_{jk}\right)^2 \qquad (3)$$

where e_{jk} are the elements of residual matrix E. Q_i indicates the distance between the actual values of the batch and the projected values onto the reduced space. The D-statistic or Hotelling T^2 statistic, measures the distance of the projection of the observation on the reduced space to the center of it:

$$D_i = t_i^T S^{-1} t_i \qquad (4)$$

where S is the estimated covariance matrix.

The statistical limits for the variable-wise unfolded models are calculated separately for each time instant K and each score. Practically the resulting scores are reorganized so that the scores of one batch form one row vector in a matrix X_T. This matrix has I rows (one per batch) and $T.K$ columns (one per variable per time). This transformed matrix X_T is used to derive confidence limits for each time instant and each score by estimation of their respective means and standard deviations. Thus, for each score an average trajectory for its upper and lower tolerance limits can be obtained. Residuals are then calculated for each time instant as:

$$e_{new} = x_{new} - t_{new} P'_{model} \qquad (5)$$

The standard deviation of these residuals RSD RSD is a measure of the difference of the new batch to the model which is calculates as follows.

$$RSD = DModX = \frac{e_{new} e'_{new}}{(M - R)} \qquad (6)$$

here R is the number of components in the model and M is the number of columns in X.

2.3 Methodology

1959 complete batches were run between from December 16th and 2003 to July 18th of 2005. Each of these batches resulted in 6 different trajectories of 300 samples each, being the weight of the reactor and the temperature, the pH, the DO, the ORP and the conductivity in the reactor. The approach taken to compare the discussed options in PCA-based monitoring is explained in the following paragraphs:

Step 1: The three-way data matrix is first unfolded in batch direction (type E). As there was no sufficient detailed knowledge available about the process behaviour or the type of faults that may have occurred,. the data **X**(*I* x *JK*) were normalized using autoscaling as suggested by MacGregor and Kourti (Westerhuis, *et* al., 1999) to construct an MPCA model for data screening. By means of the resulting MPCA-model ([number] PC's), 248 batches were identified as being abnormal and were thus excluded from the data set for future use.

Step 2: The 1711 left-over batches were used to compare different unfolding and scaling approaches to PCA-based monitoring of the pilot-scale SBR This size of data base (6x300x1711) is divided in order to develop the models with 80% of the total data with a three way data matrix of (6x300x1369) and 20% to validate the models with a size of (6x300x342). Two options in regard to the unfolding are available when considering batch process monitoring. One is to unfold the data in batch direction (type D or E) or unfold the data into the variable direction (type A). For both options; the data **X**(*I* x *JK*) are again normalized using three different options: Autoscaling (auto): the mean and standard deviation of each variable are calculated at each time in the batch over all batches. Variable scaling or group scaling (grps): the mean is calculated at each time in the batch over all batches and one standard deviation per variable is calculated over all batches. Continuous scaling (cs): It was applied by Wold (Wold, *et* al., 1987). One mean and one standard deviation per variable are calculated over all batches. In total six models for SBR monitoring are thus generated. Table 3 summarizes how the respective models were labelled in the framework of our research. Before making further inferences, the normal probability plot of the first score is made. This allows a visual inspection of validity of one of the assumptions in PCA modelling (i.e. that the

scores exhibit gaussian distributions). Immediately after, a determination's study of the number of principal components and analysis of the models are made.

Table 3 Names for each model developed

	Type A	*Type E*
auto	Model 1	Model 4
grps	Model 2	Model 5
cs	Model 3	Model 6

3 RESULTS AND DISCUSSION

3.1 Normal probability plot

Before going into detailed model comparison, it is useful to evaluate if the extent of linearization of the original data –typically non-linear and dynamic, by different combinations of unfolding and scaling approaches satisfies the assumption for the normal lineal model and if they have an approximately normal distribution. Two hypotheses have been achieved in accord with Giudici 2003. The first one state: the value of the response variable is a linear combination of the explanatory variables. This is plainly stated for the linear combinations of PCA. The second hypothesis is about the data set. If the data comes from populations with normal distribution, these could be tested using the Measures of Kurtosis and the Quantile-Quantile plot (q-q plot). Figure 4 shows the q-q plots for the first PC of all models, organized as in Table 3 (models 1 to 6 from top to bottom and from left to right). If the values of the first PC come from a normal distribution, then the plots should appear as a linear curve.

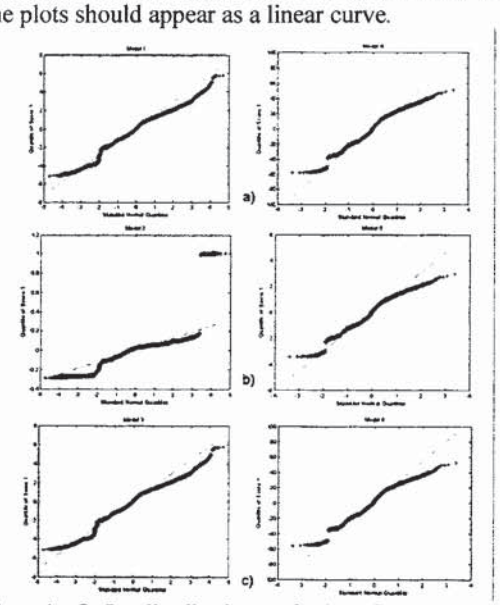

Fig. 4 Q-Q distribution of the first principal component for models which are unfolded variable-wise (left-side) and batch-wise (right-side) and scaled with a) Auto b) Group c) Continuous approaches

The plots show that the data comes from a normal distribution. The plots are approximate linear for all models except for model 2. This non-linearlities correspond to two batches which are removed. It should be noted that none of the 6 combinations unfolding type and scaling type, has lead to a acceptable removal of non-linearities in the data set. This is clearly visible at the left hand side of the q-q plot that the variable-wise unfolded models show a gap in the plot while the batch-wise unfolded models show a bump in the plot in the same region. Still, these non-linearities were not considered to be extreme violations of the assumptions on linearity; however, some errors or omissions could be expected when the process monitoring will be developed. When faced with extreme non-linearities, process monitoring may be improved by applying non-linear methods as Kernel PCA (Lee, *et* al, 2004).

3.2 Determination of the number of principal components using the contribution plots

A critical step in PCA modelling is the determination of the number of principal components to be retained in the model. Qin, *et* al., (2000) and Al-Kandari, *et* al., (2005) elaborate on this subject in the framework of PCA-based sensor validation and reconstruction. In this paper, a new method will show and evaluate an alternative based on the loading plots of the principal components. The selected principal components are limited to a number of components that capture all the present variables. The principal components are ordered along their captured variance (equivalent to ordering by their eigenvalues) from high to low and are evaluated in that order. When the dominating variables of the considered principal component are already dominating in the yet retained components, this and all following components are omitted. To illustrate this method, the selection in the case of model 4 is explained in detail here.

Fig. 5 shows the contribution plots of the first eight components (8 highest captured variances) for model 4. It can be observed that temperature and conductivity are dominating variables in PC1. The same holds for ORP in PC2 and pH in PC3. PC4 is dominated by the weight and PC5 is influenced mostly by the conductivity values. In PC6, the DO is the dominating variable. Then, with the first 6 principal components, all the variables are represented. It can be seen that PC7 and PC8 are dominated by DO again. Based on these observations, six principal components are thus selected with 86.46% of variance captured total.

The same approach was taken to select the principal components for model 5 and model 6. It was observed (not shown) that the variables could all be found as dominating variable in the first 6 components, although the component in which they are respectively dominating was not necessary the same. It was possible to observe that the three types of scaling lead to similar captured variances until component number five. It is the sixth component that magnifies the difference between continuous scaling and other scalings. It should be kept in mind that autoscaling and group scaling cause a larger decrease in total variance when compared to continuous scaling.

To find the adequate number of principal components in variable direction, authors as Kyoo

Yoo, *et* al., (2003) uses the cross validation of the prediction residual sum of squares method. In this work has been applied the contribution per variable upon the principal components therefore. In this way, Six principal components are selected for models 1, 2 and 3. That decision is due to each PC represents variables different.

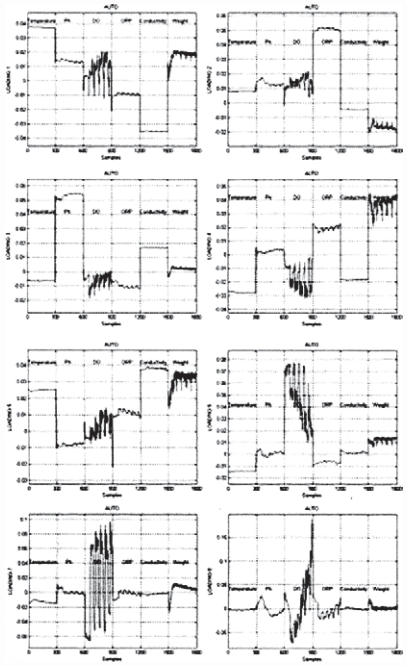

Fig. 5 Loads graphics from 1 to 8 components

3.3 Validation of the models

To evaluate the performance of the resulting MPCA models, the validation dataset was projected onto the models. Each batch in the validation set is projected on each of the 6 models, the corresponding statistics are calculated and checked against their in-control limits. That means that,, each batch was classified 6 times. First, all results are analyzed together. Then the analyses between the two unfolding types are made. Finally, each model is evaluated separately. The Hotelling T2 statistic and Q-statistic charts with 95%-confidence limit have been used for the batch-wise unfolded models. For variable-wise unfolded models, only the Hotelling T2-statistic was available since the number of PC's was equal to the number of variables, in this form the residual matrix is zero. First analysis: After projections onto the models, all observations (batches) in the validation data set were grouped into the following categories:

- Batches rejected by all models. These observations are not commun for comparison of the models and are therefore considered to be truly abnormal, i.e. detected faults.
- Batches accepted by none of the models. Also these observations are not commun for comparison of the models and are therefore considered truly normal, i.e. correct acceptance.
- Batches detected only by one up to five models are investigated in detail to understand what happened to the process.

The third category is useful common for model comparison consists of 52 batches. By detailed inspection it was found out that 47 of these batches were abnormal and 5 were normal batches. In table 4 the number of true alarms, false alarms, false acceptances, true acceptances and the total misclassifications (false alarms + false acceptances) are given. Given the results of the detailed inspection of the considered batches, the sum of the number of true alarms and the number of false acceptances is 47 for all models. Likewise, the sum of the number of false alarms and the number of true acceptances is always 5. The results are interpreted in terms of the effect of the scaling method by comparing the models within the group of batch-wise models and variable-wise models separately. Afterwards, the effect of the unfolding is interpreted by pair-wise comparison of the results for the corresponding batch-wise unfolded and variable-wise unfolded model.

Table 4 Comparative table of models

	1° Analysis						*2° Analysis*	
Models	1	2	3	4	5	6	1-3	4-6
True alarm	15	14	33	7	7	31	35	34
False acceptance	32	33	14	40	40	16	12	13
False alarm	0	0	4	0	0	1	4	1
True acceptance	5	5	1	5	5	4	1	4
Total error	32	33	18	40	40	17	17	18

As can be seen, models 1 and 2 (variable-wise unfolding with respect to AS and GS) deliver similar results. The number of true alarms (resp. 15 and 14) and false acceptances (resp. 32 and 33) differ only by one while the number of false alarms (both 0) and true acceptances (both 5) are exactly the same. When comparing the results for the variable-wise model with CS (model 3), with the other variable-wise models (model 1 and 2) considerable differences are seen. As can be seen, the false acceptance for model 3 (14) is less than half the false acceptance for models 1 and 2 (32 and 33). This lowered number of false acceptances is paid off by an increased number of false alarms (4), compared to model 1 and model 2 (both 0). The overall misclassification is however considerably lower (18) when they are compared to models 1 and 2 (resp. 32 and 33). Similar observations can be made with regard to the batch-wise unfolded models. All considered numbers are the same for models 4 and 5 and also in this case, the CS model (model 6) leads to a number of false acceptances (16) less than half the number obtained for models 5 and 6 (both 40). Again, this is paid off by a (small) increase in false alarms (1), compared to model 5 and 6 (both 0). Similarly to the variable-wise models, the overall number of misclassifications is lower for the CS model (17), when compared to models 4 and 5 (both 40). In order to investigate the effect of unfolding on the model performance, the variable-wise unfolding models and batch-wise unfolding models are compared pair-wise here, i.e. models with the same scaling are compared to each other. When comparing the results of the autoscaling (AS) models (model 1 and 4), it can be seen that variable-wise unfolding

leads to a lower number of false acceptances (32) when compared to the batch-wise unfolding (40), while the number of false alarms is the same (both 0). As a result, the total number of misclassifications is 32 and 40 in the case of variable-wise and batch-wise unfolding respectively. Very similar observations can be made when looking at the results in case of group scaling (models 2 and 5). The only difference lies in an increased (+1) number of false acceptances and the corresponding increase of the total number of misclassifications.

Second analysis: When looking at the results in the case for continuous-scaling (models 3 and 6), observations are different compared to the previous case. However the difference is rather small, variable-wise unfolding leads to a lower number of false acceptances (12) compared to batch-wise unfolding (13). In this case, this is paid off by a higher number of false alarms (4) in the case of variable-wise unfolding compared to the case of batch-wise unfolding (1). In the overall number of misclassifications, the difference is again small as the variable-wise unfolding leads to 17 misclassifications while the batch-wise unfolding leads to 18 misclassifications. In general, it is observed that effect of unfolding are smaller than the effect of scaling.

CONCLUSIONS

In this paper several approaches to PCA-based biological process monitoring that have been discussed in literature were compared to each other. The constructed models exhibited different two common types of unfolding and three conventional ways of scaling. In order to compare the performance of the models the assumption on linearity of the scores was checked for each of the models. The number of components needed for modelling was checked as well and was found out to be the same for all models. With model 4, all the periods during the trajectory are treated equally. When the data are scaled using group scaling the variability in each trajectory is loaded more than in the rest of cases. In the event of continuous scaling the variability in one trajectory for variable is calculated. In the same way, normalizing in variable direction with auto, group and continuous scaling are implemented. Making a comparison between validations of the models with equal normalizing, it is possible to observe that, similar results have been obtained for model where the auto and group scaling were applied. In variable direction more batches with AOC have been detected with a bit error in the classification. Nevertheless as much as model 3 and 6 detected more faults than the others but the false alarms are incremented. Based on the results, the scaling decision should be related to the objectives of whether the operator wants. Normalizing the data, the dynamic of the process is removed getting that the models describe close to the reality however it is necessary think about the possible false alarms or omission at the moment to select the approach.

ACKNOWLEDGEMENT

This work is part of the project "Development of a intelligent control system apply to a Sequencing Batch Reactor by loads (SBR) for the elimination of organic matter, nitrogen and phosphorus" DPI2005-08922-C02-02 supported by the Spanish Government, the FEDER Founds and the Institute for Encouragement of Innovation by means of Science and Technology in Flanders (IWT).

REFERENCES

Al-Kandari N. and I. Jolliffe (2005) "Variable selectrion and interpretation in correlation principal components" Envirometrics; 16:659-672

P. Giudici (2003) "*Applied data mining – Statistical methods for business and industry*". England: John Wiley & Sons Ltd

Kyoo Yoo, C., P. Vanrolleghem, (2003) I.B. Lee "Nonlinear modelling and adaptive monitoring with fuzzy and multivariate statistical methods in biological wastewater treatment plants" Journal of Biotechnology 105 135-163

J.M. Lee, C.K. Yoo, S.W. Choi, P. Vanrrolleghem, I.B. Lee (2004) "Nonlinear process monitoring using kernel principal component analysis" Chemical Engineering Science 59: 223-234

Nomikos, P. and J.F. MacGregor (1994) "Monitoring batch process using multiway principal component analysis" AIChE Journal Vol40 No 8

Nomikos P. and J.F. MacGregor (1995) "Multivaraite SPC Charts for monitoring batch processes" Technometrics Vol 37 No 1

Qin, S. and R. Dunia (2000) "Determining the number of principal components for best reconstruction" Journal of Process Control 10 254-250

Sin. G. (2004) "Systematic calibration of activated sludge models" Ph.D. thesis Ghent University Promotor Peter Vanrolleghem

Sung Lee, D., J. Moon Park and P. Vanrolleghem (2005) "Adaptive multiscale principal component analysis for on-line monitoring of a sequencing batch reactor" Journal of Biotechnology 116 195-210

Sung Lee, D. and P. Vanrolleghem (2003) "Monitoring of a sequencing batch reactor using adaptive multiblock principal component analysis" Biotechnol Bioeng 82: 489-497

Undey, C. and A. Cinar (2002) "Statistical monitoring of multistage, multiphase batch processes" IEEE Control Systems Magazine 22 (5) 40-52

Westerhuis, J., T. Kourti and J.F. MacGregor (1999) "Comparing alternative approaches for multivariate statistical analysis of batch process data" Journal of Chemometrics 13, 397-413

Wold, S., P. Galdi, K. Esbensen and J. Öhman (1987) "Multi-way principal components and PLS analysis" Journal of Chemometrics, Vol 1, 41-56

Wold, S., N. Kettaneh, H. Friden and A. Holmberg (1998) "Modelling and diagnostics of batch processes and analogous kinetic experiments" Chemometics and Intelligent Laboratory Systems 44 331-340