Combining multiway principal component analysis (MPCA) and clustering for efficient data mining of historical data sets of SBR processes

Kris Villez, Magda Ruiz, Gürkan Sin, Joan Colomer, Christian Rosén and Peter A. Vanrolleghem

ABSTRACT

A methodology based on Principal Component Analysis (PCA) and clustering is evaluated for process monitoring and process analysis of a pilot-scale SBR removing nitrogen and phosphorus. The first step of this method is to build a multi-way PCA (MPCA) model using the historical process data. In the second step, the principal scores and the Q-statistics resulting from the MPCA model are fed to the LAMDA clustering algorithm. This procedure is iterated twice. The first iteration provides an efficient and effective discrimination between normal and abnormal operational conditions. The second iteration of the procedure allowed a clear-cut discrimination of applied operational changes in the SBR history. Important to add is that this procedure helped identifying some changes in the process behaviour, which would not have been possible, had we only relied on visually inspecting this online data set of the SBR (which is traditionally the case in practice). Hence the PCA based clustering methodology is a promising tool to efficiently interpret and analyse the SBR process behaviour using large historical online data sets.

Key words | LAMDA clustering, MPCA, nutrient removal, on-line monitoring, SBR

Kris Villez
Gürkan Sin
Peter A. Vanrolleghem
BIOMATH, Ghent University,
Coupure Links 653, B-9000 Ghent,
Belgium
E-mail: *Kris.Villez@biomath.ugent.be*;
Peter.Vanrolleghem@gci.ulaval.ca

Magda Ruiz
Joan Colomer
eXiT, Department of Electronics, Computer Science and Automatic Control,
University of Girona,
Campus Montilivi CP 17071 Building PIV,
Girona,
Spain
E-mail: *mlruizo@silver.udg.es*

Christian Rosén
IEA, Lund University,
Box 118, SE-221 00 Lund,
Sweden
E-mail: *Christian.Rosen@iea.lth.se*

Peter A. Vanrolleghem
model*EAU*, Dépt. génie civil, Pavillon Pouliot,
Université Laval, Québec, QC, G1K 7P4,
Canada
E-mail: *Peter.Vanrolleghem@gci.ulaval.ca*

INTRODUCTION

In the past decades, the search for advanced control strategies has gained attention in wastewater treatment engineering (Olsson & Newell 1999). Despite promising results in this area, new and advanced control strategies are generally not applied in full-scale wastewater treatment plants as the required reliability of sensors and actuators is not met in many cases. In addition, changes of the microbial population, as reported in Yuan & Blackall (2002) and Sin *et al.* (2006), which might lead to lower performance of the system, limit the application of common control strategies. A systematic approach to process monitoring and diagnosis that aims to address these issues is expected to improve the control of wastewater treatment plants.

The application of process monitoring techniques for wastewater treatment processes has recently gained momentum. One of the first applications of principal component analysis (PCA) to monitor a full-scale WWTP is given by Rosén & Olsson (1998). In many other cases, the basic technique, often PCA, is extended to address typical features of biological processes. Among the most important are the *multiscale* extensions to tackle the monitoring problem at different time-scales (Rosén & Lennox 2001), *adaptive* modelling to account for changing system properties (Rosén & Lennox 2001), *multiblock* modelling to facilitate fault isolation in the context of processes exhibiting several structurally different phases (Lee & Vanrolleghem 2003) and the *Kernel* extension to tackle

doi: 10.2166/wst.2008.143

severe non-linearities (Lee *et al.* 2004). Classical PCA-based approaches to the diagnosis of faulty situations involve the visual inspection of contribution plots, in which the contribution of all or certain variables to the scores and statistics (Hotelling's T^2, Q-statistic) is evaluated. This becomes a cumbersome and time-consuming task when confronted with large data sets. Dunia & Qin (1998) discuss a method for automated diagnosis based on PCA modelling. A set of a priori identified faults needs to be identified however, which is unrealistic in wastewater treatment practice. Singhal & Seborg (2002) provide a pattern-matching tool to retrieve similar behaviour in a historical database. However, operators are supposed to select the final matching case which impedes automatic diagnosis. Case-Based Reasoning (CBR) provides a framework for fault diagnosis as well (Martinez *et al.* 2006).

Despite the available methods, a lack of automation in wastewater treatment plants (WWTPs) is still present. At the same time, the increased use of on-line sensors for monitoring of WWTPs results in large amounts of data, often not analyzed, managed or used in an efficient manner. A systematic approach for data screening and interpretation is however a crucial step for modelling and for the design of control strategies. Hence, it is our aim to make a further step towards facilitated screening and interpretation of large historical data sets from wastewater treatment facilities. To this end, a combined methodology of PCA-modelling and LAMDA (Learning Algorithm for Multivariable Data Analysis) clustering is proposed where observations are clustered using principal scores and Q-statistics as described in detail below. The methodology is evaluated at a pilot-scale SBR for nitrogen and phosphorus removal.

The paper is organised as follows. After materials, the combined PCA and LAMDA clustering are explained. Then, the results of the methodology are presented and evaluated to assess the efficiency of the proposed approach for data mining of historical data sets of SBR processes. Finally, conclusions regarding the devised method are drawn.

MATERIALS AND METHODS

Data

The data set used in this paper consists of 1959 complete batches collected from a pilot-scale SBR setup between December 16th, 2003 and July 18th, 2005. The SBR under study has a working volume of 64L and is fed with synthetic sewage resembling domestic wastewater characteristics (Insel *et al.* 2006). Detailed information on the setup can be found in Lee *et al.* (2005). It is noted here that three contiguous operational periods are discerned in the studied period. They last from December 16th, 2003 to March 3rd, 2004 (OP1), from March 3rd, 2004 to December 16th, 2004 (OP2) and from December 16th, 2004 to July 18th, 2005 (OP3). The first change in operation concerns a higher oxygen setpoint, while the second change is due to a reconsidered feeding pattern, oxygen setpoint and SRT (Sludge Retention Time).

The length of one cycle, i.e. one batch run, is 6 hours. The cycle consists of a fill/anaerobic phase (60'), 4 sequences of an aerobic (32.5′) and an anoxic phase (30′), a final aerobic phase (30′) and a settling/draw phase (60′). During the fill/anaerobic phase of the cycle, 24 L of the influent is supplied, while 10 L is equally step-fed to the anoxic phases, i.e. 2.5 L per each anoxic phase.

The on-line data of each batch consists of 6 trajectories corresponding to pH, Oxidation-Reduction Potential, Dissolved Oxygen, temperature, weight and conductivity sensors. Data collected during the settling and draw phases are excluded from these trajectories since (1) these data do not provide much information about the processes (Lee & Vanrolleghem 2003), (2) measurements are not representative since the medium is not mixed and (3) sludge settling exhibits dynamics that might lead to batch-to-batch data variation which is not straightforward to explain.

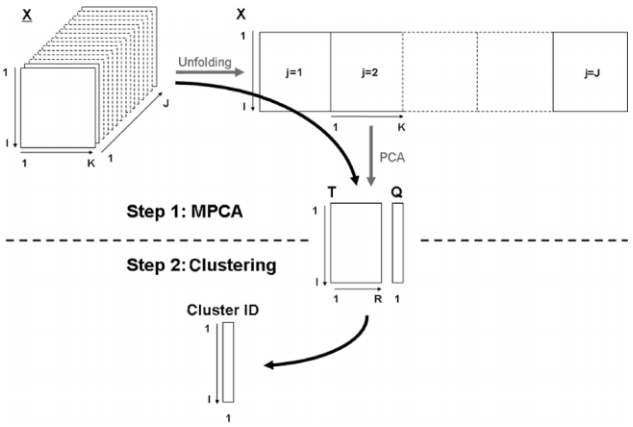

Figure 1 | MPCA-based clustering procedure.

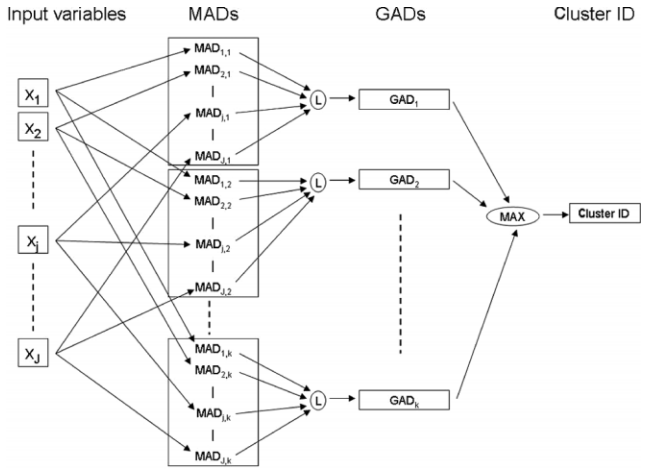

Figure 2 | LAMDA model structure: MAD = Marginal Adequacy Degree, GAD = Global Adequacy Degree, L = hybrid connective operator.

Method: combining multiway principal component analysis and clustering

The applied method consists of two steps, being (1) a data dimension reduction by means of MPCA and (2) clustering of the resulting data set with reduced dimensions, as shown in Figure 1. In this work, MPCA was performed as proposed by Nomikos & MacGregor (1994). The modelling procedure thus includes batch-wise unfolding, autoscaling and linear PCA modelling.

The resulting principal component scores and the Q-statistic are then used as input variables for clustering. The Learning Algorithm for Multivariate Data Analysis (LAMDA) is applied to do so. The structure of any resulting model is similar to that of a single neuron in a neural network with as many nodes as classes (see Figure 2). In this structure, the Marginal Adequacy Degree (MAD) is a

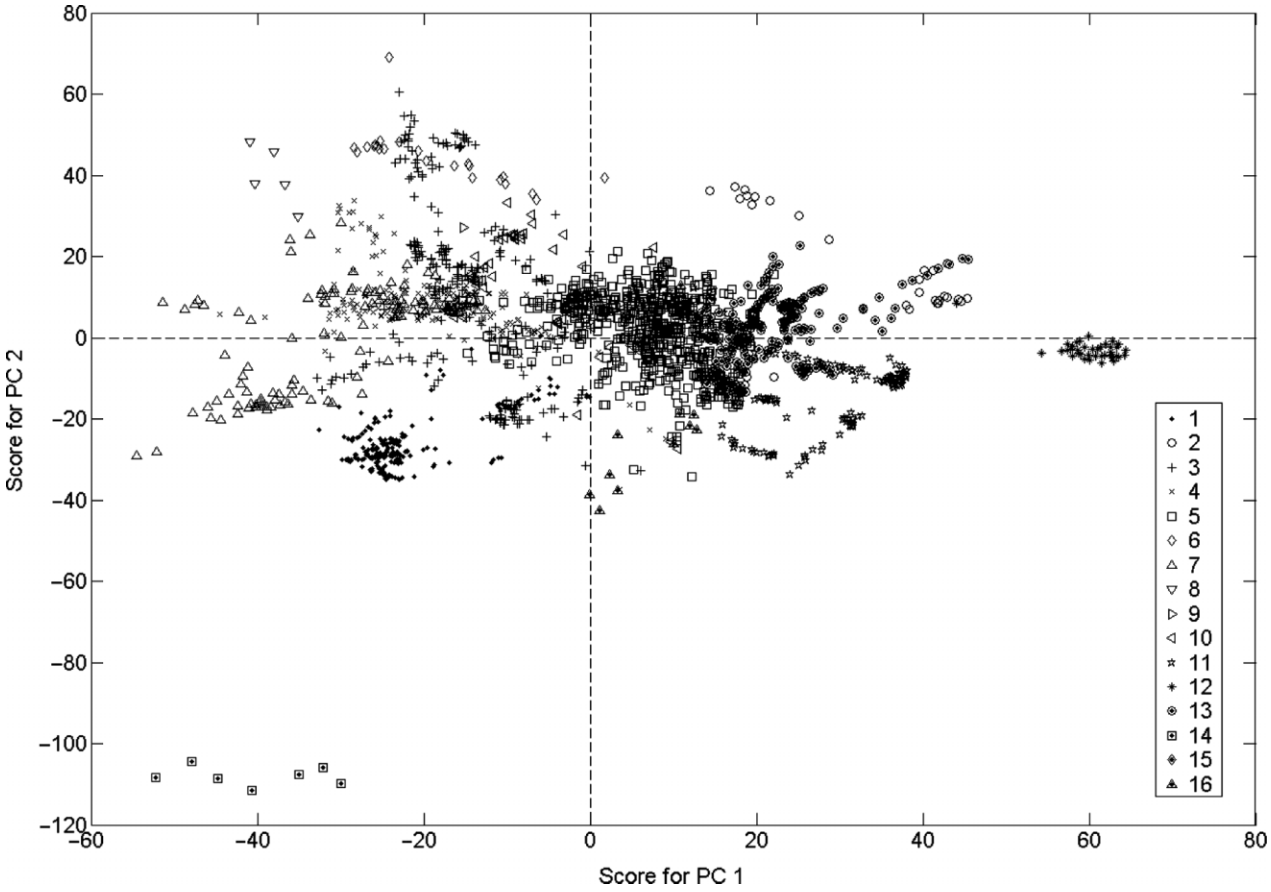

Figure 3 | Biplot of the first two scores of the first PCA model. Markers indicate the cluster ID.

measure for the possibility that an observation belongs to a class given one of the input variables. In this work, the fuzzy extension of the binomial probability function is used:

$$\mathrm{MAD}_{j,k} = \rho_{j,k}^{x_j} \cdot (1 - \rho_{j,k})^{x_j}$$

where x_j is the observed value for the input variable j and $\rho_{j,k}$ the classifier model parameter for input variable j and class k. The values for $\rho_{j,k}$ are optimized during the training process. The Global Adequacy Degree combines the information of all MAD's for a given observation and a given class by means of hybrid connectives. In this case, these connectives are established as follows (for J input variables):

$$\begin{aligned}\mathrm{GAD}_{\beta,k} &= L(\mathrm{MAD}_{1,k},\ldots,\mathrm{MAD}_{j,k},\ldots,\mathrm{MAD}_{J,k}) \\ &= \beta \cdot T(\mathrm{MAD}_{1,k},\ldots,\mathrm{MAD}_{j,k},\ldots,\mathrm{MAD}_{J,k}) + (1-\beta)\cdot \\ &\quad S(\mathrm{MAD}_{1,k},\ldots,\mathrm{MAD}_{j,k},\ldots,\mathrm{MAD}_{J,k})\end{aligned}$$

where $L(\ldots)$ denotes the connective operator, $T(\ldots)$ is a t-norm and $S(\ldots)$ is its dual t-conorm as commonly denoted in the context of fuzzy logic and β a parameter to be set by the user. In this case β was set to 1 so as to define an intersection in the fuzzy logic analogy (β set to 0 defines the union). An observation is assigned to the class k with maximal $\mathrm{GAD}_{\beta,k}$.

Important features of the LAMDA modelling algorithm are that (1) (supervised) classification, (unsupervised) clustering or mixed forms are possible, (2) input variables can be of quantitative and/or qualitative nature and (3) the sequential treatment of observations leads to fast training. For more details we refer to the work of Aguilar & Lopez de Mantáras (1982) and Moore (1995).

RESULTS

In this section, first the PCA-based clustering on the whole data set (all three operational periods) is shown and secondly the PCA-based clustering of the normal operational condition (NOC) data, identified in the first step, is presented.

PCA-based clustering of the whole on-line data set

Following the PCA modelling step, the matrix consisting of 6 retained principal scores and the Q-statistic was fed to the LAMDA clustering algorithm. The LAMDA algorithm clustered the observations (batches) into 16 clusters. A biplot of the first two scores for all data is shown in Figure 3. It can be seen that the clustering algorithm separates small groups of outliers from other clusters that are larger in number of members.

Each of the clusters was subjected to diagnosis by close investigation of the on-line data. The labels that were obtained by doing so are given in Table 1 together with the number of batches. Only one cluster (cluster 16) could not be tagged uniquely, though this cluster exhibited only abnormal batches. As can be observed, only 5 clusters (1, 3, 5, 7 and 13) were identified as normal, corresponding to 73% of the whole data set. By normal operation, it is meant that the operation of sensors and actuators is technically correct. The 10 remaining clusters could be linked uniquely to a specific problem or set of problems.

Table 1 | Interpretation of the first clustering results. "Normal" defines correct operation of hardware and software

No.	Batches	Label
1	219	Normal 1 (low DO operation)
2	31	Communication problem with balance
3	241	Normal 2
4	187	Cooler failure
5	607	Normal 3
6	23	High pH
7	126	Normal 4, recovery from cooler failure (cluster 4)
8	5	Extreme DO
9	1	Low ORP and extreme DO
10	72	Conductivity probe failure
11	144	Conductivity probe in repair
12	39	Conductivity probe in repair and communication problem with balance
13	245	Normal 5 (optimised operation)
14	7	Low ORP measurement
15	1	Multiple sensor failure (ORP, temperature and weight)
16	11	Abnormal (no unique fault)

PCA-based clustering of the NOC on-line data set

As outliers can have a large influence on PCA models, the PCA-based clustering of the whole data set might have impaired the discrimination between different types of normal behaviour. Therefore, the PCA-based clustering was repeated on the data of the batches assessed to be normal only. This means that the data corresponding to clusters 1, 3, 5, 7 and 13 in the former clustering procedure were used for PCA model training and consequent clustering. All following graphs and results correspond to this "NOC" data set only.

The 7 retained principal scores and the Q-statistic, obtained by PCA modelling, were fed to the LAMDA-algorithm as in the previous section. An unexpected large number (17) of clusters were hereby obtained. Figure 4 shows the biplot of the first two scores for all batches under study. These clusters were investigated again in detail in order to label them. In Table 2, the obtained clusters are given with their number of batches and label, ordered in order of appearance. Now, 10 out of 17 clusters are identified as normal and represent 97% of the NOC data set (1399 batches out of 1438). As such, the newly found abnormal batches, representing 3% of the NOC data set and 2% of the whole data set. This indicates that a large part of the identified abnormal batches can be identified by a single application of the combined PCA and clustering procedure. In a classic PCA application, 362 faulty batches (65% of all identified faulty batches) would not be selected for investigation on the basis of 95%-levels for either Hotelling's T^2 or the Q-statistic. As such, the MPCA-based clustering is shown to be beneficial for data screening.

In addition to the lower coverage of abnormal data in the cleaned up data set, the "normal" data is now split up into 10 clusters instead of 5, as in the first iteration of the PCA-based clustering. In Figure 5 the class numbers for all

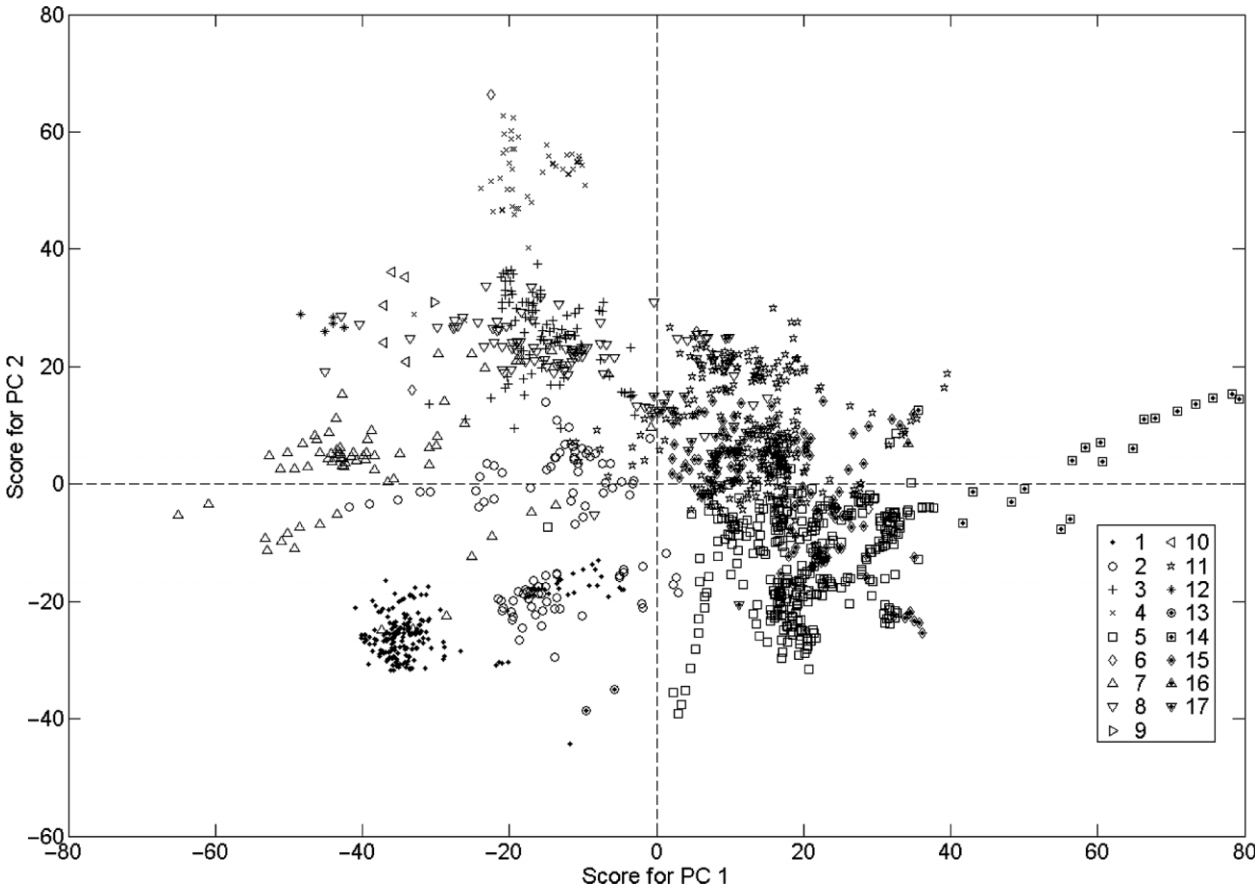

Figure 4 | Biplot of the first two scores of the second PCA model. Markers indicate the cluster ID.

Table 2 | Interpretation of the second clustering results

No.	Batches	Tag	Label
1	210	Normal 1	Low DO set point, transient operation
2	95	Normal 2	Low DO set point, steady operation
3	98	Normal 3	High SV_{30} (bad settling), low NO_3—N
4	49	Normal 4	Decreasing SV_{30} (improving settling), increasing NO_3—N
5	334	Normal 5	Filamentous bulking, decreasing/low NH_4—N, high NO_3—N
6	4	Abnormal 1	High DO in anoxic phases
7	71	Normal 6	Increasing SV_{30} (worsening settling), high NO_3—N (10–20), increasing COD
8	94	Normal 7	High SV_{30} (bad settling), high NO_3—N (>20), increasing COD
9	3	Abnormal 2	High DO in aerobic phases
10	5	Abnormal 3	High DO in anaerobic phase (mixing too intense)
11	289	Normal 8	Filamentous bulking, high NH_4—N, decreasing/low NO_3—N
12	5	Abnormal 4	Cooling system failure
13	2	Abnormal 5	Pump control error: feeding too high in anaerobic phase
14	19	Abnormal 6	Pumping failure
15	116	Normal 9	Filamentous bulking, decreasing/low NH_4—N, high NO_3—N
16	1	Abnormal 7	High DO in aerobic phases
17	43	Normal 10	Filamentous bulking, high NH_4—N, low NO_3—N

NOC batches are shown as a function of time. It can be observed that the normal clusters (1, 2, 3, 4, 5, 7, 8, 11, 15 and 17) appear in different frames of the studied time window. Some of the differentiation by clustering is due to enforced operational changes according to the SBR research agenda at that time. For instance, clusters 1 and 2 represent batches of the first operational period (OP1), whereas all other clusters represent bathes in OP2 and OP3. Next to this, clusters 1, 2, 3, 4, 7 and 8 represent data logged before December 16th, 2004 (OP1 and OP2), when a major change in the operation was implemented, after the model-based optimisation results of Sin *et al.* (2004). The clusters 5, 15 and 17 appear almost exclusively in OP3. By exception, cluster 11 shows exceptionally large numbers of batches in both OP2 and OP3. The labels provided in Table 2 for the normal clusters concern a qualitative assessment of the performance of the system in terms of settling properties and nutrient removal performance. While their assessment was of a subjective nature, it can be seen that the clusters differentiate between relevant process characteristics.

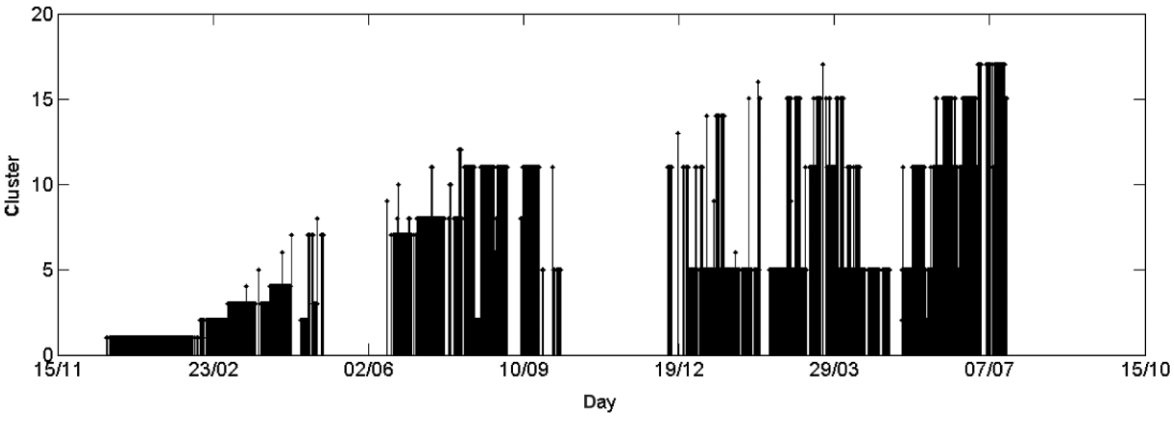

Figure 5 | Cluster number as function of time indicating how clustering memberships exhibit a pattern in time.

A more detailed interpretation of the relations between the clustering results and off-line measurements is however not the subject of this paper.

DISCUSSION

It was shown that, by means of clustering, batches with the same or similar behaviour are grouped together. Labelling is then done by investigation of a limited number of batches of those clusters. Combining MPCA and clustering therefore provides an efficient and effective tool for data screening and interpretation of historical data of batch processes.

The first application of the combined MPCA and clustering approach led to the discrimination of several clusters that uniquely correspond to a certain fault or set of faults. These clusters represented 93% of all abnormal batches found. On the contrary, 35% of all abnormal batches could be identified by means of the classic approach. As such, PCA-based clustering was shown to be a far more effective tool for data screening of large historical data sets when compared to the classic approach based on inference statistics.

After the selection of the normal data, the PCA-based clustering method was repeated on the cleaned-up data set. By doing so, an increased level of differentiation within this dataset was observed (10 normal clusters were found instead of only 5 in the first analysis). Also, it was possible to link the clusters to certain temporal process behaviour. Detailed investigation showed that by means of the combined PCA and clustering methodology both intended and unintended changes in process behaviour of the biological system could be discriminated. Importantly, this implies that the resulting clusters reflect meaningful changes in the process behaviour.

CONCLUSIONS

It is shown that data mining of historical data sets on the basis of PCA modelling can be significantly improved by the use of clustering techniques, such as the LAMDA clustering algorithm. Firstly, the PCA-based clustering is shown to be a fast and robust tool for data screening. Not only does it allow removing the larger part of abnormal batches in a single iteration, it also leads to a robust discrimination between different anomalies. The latter may help process operators to understand and interpret the corresponding failures in a fast way.

Secondly, the PCA-based clustering was repeated on the data assigned to be normal in the first LAMDA clustering application. Results showed that the combination of PCA modelling and the LAMDA algorithm allowed a clear-cut discrimination of applied operational changes to the SBR system. More important, a priori unknown but meaningful variation in the data set was revealed by means of the method presented.

ACKNOWLEDGEMENTS

This work was supported by the Institute for Encouragement of Innovation by means of Science and Technology in Flanders (IWT) and the research project Development of an 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. Peter Vanrolleghem holds the Canada Research Chair in Water Quality Modelling.

REFERENCES

Aguilar-Martin, J. & López de Mántaras, R. 1982 The process of classification and learning the meaning of linguistic descriptors of concepts. In: *Approximate Reasoning in Decision Analysis*, Gupta, M. M. & Sanchez, E. (eds). North-Holland Publishing Company, New York, pp. 165–175.

Dunia, R. & Qin, S. J. 1998 Joint diagnosis of process and sensor faults using principal component analysis. *Control Eng. Pract.* **6**, 457–469.

Insel, G., Sin, G., Lee, D. S. & Vanrolleghem, P. A. 2006 A calibration methodology and model-based systems analysis for SBR's removing nutrients under limited aeration conditions. *J. Chem. Technol. Biotechnol.* **81**(4), 679–687.

Lee, D. S. & Vanrolleghem, P. A. 2003 Monitoring of a sequencing batch reactor using adaptive multiblock principal component analysis. *Biotechnol. Bioeng.* **82**(4), 489–497.

Lee, J.-M., Yoo, C. K., Choi, S. W., Vanrolleghem, P. A. & Lee, I. B. 2004 Nonlinear process monitoring using kernel principal component analysis. *Chem. Eng. Sci.* **59**, 223–234.

Lee, D. S., Park, J. M. & Vanrolleghem, P. A. 2005 Adaptive multiscale principal component analysis for on-line

monitoring of a sequencing batch reactor. *J. Biotechnol.* **116**, 195–210.

Martinez, M., Sànchez-Marrè, M., Comas, J. & Rodríguez-Roda, I. 2006 Case-based reasoning, a promising tool to face solids separation problems in the activated sludge process. *Water Sci Technol.* **53**(1), 209–216.

Moore, K., Burback, R. & Heeler, R. 1995 Using neural nets to analyse qualitative data. *Mark. Res. Mag. Manage. Appl.* **7**(1), 35–39.

Nomikos, P. & MacGregor, J. F. 1994 Monitoring batch process using multiway principal component analysis. *AIChE J.* **40**(8), 1361–1374.

Olsson, B. & Newell, B. 1999 *Wastewater Treatment Systems – Modelling, Diagnosis and Control*. IWA Publishing, London, UK, pp. 750.

Rosén, C. & Olsson, G. 1998 Disturbance Detection in Wastewater Treatment Plants. *Water Sci. Technol.* **37**(12), 197–205.

Rosén, C. & Lennox, J. A. 2001 Multivariate and multiscale monitoring of wastewater treatment operation. *Water Res.* **35**(14), 3402–3410.

Sin, G., Insel, G., Lee, D. S. & Vanrolleghem, P. A. 2004 Optimal but robust N and P removal in SBRs: a systematic study of operating scenarios. *Water Sci. Technol.* **50**(10), 97–105.

Sin, G., Govoreanu, R., Boon, N., Schelstraete, G. & Vanrolleghem, P. A. 2006 Evaluation of the impacts of model-based operation of SBRs on activated sludge microbial community. *Water Sci. Technol.* (in Press).

Singhal, A. & Seborg, D. E. 2002 Pattern matching in multivariate time series databases using a moving-window approach. *Ind. Eng. Chem. Res.* **41**, 3822–3838.

Yuan, Z. & Blackall, L. 2002 Sludge population optimisation, a new dimension for the control of biological wastewater treatment systems. *Water Res.* **36**, 482–490.