International Environmental Modelling and Software Society (iEMSs)
2012 International Congress on Environmental Modelling and Software
Managing Resources of a Limited Planet, Sixth Biennial Meeting, Leipzig, Germany
R. Seppelt, A.A. Voinov, S. Lange, D. Bankamp (Eds.)
http://www.iemss.org/society/index.php/iemss-2012-proceedings

Efficient data quality evaluation in automated water quality measurement stations

Janelcy Alferes[1], Pascal Poirier[1], Peter A. Vanrolleghem[1]
[1]model*EAU*, Département de génie civil et de génie des eaux. Université Laval, Québec, QC, Canada. (Email: Janelcy.Alferes@gci.ulaval.ca)

Abstract: In this paper, software tools for automatic data quality assessment with a practical orientation are proposed. Two different approaches are presented that use time series information. First, univariate methods based on autoregressive models are applied for data correction (outliers detection for data replacement). Faults are detected by defining acceptable thresholds to data features and to the residuals' standard deviation (RSD). Second, multivariate statistical methods based on Principal Components Analysis are used to extract correlations between variables from data sets and performing fault detection using the T^2 and Q statistics. The proposed tools are successfully tested on river water quality time series obtained from *in situ* monitoring stations collecting a large amount of physical and chemical variables.

***Keywords**:* in situ monitoring stations; data quality assessment; fault detection

1 INTRODUCTION

Effective management of water bodies requires having reliable information about water quality. Nowadays, implementation of *in situ* continuous monitoring at high frequency is being used to collect water quality information of surface waters. Along rivers and water networks, on-line measuring campaigns over long periods are conducted on to identify spatial and temporal variations in water quality, trends and also analyze the variability of the polluting sources [Langeveld et al., 2011]. Additionally, the use of automated on-line measuring systems can reduce the total monitoring costs [Pressl et al., 2004].

An important change can be noticed from having not sufficient data (grab or composite samples) to huge and complex data sets consisting of a large number of physical-chemical parameters. Such data usually is affected by different sources of errors and uncertainties [Rieger and Vanrolleghem, 2008]. Since measurements are carried out in a very severe and difficult environment, sensors are subject to many functional, technical and operational constraints. Despite of the important efforts of manufacturers including for example self-cleaning systems, the reliability of sensors remains frequently insufficient. Degradation of measuring conditions, clogging and progressive fouling by grease, solids and other wastes is usual (Yoo et al., 2007). Additionally, real hydrological data are mostly noisy, not normally distributed, and often co-linear or autocorrelated. Efficient monitoring and proper understanding and use of collected measurements in further applications depends therefore on careful data evaluation and validation to ensure data quality. Detection of corrupted, doubtful and/or unreliable data, outliers, noise, missing values and potential sensor faults becomes crucial.

Corrupt data can be identified and replaced/removed by different methods ranging from logical algorithms to more sophisticated statistics or model-based methods. However, in current practice data validation is most often carried out with a time-consuming and inefficient manual procedure based on basic data and visualization tools. Some software tools for data quality evaluation in urban hydrology can be

found in the literature based on rules to detect doubtful and/or unreliable data using parametric tests [Mourad and Bertrand-Krajewski, 2002; van Bijnen and Korving, 2008]. Some statistical methods have been developed for the same purpose in the last years, but only few of them have been implemented in software platforms for practical use in the water sector [Branisavljevic et al., 2010]. A lot of work has been done in the fault detection and diagnosis field covering model-based and data-based methods [Venkatasubramanian et al., 2003]. However, how to integrate them effectively for practical applications still remains an important topic for further research. Multivariable methods have also been used for analysing environmental data and drawing meaningful information [Alkarkhi et al., 2008]. Concerning water quality, these studies have been focused only on water samples taken in surface waters according a monitoring plan.

Given the large amount of data typically collected with continuous monitoring, in this paper, software tools for automatic data quality assessment with a practical orientation are proposed. Using time series information, the laborious manual validation procedure is replaced by automatic methods for data correction and fault detection. While univariate analysis based on autoregressive models is used for detection and replacement of doubtful data, multivariable analysis based on principal component analysis (PCA) is used to extract correlation and significant information between variables.

2 IN SITU MONITORING STATION

A Primodal Systems' RSM30 Monitoring Station (Figure 1a) has been installed to measure the water quality dynamic of the small urban river Notre Dame located in Ancienne-Lorette, Quebec, Canada. The measurement campaigns of this study covered the summer periods of 2010 and 2011. The measurement station comprises sensors for conventional, physical-chemical parameters (temperature, dissolved oxygen…) as well as innovative sensors like a UV spectrometer and an ion selective device (ISE-sensor). Water level is also recorded. All sensors are permanently submerged in a secured cage dropped on the river bed (Figure 1b).

(a) (b)

Figure 1. (a) Data acquisition cage, (b) Probe-holder cage

Table 1 gives a detailed list of the sensors installed at the monitoring station. The location of the monitoring station allows both continuous monitoring of the Notre Dame River and the impact of the urban area discharge on the river water quality. For that, all sensors recorded data at short intervals (between 5-60 seconds) generating rich-information data sets. To achieve a good data quality of the on-line measurements a systematic calibration and maintenance routine is critical. However, due to environmental conditions access to the probe-holder cage was not always possible, especially after rainy periods. Missing calibration steps and other technical problems like faulty probes and vandalism actions resulted in some gaps in the data.

Table 1 Measured water quality parameters

Probe	Parameter	Unit	Sampling (sec)
Hach pHD sc	pH	-	5
	Temperature	℃	
Hach sc100 Inductive	Conductivity		5
conductivity	Temperature	℃	
Hach LDO	Dissolved oxygen (DO)	mg/l	60
	Temperature	℃	
s::can ISE	Potassium (K^+)	ppm	60
	Ammonia (NH_4)	ppm	
	Temperature	℃	
	pH	-	
s::can spectro::lyser	Nitrates (NO_3)	mg/l	60
	Total organic carbon (TOCeq)	mg/l	
	Dissolved organic carbon (DOCeq)	mg/l	
	Turbidity	FTUeq	
Hach Solitax sc	Total suspended solids (TSS)	g/l	5
Sigma 950 flow meter	Level	m	60

3 AUTOMATIC DATA QUALITY ASSESSMENT TOOLS

Two different approaches for automatic data quality assessment are presented that use on-line time series information (Figure 2) in the absence of exact process knowledge. On the one hand, univariate methods are aimed to extract information from single measurement variables. Their proposed implementation can be divided in two main steps: outliers detection and fault detection. On the other hand, given the high dimensional measurement space, multivariable methods are used first to detect and remove correlations among variables and reduce their dimensionality and are then used for fault detection. It is important to note that both methods can be tuned to provide a more or less restrictive performance.

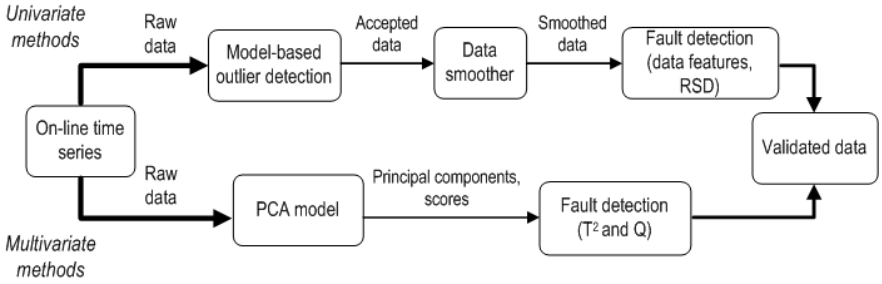

Figure 2. Proposed software tools for data series validation

3.1 Univariate analysis

The proposed tool is based on forecasting of time series data by means of autoregressive models. The first step includes the outlier detection and data replacement to generate a proper time series that can be effectively used in further steps as shown in Figure 2. An outlier is a sample value that differs notably from the mean of the measurement series. Since it could significantly affect data features, outliers must be removed or replaced. The proposed outlier detection method compares measured values with calculated forecast values by defining a dynamic prediction interval.

At time T, the forecast value of the data *x* in the next time unit, T+1, is predicted using a third-order exponential smoothing model. To give better estimations of the local variance, which is translated in more reliable prediction intervals, a simple exponential smoothing model is also defined to predict the standard deviation of the forecast error. For estimation of the variance of the forecast errors, σ_e^2, a mean absolute deviation Δ is first defined. Then, according to Montgomery et al. [2009],

the estimate of σ_e^2 at time T is given by $\hat{\sigma}_{e,T} = 1.25\hat{\Delta}_T$ and the estimates of Δ are calculated as follows:

$$\hat{\Delta}_T = \delta|e_T(1)| + (1-\delta)\hat{\Delta}_{T-1} \quad (1)$$

where $e_T(1)$ is the one-step-ahead forecast error calculated as $e_T(1) = x_T - \hat{x}_T(T)$. The term $\hat{x}_T(T)$ represents the one-step-ahead forecast value made at time T. The factor δ is a smoothing parameter, typically between 0.01 and 0.3, which controls to what extent the past observations influence the forecast. The prediction interval x_{lim} is then defined by adding or subtracting a multiple of the standard deviation of the forecast error to the forecast data value as follows:

$$x_{lim} = \hat{x}_T(T) \pm K\hat{\sigma}_{e,T} \quad (2)$$

where $\hat{\sigma}_{e,T}$ represent the one-step-ahead forecast standard deviation made at time T. K is a multiplicative factor that can be adjusted to make the model more or less restrictive. If the measurement data falls outside the prediction interval, it is considered as an outlier. In this case the outlier is replaced by the forecast data value. The resulting data series is called *accepted data.* For data validation purposes, the *accepted data* is then smoothed using a kernel smoother [Schimek, 2000] with a 13-samples bandwidth. More effective results without the corruption with signal noise are obtained when smoothed data is used to calculate features in the data.

For model evaluation and fault detection purposes, some data features are calculated. Faults are subsequently detected by applying acceptable limits to data features. To give an indication about the goodness of the smoothed data (once the outliers have been replaced for the forecast values) the fraction of forecast values used by the smoother is represented. The slope in the smoothed data is also calculated to provide information about the rate of change of the variable. Errors in the model are assumed to be normally and independently distributed with mean zero and constant covariance. Diagnosing the residuals (calculated as the difference between the accepted and smoothed data) is useful to check normality and good fit of the model to the raw data. When autocorrelation in the residuals time series is detected either the smoothed data is not representative of the real measurements or the noise presents a non-random distribution. Autocorrelation of the residuals is analysed by carrying out a runs test on a 30-samples moving window [Dochain and Vanrolleghem, 2001]. Finally, the variance of the data is determined by calculating the residuals standard deviation (RSD). Horizontal lines in Slope and RSD plots represent the determined acceptability limits for each feature according to expected realistic values in the field. Concerning the Runs test value plot, the limits correspond to the 95% confidence interval.

3.2 Multivariate analysis

Real water quality data are mostly redundant, non-stationary and often auto and cross-correlated. For exploration and interpretation of large-dimensional multivariate datasets that are highly correlated, multivariate methods can be applied to reduce the dimension of the variable matrix space into a more accessible low-dimensional space identifying key variables.

Proposed multivariate statistical methods are based on principal components analysis (PCA). Unlike other methods, PCA has been revealed as a robust technique with a low computational demand and straightforward use [Villez et al., 2009]. This technique searches a new set of uncorrelated and orthogonal variables, called principal components (PCs) which explain most of the data variability in a new coordinate system. Each principal component is a linear combination of the original variables and describes the largest process variability in a space of fewer dimensions than the original one. With X an autoscaled [*m x n*] matrix of

measurement values for *n* variables and *m* samples, the covariance matrix Cx [*m x m*] is computed as follows:

$$Cx = \frac{1}{m-1} X^T X \tag{3}$$

Cx captures the covariance between all possible pairs of measurements. Performing the singular value decomposition (SVD), Cx is diagonalized by the orthogonal matrix of its eigenvectors $P = [p_1\ p_2\ \dots\ p_n]$, called *loadings*. The principal components of X are the columns of P and the corresponding eigenvalues $[\lambda_1\ \lambda_2\ \dots\ \lambda_n]$ represent the variance of X along each principal component p_i. In the new coordinate system, the transformed data, called *scores*, are represented by $T = XP$. A graphical representation of the PCA projection is shown in Figure 3.

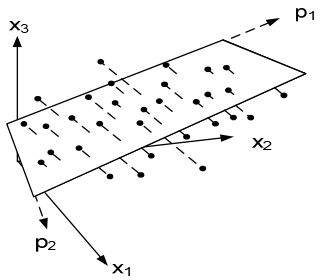

Figure 3. Dimensionality reduction using PCA

Sorting the columns of P in decreasing order, p_1 corresponds to the largest eigenvalue λ_1 and it is oriented in the direction of the largest variation of the original variables capturing the largest fraction of the data variance. A dimension reduction can be obtained by retaining a number of components $a < n$. In this case, the original data space can be expressed as:

$$X = TP^T + E \tag{4}$$

where E represents the residual matrix which contains the components corresponding to the less significant eigenvalues. Choosing the number of principal components α is crucial to obtain a descriptive PCA model as a trade-off between dimension reduction and variability captured by the model. The method based on the eigenvalue scree plot [Jolliffe, 2002] is used. Once the PCA model is obtained new data can be projected onto the existing model preserving the matrix P.

In order to properly interpreting the PCA results and for fault detection purposes two statistics which describe the statistical fit of the model are calculated. Plotting these statistics with appropriate confidence limits allows detecting deviations of the measurements from the normal behavior. The first statistic, called T^2, which is the normalized sum of scores, captures the variations in the reference data model. At time *k*, T^2 is calculated as:

$$T^2(k) = x^T(k) P \Lambda^{-1} P^T x(k) \tag{5}$$

where x(k) is the measurement vector and Λ the diagonal matrix of the eigenvalues associated with the retained principal components. The confidence limit T_α^2 for T^2 is obtained using the F- distribution [Yoo et al., 2007]. The second statistic, called Q, is defined as the sum of squared residuals of the active principal components. At time *k*, Q is calculated as:

$$Q(k) = x^T(k) \left(I - PP^T \right) x(k) \tag{6}$$

The confidence limit Q_α for Q is computed under the assumption of normally distributed scores according to Montgomery [2009]. In general, Q captures the variation in the residual space not accounted for by the PCA model. For a new measurement, if $T^2 < T_\alpha^2$ and $Q < Q_\alpha$ it is considered that the process is in control with 100(1-α) % of confidence, α being a level of significance.

4 RESULTS

Some results about the application of the described methods are shown in this section. The univariate methods have been successfully tested on on-line time series of the different water quality parameters in Table 1. Figure 4 shows the behaviour of the outlier detection method for a short period of level measurements. The dynamic calculation of the prediction interval lets the algorithm adapt to the time-varying hydraulic behaviour visible in the data. Most of the data falls into the prediction interval with Raw and Accepted data almost coinciding, except for some periods associated with important changes in the level measurements as indicated for example in Figure 4a. A less restrictive version of the model in Figure 4b leads to accept the raw data initially rejected.

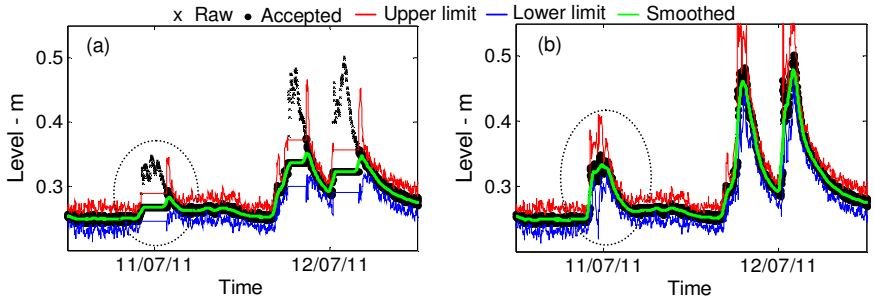

Figure 4. Outlier detection method for on-line level measurements. (a) Restrictive case. (b) Less restrictive case.

Figure 5 shows the overall results over a short TSS time series. The impact of two rain events on the TSS behaviour is clearly observed. Hydraulic variations due to rain events directly affected the mixing conditions and the suspended solids concentration in a significant way. The TSS concentration was increased almost tenfold from the normal values in the first rain event.

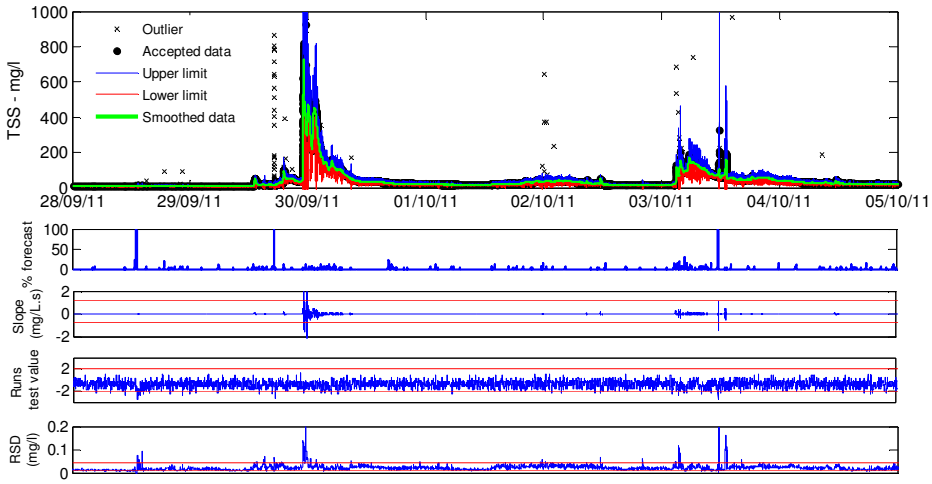

Figure 5. Application of univariate methods over TSS on-line measurements

It can be seen how several outliers were detected by the algorithm around these episodes increasing the fraction of forecast data used by the smoother. Although the slope in the smoothed data remained inside the limits during most of the analysed period, larger slope values were detected around rain events evidencing the important dynamics in the variable. Abnormal slope values were observed around September 30th. Run tests have shown that most of the data fall into the 95% interval (-2, 2) suggesting the adequacy of the model. Periods in which some residuals correlation was detected are related with an insufficient performance of the smoother which is averaging important peaks in the data. This coincides with a

higher percent of forecast values used in the smoother and larger slope values. The RSD values also confirm the high variance in the data around the rain events. It is important to highlight that acceptability limits and also the model and smoother parameters can be adjusted to make them more or less restrictive in the fault detection phase.

Concerning the multivariate methods, the tool has been tested using different groups of on-line variables in Table 1. The following examples show the results obtained considering time series of Turbidity, NO_3, TOCeq, DOCeq, pH, K^+, NH_4 and Temperature. Before the application of the PCA algorithm, all variables have been properly autoscaled (mean centering and variance scaling). Figure 6 shows the percentage of the total variability explained by each principal component for this data matrix. It can be seen in Table 2 that the first four principal components capture more than 90% of the variance in the process.

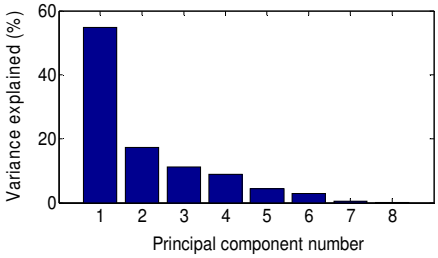

Figure 6. Relation between the principal components and the captured variance

Table 2. Results of the PCA model

p_i	λ_i	% explained variance	%cum. variance
1	4.38	54.77	54.77
2	1.39	17.36	72.12
3	0.91	11.33	83.45
4	0.70	8.79	92.24
5	0.35	4.36	96.60
6	0.23	2.86	99.46
7	0.04	0.51	99.97
8	0.00	0.03	100.00

To illustrate the capacities of the multivariate methods, the following figures show some results for short on-line time series. Figure 7a shows the scores and the coefficients for the two first principal components for each observation. Each variable is represented by a vector and its length and direction indicate the contribution of the variable to the two principal components. Each point in the plot corresponds to a sample and its location indicates the score of each sample in the two principal components space. Points that cluster represent similar behaviour, and deviating points indicate process changes. Due to the mean centering of data, under normal operation points should be close to the origin. Some outlying points can be for example identified in the marked areas (I, II) suggesting an abnormal behaviour or disturbance for these samples. Graphical representation of T^2 and Q statistics in Figure 7b also illustrate some fault situations in the process. In period I for example T^2 accounted for a fault associated with abnormal variations within the model subspace in the NO_3 measurements. In period II T^2 revealed some abnormal variations in the Turbidity, DOCeq and TOCe measurements; but Q also identified events not taken into account in the current realization of the model.

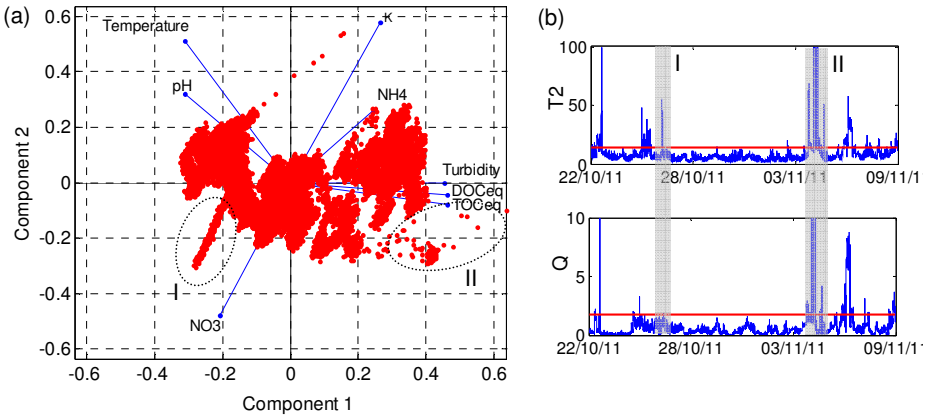

Figure 7. (a) Scores for the first and second principal components. (b) T^2 and Q statistics. Abnormal behaviour occurs in periods I and II (see text)

5 CONCLUSIONS

Automatic data quality and assessment tools for analysis of time series, based on univariate and multivariate methods, have been presented and successfully validated on complex data sets obtained from automated water quality measurement stations. The application of the univariate methods for identification and replacement of outliers allowed creating "good" time series that can be properly used in further analysis steps. Calculation of data features using smoothed data allows model evaluation and a better understanding of the time series, making possible the identification of possible faults or abnormal behaviours. The application of the multivariate methods has allowed dimension reduction and the identification of key variables that capture the most significant variability in the complex data set. Monitoring of the different data quality statistics has resulted effective and applicable to detect multiple sensor faults and also the statistical fit of the model.

ACKNOWLEDGMENTS

John Copp (Primodal) is thanked for his overall support with the station and some software upgrades. Peter Vanrolleghem holds the Canada Research Chair in Water Quality Modelling. The CFI Canada Research Chairs Infrastructure Fund project (202441) provided the monitoring stations.

REFERENCES

Alkarkhi A., A. Ahmad., N. Ismail, A. mat Easa, and K. Omar, Assessment of surface water through multivariate analysis. *J. Sust. Develop., 1*(3), 27-33, 2008.

Branisavljevic N., D. Prodanovic and D. Pavlovic, Automatic, semi-automatic and manual validation of urban drainage data, *Wat. Sci. Tech.*, *62*(5), 1013-1021, 2010.

Dochain D. and P.A. Vanrolleghem, Dynamical Modeling and Estimation in Wastewater Treatment Processes, IWA Publishing, London, UK, 2001.

Jolliffe, I.T., Principal Component Analysis, second ed. Springer, Berlin, 2002.

Langeveld J.G., R.P.S. Schilperoort, S.R. Weijers, J. De Jonge and T. Flameling, Climate change and urban wastewater infrastructure: there is more to explore, 8th IWA Leading-edge Conference on Water and Wastewater, Amsterdam, 2011.

Montgomery D.C., Introduction to statistical quality control, 6th edition, John Wiley&Sons, New York, 2009.

Mourad M. and J.L. Bertrand-Krajewski, A method for automatic validation of long time series of data in urban hydrology, *Water Sci Tech.,* 45(4-5), 263-270, 2002.

Pressl A., S. Winkler and G. Guber, In-line river monitoring – new challenges and opportunities, *Wat. Sci. Tech., 50*(11), 67-72, 2004.

Rieger L. and P.A. Vanrolleghem, monEAU: a platform for water quality monitoring networks, *Wat. Sci. Tech. 57*(7), 1079-1086, 2008.

Schimek M.G., Smoothing and Regression: Approaches, Computation and Application, John Wiley&Sons, New York, 2000.

van Bijnen M. and H. Korving, Application and results of automatic validation of sewer monitoring data. 11th International Conference on Urban Drainage, Edinburgh, Scotland, UK, 2008.

Venkatasubramanian V., R. Rengaswamy, S.N. Kavuri and K. Yin, A review of process fault detection and diagnosis. Part III. Process history based methods, *Comp. Chem. Eng., 27*(3), 327-346, 2003.

Villez K., M. Ruiz, G. Sin, J. Colomer, C. Rosén and P.A. Vanrolleghem, Combining multiway principal component analysis and clustering for efficient data mining of historical data sets of SBR processes, *Wat. Sci. Tech., 57*(10), 1659–1666, 2008.

Yoo C.K., K. Villez, S.W.H Van Hulle and P.A. Vanrolleghem, Enhanced process monitoring for wastewater treatment systems, *Environmet.*, *19*, 602-617, 2007.