Multivariate on-line monitoring: challenges and solutions for modern wastewater treatment operation

C. Rosen*, J. Röttorp and U. Jeppsson***

* Industrial Electrical Engineering and Automation, Lund University, Box 118, SE-221 00 Lund, Sweden (E-mail: *christian.rosen@iea.lth.se*; *ulf.jeppsson@iea.lth.se*)

** IVL, Swedish Environmental Research Institute, SE-100 31 Stockholm, Sweden (E-mail: *jonas.rottorp@ivl.se*)

Abstract In this paper, a number of challenges, which need to be overcome if multivariate monitoring of wastewater treatment operation is to be successful, are presented. For each challenge, one or several solutions are discussed. The methodologies are illustrated using an example from full-scale wastewater treatment operation. Some guidelines regarding choices of methods and implementation aspects are given.
Keywords Adaptive monitoring; detection; principal component analysis (PCA); wastewater treatment operation

Water Science and Technology **Vol 47 No 2 pp 171–179** © IWA Publishing 2003

Introduction

On-line monitoring of industrial processes is carried out to ensure that process outputs meet requirements on product quality, process safety and efficient use of resources. Modern wastewater treatment (WWT) plants collect large numbers of on-line measurements as more and more process variables can be measured. The on-line measurements have become an important source of information in the effort to achieve efficient operation and management of WWT plants. The large number of measured variables together with the nature of WWT processes put high demands on the techniques to extract on-line information from the data.

In most process industries, monitoring of the process and its outputs is an important part of the operation. Monitoring can be said to consist of three phases:

1. Detection – recognising that there is a deviating event or that the process is not operating at its normal operational point;
2. Isolation – finding the deviating measurement variables that have triggered the detection;
3. Interpretation – finding the physical causes of the deviation and assessing its impact on the process.

The first phase is a task well suited for computers, as it is monotonous and quantitative. The second phase is normally integrated with the first task and is, thus, also suitable for computers. However, the third phase requires process knowledge and is mainly a task for the operator relying on his experience (and possibly intuition), although attempts have been made using knowledge based systems, such as expert systems, for the interpretation. In order to facilitate the third phase, the methods used for the first two tasks must extract and organise the information appropriately and present the information in an easily interpretable way.

During the last two decades, techniques based on multivariate statistics have become increasingly popular within many industrial fields. Principal component analysis (PCA) and developments based on PCA, such as principal component regression (PCR) and projection to latent structures (PLS), have been applied successfully to various industrial

processes (e.g. Kourti *et al*., 1996), including wastewater treatment (Rosen and Olsson, 1998; Teppola *et al*., 1998; Rosen and Lennox, 2001). However, while these methods have worked well in some applications, they have worked poorly in others. There are many reasons for this, of which lack of understanding for the techniques and poor choice of methods are probably the most important. Unsatisfactory experiences have in some cases given multivariate monitoring an undeservedly bad reputation. It is the authors' intention to provide some insights as to why these methods sometimes fail and how some common problems can be overcome to obtain a powerful monitoring method, adapted to the needs of WWT operation. General guidelines for the choice of methods and data pre-processing to obtain a maximum of information will be discussed. Often-encountered problems will be exemplified and discussed using real WWT operation data.

Challenges

On-line monitoring of industrial processes implies a number of challenges. Kresta *et al*. (1991) list some general ones for a monitoring method:

1. The method must be able to deal with collinear data of high dimension, in both independent and dependent variables;
2. The method must reduce the dimension of the problem substantially and allow for simple graphical interpretations of the results;
3. If both process and quality variables are present, the method must provide good predictions of the dependent variables.

These challenges are normally used to justify the use of methods based on multivariate statistics such as PCA or PLS. In multivariate statistics, the collinear nature of data is utilised to reduce the dimensionality of the monitoring problem. Moreover, PLS and PCR, being regression techniques, can also be used for predictive purposes. Thus, the multivariate statistics framework provides a basis to meet the above challenges. However, to fulfil the more specific challenges in WWT monitoring, a monitoring method must also meet some additional requirements to be successful:

4. Poor data quality and reliability due to the hostile environment in which the measurement equipment has to function;
5. Non-linear relationships between variables;
6. Changing conditions due to, for instance, diurnal and seasonal changes, which makes the process far from stationary;
7. Dynamic relationships between variables with a wide range of time constants.

All these challenges are not necessarily encountered in the same process or at the same time, since the aim of the monitoring task may differ depending on the application. However, for an all-embracing monitoring system at a WWT plant most of the listed challenges must be met and overcome. If this is not achieved, it will be difficult to obtain end-users' or operators' acceptance, which is a necessity for success.

Basic multivariate monitoring

Multivariate statistics can be used to meet the general challenges (1–3) listed above. In this paper, we will only discuss PCA and its extensions, as it, through its simplicity, serves as a good introduction to multivariate statistics (the same methodology is extendable to PCR and PLS).

Principal component analysis

The basic idea behind PCA is that the collinear nature of data is utilised to reduce the dimensionality of the measurement space by introducing a number of pseudo-variables (principal components). These pseudo-variables describe the main mechanisms that drive

the process and are normally significantly fewer than the number of measured variables. PCA can be described as a method to fit a line (or component) in the direction of greatest variability of the measured variable space. Next, a line is fitted in the second greatest direction of variability orthogonal to the first line and, thus, a plane is obtained. The following line is fitted in the third greatest direction of variability, orthogonal to the plane. This is continued until it is established that no systematic variability remains (Figure 1).

In mathematical terms, PCA is obtained by singular value decomposition of the covariance or correlation matrix of the process data. By doing so, a subspace (process subspace) containing the true (non-random) variation is identified. Complementary to this subspace is the noise subspace, which ideally contains only noise. In matrix form, PCA is written as $\mathbf{X} = \mathbf{T}\mathbf{P}_a^T + \mathbf{E}$, where $\mathbf{X}$ is the original data set of size $[m \times n]$, $\mathbf{T}$ is called score $[m \times a]$, $\mathbf{P}$ is called loading $[n \times a]$ and $\mathbf{E}$ is the model residual (or noise subspace). If $a = n$ then $\mathbf{E} = 0$, as all the variability directions are described. However, if $a < n$, i.e. less principal components than original variables are retained, then $\mathbf{E}$ describes the variability not described by the sum of the matrices $\mathbf{T}\mathbf{P}_a^T$. In general, $a << n$ is true for industrial applications. Note that variables generally must be scaled so that they will have similar influence on the model regardless of their original numerical values and engineering units. A common solution is to scale every variable to unit variance after mean centring (autoscaling). More elaborate discussions on PCA and other multivariate statistics based methods are found in most of the references listed in this paper.

On-line monitoring

The basic principle for process monitoring using PCA is that a training set of data representing normal operational conditions is decomposed and a process subspace is identified. When new process data are obtained, they are projected onto the process subspace and noise space, respectively, according to $\mathbf{T}_{new} = \mathbf{X}_{new}\mathbf{P}_a + \mathbf{E}$ or $\mathbf{t}_{new} = \mathbf{x}_{new}\mathbf{P}_a + \mathbf{e}$. Through investigation of the projected data $\mathbf{T}_{new}$ and the residual $\mathbf{E}$, process deviations and disturbances can be detected utilising various techniques (Jackson and Mudholkar, 1979; Kresta *et al.*, 1991; Yoon and MacGregor, 2001). By analysing the sum of the squared prediction error (*SPE*), the current model fit is determined. If the current operation display poor fit, the current operational state is obviously different from that of the training data set. Hotelling's T^2 is a summarised way of surveying the scores and is used to assess the variations within the model. When a significant deviation is established, backtracking through the model (contribution analysis) is carried out to isolate what variables contribute to the deviation (MacGregor *et al.*, 1994; Teppola *et al.*, 1998).

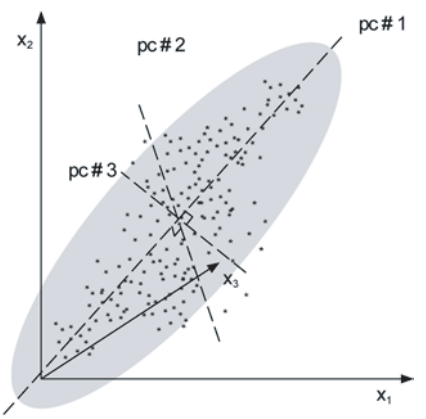

Figure 1 PCA as a successive fitting of lines (or components) in the directions of the largest variability

Improvements

To overcome challenges 4–7, some modifications to the basic multivariate monitoring approach must be done. In this section, some extensions to basic PCA are discussed that makes the approach applicable to WWT process monitoring.

Poor data quality

The quality of data is crucial to the outcome of the monitoring task. All measurements are affected by noise. This may significantly distort the information carried by a signal. However, in WWT monitoring, the time limitation is generally not a crucial factor. Thus, there is ample time to scrutinise data before they are used and this is to an increasing extent already done in state-of-the-art sensor equipment. Noise will still be present in data and here an appealing feature of multivariate statistics comes into play. The decomposition of data is carried out so that two subspaces are obtained: the process and the noise subspaces. If this partition is done correctly (i.e. a suitable choice of a), the process subspace will contain the actual changes in the process, whereas the noise subspace will contain mainly noise.

An often encountered problem is that of missing data. A missing data point is normally a symptom of problems in the sensor or the measurement system. However, if the problem is known (and thus not an issue for detection), missing values will severely distort the monitoring results unless the value is replaced by estimates. Replacements of only a few missing values can be achieved by using the last known value or prediction algorithms, but a prolonged time period without values, leads to estimates that become less and less accurate. To minimise the damage to the overall monitoring, the value can be replaced so that it fits the monitoring model. There are many ways to deal with missing data in PCA (Nelson *et al.*, 1996; Grung and Manne, 1998). Perhaps the most basic way is to replace the value with its expected mean. This will not take the internal relations within the data into account, since the variable will not covariate with the other variables. A more accurate estimate can be obtained if the value is replaced using the internal relationships captured in the model. The estimate is calculated using all the available data to produce an estimate that fits the model. In this case, the estimated variable covariates with the other variables. Let $\chi^T = \mathbf{x}$, $\tau^T = \mathbf{t}$ and χ^{*T} be a row vector of the available values. Then, the score vector is estimated as $\tau = (\mathbf{P}_a{}^{*T}\mathbf{P}_a{}^{*})^{-1}\mathbf{P}_a{}^{*T}\chi^{*}$, where $\mathbf{P}_a{}^{*T}$ are the loading vectors associated with χ^{*}. This estimate can easily be incorporated in the monitoring algorithm, since if no missing values are detected (i.e. $\chi^{*T} = \chi^T$) the above equation is identical to the basic representation. However, it may be wise to notify the operators if estimations are carried out. It should be pointed out that in some cases $\mathbf{P}_a{}^{*T}\mathbf{P}_a{}^{*}$ is ill-conditioned but this problem can be solved by biased regression (Nelson *et al.*, 1996).

Nonlinear relationships

Many subprocesses within wastewater treatment display nonlinear behaviour and nonlinear methods may provide a remedy if these are to be modelled. However, from a macro point of view, which is normally the view of the operators, wastewater treatment processes display surprisingly linear behaviour. There are important exceptions, for example sludge loss, but for a plant in a normal operational state the nonlinearities are often well behaved. By well behaved we mean nonlinearities that display smooth and monotonic behaviour. Monitoring is often a case of classification of the current operational state into one of two classes: normal or abnormal. For example, let a linear model approximately describe the normal region of the operational space. Deviations outside this region are driven by linear and/or nonlinear mechanisms. When a deviation is established it is often of less importance "how much" abnormal the current state is. The very fact that a deviation has been

established is serious enough to invoke actions. The situation is different when a prediction model is the objective or when there is obvious nonlinear behaviour in the normal operational region. In such cases, the only feasible solution is nonlinear modelling. Different nonlinear PCA algorithms have been proposed in the literature, e.g. Kramer (1991), Dong and McAvoy (1996) and Jia *et al*. (1998).

Changing operational conditions

Operational conditions change due to reasons such as varying raw material quality, surrounding temperature, varying process load and equipment wear. This is certainly not an ideal situation for the basic multivariate approach, which relies on the assumption that data are stationary in the time scale of interest. The way to address this problem depends on the nature of the process drift and two major cases can be distinguished. The first case originates from univariate changes in mean and variance, i.e. mean and variance are varying but the qualitative relations between variables stay the same. In this case, it is sufficient to update the scaling parameters (mean and variance) of the data as shown in Rosen and Lennox (2001).

The second case involves changes in the relations between the variables (covariance structure) in addition to changes in the mean and variance. Here, the covariance structure of the model must be updated. A straightforward way is to use a moving (rectangular) time window, on which the model is based. A more sophisticated way is by recursive means. The principle of such updating schemes is that when new data are available they are included in the data matrix according to certain weights. A number of updating schemes can be found in the literature (see e.g. Wold (1994), Dayal and MacGregor (1997) and Li *et al*. (2000)). A straightforward method is to update the covariance or correlation matrix recursively: $(\mathbf{X}^T\mathbf{X})_k = \alpha(\mathbf{X}^T\mathbf{X})_{k-1} + \mathbf{x}_k{}^T\mathbf{x}_k$, where the $(\mathbf{X}^T\mathbf{X})_k$ is the covariance structure at time k and $\mathbf{x}_k$ is the new observation vector. Here, the history is increasingly disregarded as the monitoring progresses and the speed at which this is done is dependent on the choice of the forgetting factor α. Scaling parameters must also be updated, and this is preferably done in the same manner (Rosen and Lennox, 2001). There is usually a need for an updating criterion to ensure that only data representative for the process are used in the updating of the model (Rosen, 2001).

Dynamic processes

Challenge 7 addresses the dynamics of the system. It is clear that WWT processes are dynamic processes and basic PCA does not take this into account. On the contrary, it is assumed that data is static (no autocorrelation). This introduces mainly two problems. Firstly, vital information about the internal dynamic relationships among the variables is lost. Secondly, autocorrelation will distort the model leading to less reliable monitoring results. If no empirical knowledge on time lags is available, investigation of the cross-covariance function between the variables can be used to estimate suitable time lags. Some of the variables are, thus, purposely delayed according to the cross-covariance functions to obtain a more accurate model. This is recommended if there are obvious delays in the system, such as in the plug flow case. The method can be extended to using an a priori model for the time lag of every relation, for example, depending on the retention time. By doing so, the time lag of the variables changes dynamically as the flow rate changes and we will obtain a quasi-dynamic representation of the flow rate dynamics (Röttorp, 2001). By introducing ideas from time series modelling, PCA modelling can be made truly dynamic. The $\mathbf{X}$ matrix is augmented by lagged variables ($\mathbf{X} = [\mathbf{X}_k\ \mathbf{X}_{k-1}\ \ldots\ \mathbf{X}_{k-l}]$) in such a way that the model describes the dynamics (e.g. Ku *et al*. (1995) and Tsung (2000)). However building a lagged version of the covariance structure includes the identification of many parameters

and in large data sets this becomes cumbersome. Further, in the case of stiff systems (i.e. wide range of time constants) the augmented matrix becomes huge, since slow changing variables will require a large number of lagged variables.

Multiresolution analysis (MRA) is a method based on wavelet theory for decomposing signals into separate time scales (see e.g. Strang and Nguyen (1996)). This can be utilised for multivariate monitoring purposes by decomposing variables into time scales corresponding to the major time scales of the system (Rosen and Lennox, 2001). Thus, different multivariate models can be identified for different time scales. Now, a static representation is better as an approximation of the process at each scale as a portion of the autocorrelation is removed. Also, the decomposition involves downsampling of data by a factor 2^j for scale j ($j = 1$ for the fastest scale). This facilitates the use of time series techniques, since the augmented matrix does not explode in size when slow dynamics are considered.

Example

In the following section, the above discussion is exemplified through an example. The example is taken from about 40 days of operation at Ronneby WWT plant in Sweden. To highlight the difficulties discussed earlier, the period is chosen so that it spans over significant changes in external operational conditions. In the example, 12 on-line variables are used: suspended solids, aerator valve opening (the oxygen concentrations are successfully controlled to constant set points), pH, ammonia, conductivity and flow rate, of which many are measured at several locations. The training period (10 days of dry weather – late summer) is used to identify a PCA model. The remaining 30 days (early autumn – several rain events and a temperature drop) are used for on-line monitoring (although this example is produced off-line, it is implemented in an on-line fashion, i.e. no knowledge of future data is available).

A PCA model with 5 principal components ($a = 5$) is identified. In Figure 2, basic PCA monitoring is applied. It is evident that the basic approach is not useful. Apart from the fact that the model does not stay valid for long (*SPE* exceeds its limit), the T^2 and score plots display large deviation from the normal operational region (shaded area). Many of the fast transients are due to calibration and cleaning of on-line sensors or by missing data (here represented by zero value). The prolonged deviations in *SPE* and T^2 between days 10 and 22 are caused by a rain. However, this is "normal" during the autumn and, consequently, a monitoring method must handle such incidents if it is going to be an asset in the daily operation.

A first step to improve the monitoring results is to remove non-representative data from sensitive and less reliable sensors. Another PCA model, based on the derivatives of the sensor signals not considered fully reliable, is used for the screening. This makes it possible to discern deviations caused by calibration/cleaning activities from those that may be true problems (the calibration/cleaning is carried out at the same location and generally at the same time). When the screening is applied, the data are only complete at about 50% of time, which is probably typical for this type of application. Thus, the incomplete data have to be complemented by estimation and this done according to the discussion above. In Figure 3, PCA is applied to the data containing estimates for unreliable and missing values. Although better than without screening, the T^2 and score plots still reveal far to many deviations from normal operation due to the changing operational conditions rather than abnormal events. Especially evident is the rainy period. Here, the model shows poor fit and it is not possible to discern any deviations except the one caused by the rain (high flow rate).

Implementing adaptive scale parameters improves the results considerably (Figure 4). The adaptation speed is chosen so that the model adapts to slow variations but not to diurnal variations. Here, the SPE and T^2 are within their detection limits most of the time. Also note

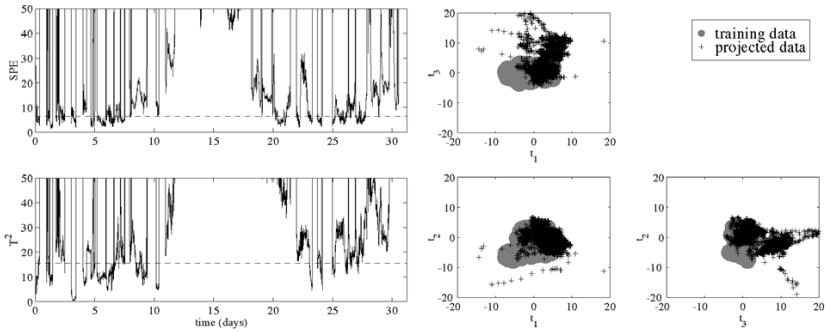

Figure 2 Basic PCA applied to the data from Ronneby WWT plant. *SPE* and T^2 with approximate 99% detection limits (left) and score plots (right). Note that some observations are far outside the range of the *SPE*, T^2 and score plots

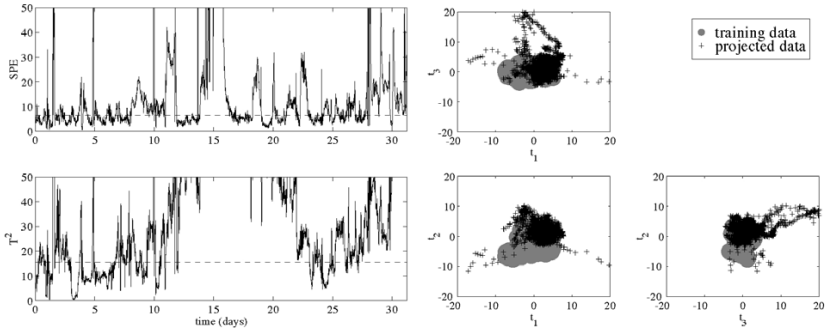

Figure 3 PCA with missing and uncertain values replaced by estimations. *SPE* and T^2 with approximate 99% detection limits (left) and score plots (right)

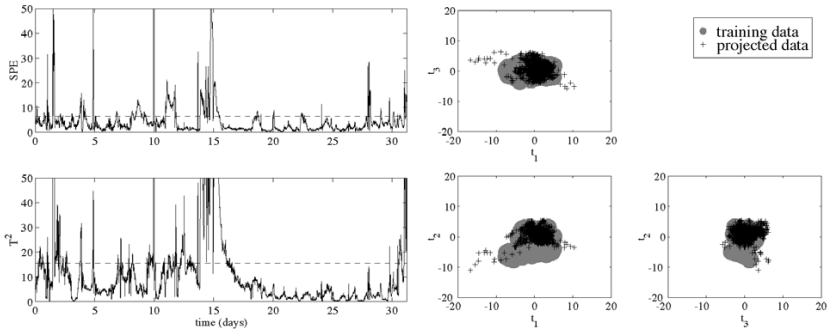

Figure 4 PCA with adaptive scaling parameters. *SPE* and T^2 with approximate 99% detection limits (left) and score plots (right)

that the score plots display only a few deviations from the region defined by the training data. It is clear that PCA with adaptive scaling parameters best describes the process in terms of number of true deviations. In hindsight, it is possible (through manual analysis) to distinguish a few events that clearly deviate from normal operational behaviour. These events are all detected by the PCA and by contribution analysis the deviating variables are isolated. However, the extended PCA algorithm also detects some additional events, especially close to the alarm limits. These can be removed by changing the limits, which can be justified since the assumptions for calculating limits can only be regarded as partly fulfilled.

On-line monitoring using multivariate methods is often an issue of minimising the number of false alarms while true deviations are retained and detected. This is achieved through replacement of missing and erroneous data and by implementing adaptive scaling parameters. Here, no consideration is taken to the dynamics. This may in some cases further improve the model and its ability to detect deviations. However, we have shown that good operational monitoring results can be obtained if basic PCA is extended by these rather simple means.

Conclusions

In this paper, we have tried to highlight some difficulties with multivariate monitoring of WWT operation. We have also suggested remedies, originating from recent research within the multivariate research community. What at a first glance may seem rather hopeless can, assuming that some of the ideas presented in this paper are implemented, turn out to be fruitful. A number of conclusions, or rather guidelines, can be discerned:

- Data quality improvement
  - Omit variables that are extremely noisy or relatively constant (no or little information content).
  - Implement data screening (using another PCA model is one possible method).
  - Implement the monitoring model so that it handles missing data with notification to the user.
  - Use filtered signals only if the detection speed is of less importance, since multivariate models has an inherent noise reduction capability.
- Use a non-linear model if a linear model does not suffice (but start with a linear model).
- If the process is non-stationary
  - Implement a scale parameter adaptive model to retain the possibility to use score plots.
  - Implement a fully adaptive model if the above is not sufficient.
- To represent relationships quasi-dynamically
  - Estimate time lags using experiments.
  - Estimate time lags using cross-covariance functions.
- To represent relationships dynamically
  - Implement a time lagged augmented covariance structure if the time constants of interest do not range over too wide a time span.
  - When dynamics with a wide range of time constants are to be modelled, decompose the data into two or more time scales and identify static or dynamic models for each scale.

Most of these suggestions imply a complexity increase of the monitoring task. If the authors were to choose only one or two improvements, missing data and adaptive capabilities would be prioritised.

References

Dayal, B.S. and MacGregor, J.F. (1997). Recursive exponentially weighted PLS and its application to adaptive control and prediction, *J. Process Control*, **7**(3), 169–179.

Dong, D. and McAvoy, T.J. (1996). Nonlinear principal component analysis – based on principal curves and neural networks, *Comput. Chem. Eng.*, **20**(1), 65–78.

Grung, B. and Manne, R. (1998). Missing values in principal component analysis, *Chemometrics Intell. Lab. Syst.*, **42**, 125–139.

Jackson, J.E. and Mudholkar, G.S. (1979). Control procedures for residuals associated with principal component analysis, *Technometrics*, **21**(3), 341–349.

Jia, F., Martin, E.B. and Morris, A.J. (1998). Non-linear principal component analysis for process fault detection, *Comput. Chem. Eng.*, **22**(Suppl.), S851–S854.

Kourti, T., Lee, J. and MacGregor, J.F. (1996). Experiences with industrial applications of projection methods for multivariate statistical process control, *Comput. Chem. Eng.*, **20**, 745–750.

Kramer, M.A. (1991). Nonlinear principal component analysis using autoassociative neural networks, *AIChE J.*, **37**(2), 233–243.

Kresta, J.V., MacGregor, J.F. and Marlin, T.E. (1991). Multivariate statistical monitoring of process operating performance, *Can. J. Chem. Eng.*, **69**, 35–47.

Ku, W., Storer, R.H. and Georgakis, C. (1995). Disturance detection and isolation by dynamic principal component analysis, *Chemometrics Intell. Lab. Syst.*, **30**, 179–196.

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

MacGregor, J.F., Jaeckle, C., Kiparissides, C. and Koutoudi, M. (1994). Process monitoring and diagnosis by multiblock PLS methods, *AIChE J.*, **40**(5), 826–838.

Nelson, P.R.C., Taylor, P.A. and MacGregor, J.F. (1996). Missing data methods in PCA and PLS: score calculations with incomplete observations, *Chemometrics Intell. Lab. Syst.*, **35**, 45–65.

Rosen, C. (2001). *A Chemometric Approach to Process Monitoring and Control – With Applications to Wastewater Treatment Operation*, PhD. thesis, Dept. of Industrial Electrical Engineering and Automation, Lund University, Lund, Sweden.

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

Rosen, C. and Olsson, G. (1998). Disturbance detection in wastewater treatment plants, *Wat. Sci. Tech.*, **37**(12), 197–205.

Röttorp, J. (2001). Implementation of multivariate real-time methodologies for industrial process control, *Proc. 7th Scandinavian Symposium on Chemometrics*, Aug. 19–23, 2001, Copenhagen, Denmark.

Strang, G. and Nguyen, T. (1996). *Wavelets and filter banks*, Wellesley-Cambridge Press, Wellesley, Massachusetts, USA.

Teppola, P., Mujunen, S.-P. and Minkkinen, P. (1998a). A combined approach of partial least squares and fuzzy c-means clustering for the monitoring of an activated-sludge waste-water treatment plant, *Chemometrics Intell. Lab. Syst.*, **41**, 95–103.

Teppola, P., Mujunen, S.-P., Minkkinen, P., Puijola, T. and Pursiheimo, P. (1998b). Principal component analysis, contribution plots and feature weights in the monitoring of sequential process data from a paper machine wet end, *Chemometrics Intell. Lab. Syst.*, **44**, 307–317.

Tsung, F. (2000). Statistical monitoring and diagnosis of automatic controlled process using dynamic PCA, *Int. J. Prod. Res.*, **38**(3), 625–637.

Wold, S. (1994). Exponentially weighted moving principal components analysis and projection to latent structures, *Chemometrics Intell. Lab. Syst.*, **23**, 149–161.

Yoon, S. and MacGregor, J.F. (2001). Fault diagnosis with multivariate statistical models part I: using steady state fault signatures, *J. Process Control*, **11**, 387–400.