International Environmental Modelling and Software Society (iEMSs)
7th Intl. Congress on Env. Modelling and Software, San Diego, CA, USA,
Daniel P. Ames, Nigel W.T. Quinn and Andrea E. Rizzoli (Eds.)
http://www.iemss.org/society/index.php/iemss-2014-proceedings

Automated data quality assessment: Dealing with faulty on-line water quality sensors

Janelcy Alferes[1] **and Peter A. Vanrolleghem**[1]

[1]*modelEAU, Département de génie civil et de génie des eaux. Université Laval*
Québec, QC, Canada, Janelcy.Alferes@gci.ulaval.ca

Abstract: Compared to traditional grab sampling modern measurement systems enable continuous water quality monitoring of water systems at high frequency. However, in real world applications on-line sensors are still subject to functional, technical and operational constraints. Challenges thus remain associated with the automation of data collection and especially data validation to ensure proper use and interpretation of the data and avoid the danger of building data graveyards. Poor quality data could drastically affect the results of their application, e.g. water quality models for river basin management, model-based control, WWTP control rules, decision making, etc. For practical fault detection purposes, in this paper, a data-driven tool that attempts to extract useful information from the time series of multiple measurement signals, in the absence of exact process knowledge, is presented. The proposed tools are successfully tested on on-line water quality time series from different applications including sewers, wastewater treatment plants and receiving waters.

Keywords: data quality assessment; fault detection; on-line instrumentation.

1. INTRODUCTION

One of the obstacles for the joint use of monitoring and modelling has been the lack of sufficient and good data which affects the applicability of models that rely on them. Recent developments regarding advanced on-line water quality instrumentation and data acquisition systems have contributed to the gradual implementation of automatic in-situ monitoring systems for generation of high frequency data. Nevertheless, the intrinsically challenging measurement conditions of the water environments monitored makes that on-line sensors are still affected by many faults resulting in large amounts of data with doubtful quality being collected (Branisavljevic et al., 2010). Data will only be useful for their meant application if it is reliable and correctly validated.

Additional and important efforts are then required to assess the quality of the data to ensure that a fully utilizable database of meaningful values is constructed. In this data validation process the detection and isolation of potential sensor faults becomes crucial. In practice, typical faults that can affect water quality sensors (bias, drift, precision degradation, complete failure...) are not easy to detect (Yoo et al., 2008). Inefficient manual inspection and visualization are still the most applied procedures. Such approach becomes unachievable when large data sets need to be scrutinized (Thomann, 2008).

In this paper a novel multivariable model-based method with a practical orientation is presented in order to automatically detect possible deviations of water quality sensors from their normal operation. A principal component analysis (PCA) scheme is used for diagnosis purposes to first detect and then isolate and identify the cause of the fault by using different statistical metrics and the contribution by the different variables. The proposed tools have been successfully tested on water quality time series collected in different water systems. Assessed data is then available for different purposes, among others, for modelling of different biological processes, resource management and pollution description.

2. METHODS

2.1 On-line water quality sensors

An automated monitoring system encompasses sensors and recording systems to measure physical and chemical water quality variables at discrete time intervals at point locations. In this way, a water quality monitoring station provides a nearly continuous record of water quality variables, at high time resolution, that can be used among others to describe changes and dynamics in pollution, constituent loads, identify cause-and-effect relationships and trends, understanding the effect of loads on receiving waters, model adaptation, calibration and validation for wastewater systems and receiving waters and finally for on-line control purposes (Sonnenberg et al., 2010; Caradot et al., 2011). Data from on-line sensors can also be used to estimate other constituents if a significant correlation can be established, often by regression analyses.

The challenge of automated monitoring programs is to collect data that consistently represent the water quality. This is exactly the operational goal of a water quality monitoring program: to obtain the most accurate and complete record possible. This requires clear protocols for data collection, quality assurance and quality control. On-line water quality sensors for field deployment require careful maintenance routines, as well as systematic procedures for the storage and handling of data records (Wagner et al., 2006). Even if exhaustive practical procedures can be applied to measuring systems to avoid degradation of the measuring quality, many conditions can influence the quality of measurements and lead to wrong values or faults. In addition to fouling problems, sensors can be affected by sedimentation, debris, ice, clogging and equipment malfunctioning. A validation step of the measurements being collected is then required to detect abnormal values and separate them from valid values. Given the large amount of data, automated tools for data validation need to be applied.

Figure 1. Examples of faulty sensors.

According the mon*EAU* vision (Rieger and Vanrolleghem, 2008) proposing a framework for a new generation of monitoring stations (Figure 2) the “data quality validation module” is a key component to ensure that the quality of the data being collected is sufficient for the intended application. Recent steps forward have been made concerning the development of practical software tools for fault detection purposes as described in the next section.

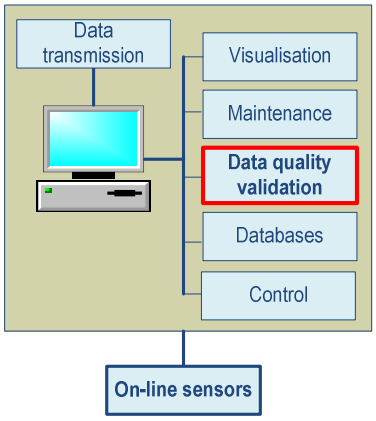

Figure 2. Software framework monitoring station (Rieger and Vanrolleghem, 2008).

2.2 Faults monitoring for on-line water quality sensors

Although different methods have been developed for fault isolation (Venkatasubramanian et al. 2003; Alcala and Qin, 2010), statistical process monitoring (SPM) has become a widely used technique, mainly in the chemical process industry. Among those methods, the principal components analysis (PCA) has won popularity thanks its data-driven characteristics, without the need of exact process knowledge. Basically, this technique looks for extracting a few independent components from highly correlated data, preserving the most relevant information of the original data set. Those key components can posteriorly be used to monitor the process operation (Lee and Vanrolleghem, 2003).

In comparison to the conventional univariate statistical process usually based on control charts to monitor individual variables, the application of the PCA technique to typically highly correlated water quality data represents a significant step forward when analyzing the data quality of multiple variables. Moreover, since PCA uses historical data to build an approximate model to summarize the measured process data, it becomes especially applicable for monitoring stations that generate large water quality data sets. For monitoring purposes, a PCA model is first obtained by using the collected process data under "normal" operating conditions. On the basis of these data control limits can be set for certain monitoring statistics and when new process data come in, they can be evaluated by monitoring the violation of these statistics to their control limits.

The applied method is shown in Figure 3. Two phases can be discerned. While the first, off-line stage is aimed at obtaining the PCA model based on a training data set, the second on-line stage intends to detect abnormalities for a new observation data vector. In the first phase, once outliers have been removed from the raw data (details can be found in Alferes et al., 2013a), the resulting data matrix X $[n \text{ x } m]$ of *n* regular-sampled observations and *m* process variables X is first normalized to a matrix $\bar{X}$ with zero mean and unit variance. The normalized matrix is decomposed as $\bar{X} = TP^T + \tilde{X}$, where T $[n \text{ x } a]$ and P $[m \text{ x } a]$ represent the scores and loadings respectively, and $\tilde{X}$ the residual matrix. Original data is transformed in this way into a new reduced dimension space characterised by a principal components. The key of the method lies therefore in the proper estimation of the transformation matrix P. In fact, the columns of P are actually the eigenvectors with the a largest eigenvalues $[\lambda_1 ... \lambda_a]$ of the correlation matrix R of the variables, which can be approximated as: $R \approx (n-1)^{-1} \bar{X}^T \bar{X}$. Each chosen eigenvector or principal component (PC) captures the maximum amount of variability in the data in an ordered manner. The residual matrix that contains the remaining components then represents the variability due to process noise. Once the PCA model has been set and data have been transformed, multivariate statistics can be calculated and multivariate control charts can be built for fault detection purposes.

More specifically, two statistics, the Hotelling t^2 and the squared prediction error Q are computed based on the projections of the data in the model and residual subspace respectively. While the Q index indicates the extent to which each sample conforms to the PCA model (measure of the amount of variation not captured by the model), the t^2 index measures the amount of variation in the model subspace. Details on their calculations can be found in Alferes et al. (2013b). The approximated control limits with a certain confidence interval α are determined from the "normal" operating data by applying probability distribution assumptions. Limits are then calculated as: $t_\alpha^2 = \left(a(n^2-1)/n(n-a)\right) F_\alpha(a, n-a)$ and $Q_\alpha = \theta_1 \left(c_\alpha h_0 \sqrt{2\theta_2}/\theta_1 + 1 + \theta_2 h_0 (h_0 - 1)/\theta_1^2 \right)$; with a and $n-a$ degrees of freedom, $F_\alpha(a, n-a)$ the upper limit of the Fisher distribution and c_α the normal distribution with α level of significance. θ_i is given by $\theta_i = \sum_{j=a+1}^{m} \lambda_j^i$ with $i = 1, 2, 3$ and $h_0 = 1 - 2\theta_1\theta_3/3\theta_2^2$.

It is expected that an abnormal or faulty situation will cause at least one of the two indices to exceed the control limits. Since the confidence limits are obtained in a statistical sense, the number of "normal" observations must be sufficiently large to result in consistent thresholds. For monitoring purposes, called on-line analysis, a new normalized data observation vector $\bar{x}_k$ collected at time instant k is projected onto the obtained PCA model and faults are detected by evaluating whether the two statistics (based on these projections) fall in the in-control region defined by the thresholds previously established. In case a fault is detected, a consecutive analysis is carried out to identify and isolate the root cause of the faulty situation in both statistics. In this case the individual contribution of

each variable to the fault detection indices is calculated. For the Q statistic, which measures the lack of fit of the sample to the model, the individual contributions for each data observation vector $\bar{x}_k$ are calculated as the k row vector of the residual matrix $\tilde{X}$. For t^2 statistic, the vector of individual contributions $t^2_{cont,k}$ is calculated as the weighted contribution of each score by $t^2_{cont,k} = T_k \Lambda^{-1/2} P_k^T$.

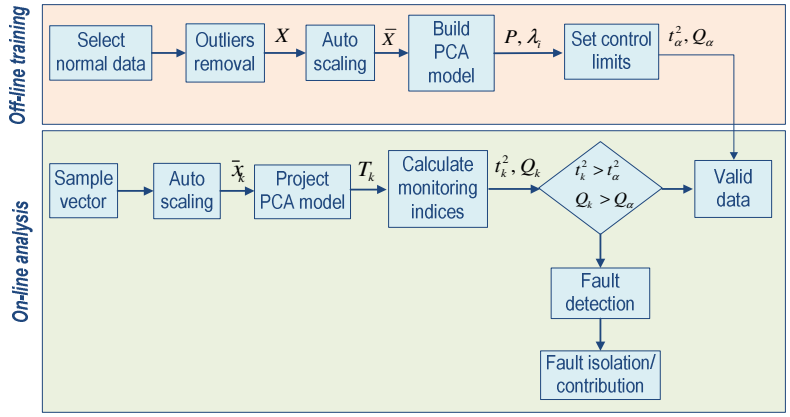

Figure 3. Scheme of setting up and running a PCA-based monitoring system.

Due to changes in operational conditions, mean, variance and correlation structure among variables could change in time. In that case possible false alarms could be generated. Currently, an adaptive approach is investigated to recursively determine the PCA model and the control limits that best fit the data under study by using a data monitoring window (Lee and Vanrolleghem, 2003). However, keeping the computational burden limited is the key point to consider for an on-line implementation.

3. RESULTS

To illustrate the potential of the proposed approach, the PCA method has been applied to different case studies with groups of on-line water quality variables. Figure 4 shows the results obtained for a set of eight on-line time series (three temperature signals coming from three different sensors, conductivity, turbidity, pH, ammonia and chloride) collected at the inlet of the Lynette wastewater treatment plant (Copenhagen, Denmark). Data has been recorded at 5 second intervals.

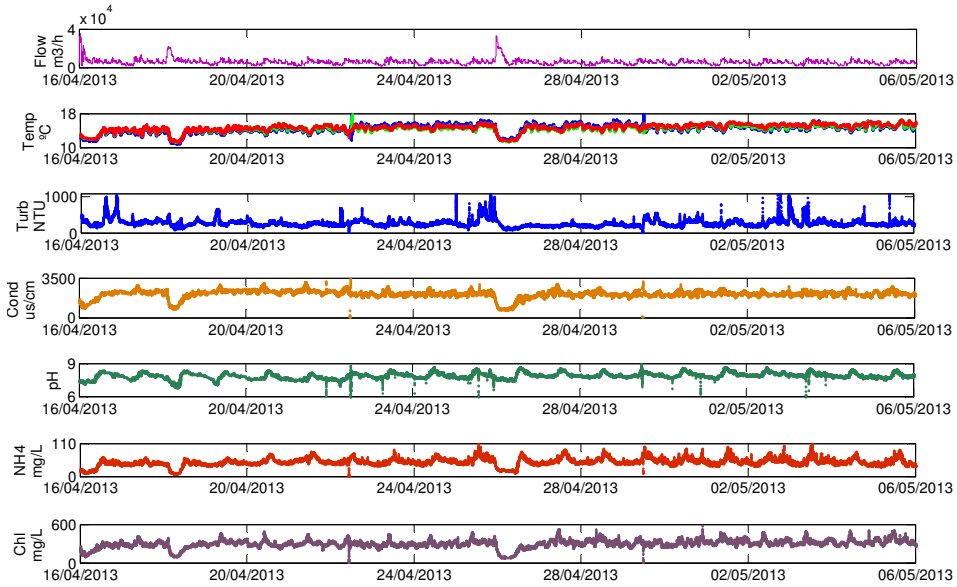

Figure 4. On-line water quality measurements at the Lynette WWTP (DK).

A 4-day data set collected under what could be considered normal operating conditions (period between April 18th and April 22nd in Figure 4), has been used for off-line training purposes, resulting in a PCA model where three principal components (PC1, PC2, PC3) capture around 90% of the variability in the data. Calculation of Q, and its respective control limit Q_α, has revealed that around 3% of the data was considered abnormal, proving the goodness of the model to describe the main sources of variability in the data.

Once the PCA model was obtained, it was applied to the whole time series for identification of faulty data. Figure 5 shows the scores (data transformed in the PCA model) for the two first principal components, for a 2-days data set (period between April 24nd and April 26th). While each point in the plot represents a data observation vector $\bar{x}_k$, the vectors represent each variable and its contribution to PC1 and PC2. A first visual analysis reveals that the three temperature signals collected by a conductivity (ConTemp), pH (pHTemp) and NH4 (AniseTemp) sensors behave similarly. A strong correlation is also noticed between the chloride (Chl) and conductivity measurements (Cond). This finding is confirmed by observing the time series (Figure 4) where rain events around April 18th and April 26th caused similar dynamics in both variables.

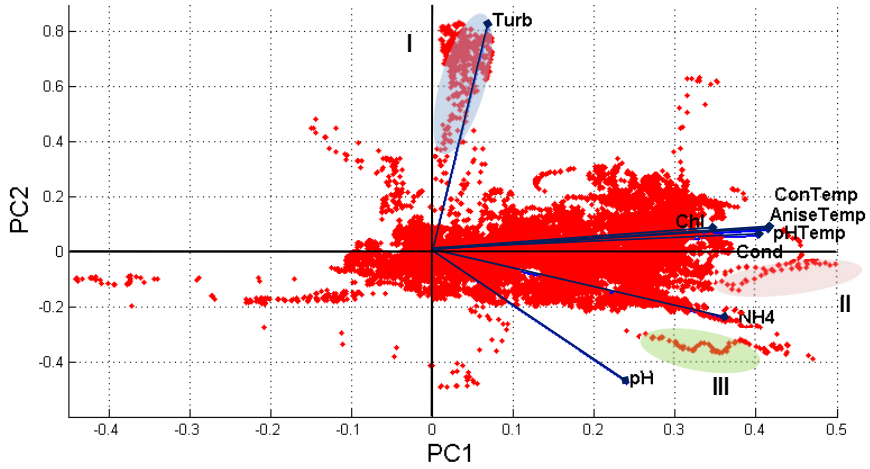

Figure 5. Scores of the PCA representation of the Lynette data.

Inspecting the scores representation in Figure 5 it can be noticed that most of the data cluster close to the origin but some deviations in the direction of the different variables' vectors are also present. For example, data clusters I, II and III suggest an abnormal behavior for these corresponding sensors. Monitoring of the t^2 and Q statistics allow for the detection and isolation of some fault situations in the process. Some examples will be discussed below.

For observations in cluster I in the direction of the turbidity measurements (Turb – Figure 5) the Q statistic detected unusual behavior that changes the normal correlation between the variables and the t^2 statistic detected abnormal variations within the model subspace (see Figure 6a). Calculation of the individual score contributions to t^2 for one sample in period I suggests that the turbidity measurements is the probable cause of the fault (variable 1, last subplot in Figure 6a). Moreover, the time series for turbidity measurements indeed revealed abnormal behavior for that variable.

Figure 6b shows an example of coinciding multiple sensors faults. Time series for period II revealed some abnormal behaviour for the conductivity measurements and the temperature data recorded with that sensor. Although t^2 remained inside the in-control region, the Q statistic revealed a new source of variance and a bad correspondence of these data to the PCA model. In this case, the individual contributions to the Q statistic for one sample in that period (variable 4 and variable 5, last subplot in Figure 6b) showed that the conductivity and conductivity temperature measurements were the variables responsible for these large Q values. For period III, both t^2 and Q statistics were outside the control limits suggesting (1) phenomena not taken into account in the model and (2) a higher than normal variability in the data. Time series for that period indeed exhibited abnormal behaviour for pH and pH temperature measurements (Figure 6b).

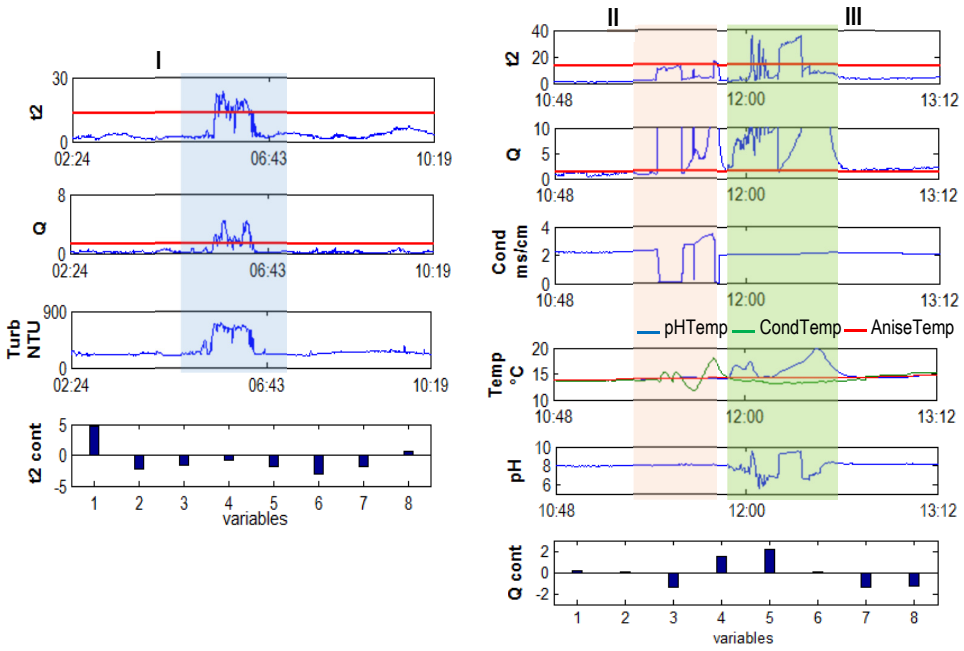

Figure 6. Faulty measurements on April 22nd. (a) Period I: From top to bottom, t^2 and Q statistic, Turbidity time series and t^2 contribution plot by the 8 sensors, (b) Period II and III: From top to bottom, t^2 and Q statistic, Conductivity, temperature and pH time series and Q contribution plot by the 8 sensors.

Summarizing, for each observation data quality was monitored by calculating the corresponding Q and t^2 statistics and their violation to the established control limits. Figure 7 shows the monitored statistics for the complete validation data set and their respective control limits (horizontal red lines). For the period under study Q and t^2 revealed respectively around 7% and 2% of the data to be abnormal or faulty. This is quite good compared to typical data rejection percentages between 5 and 50% (van Bijnen and Korving (2008): 40%; Thomann (2008): 5-15%; Métadier (2011): 40-60%; Schilperoort (2011): 25-50%).

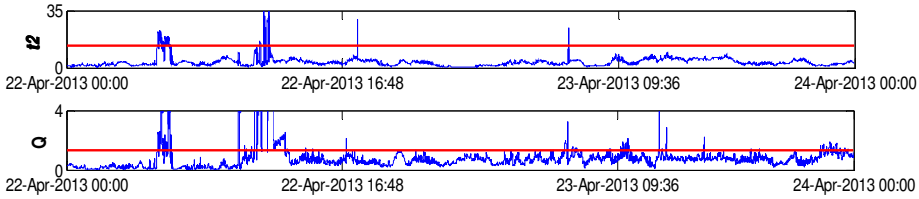

Figure 7. Q and t^2 statistics for the validation data set collected at Lynette (DK)

4. CONCLUSIONS

Dealing with on-line sensors to make water quality monitoring networks useful in practice still represents an important challenge. Data collected with in situ monitoring systems are not without errors due to the challenging measurement conditions that prevail in wastewater and other water system environments. In that sense, efficient monitoring will depend on careful data quality assessment. With this in mind, an automatic data quality evaluation tool for analysis of multivariate on-line time series, based on statistical process monitoring, has been presented and successfully validated on complex data sets collected at the inlet of a treatment plant. The method, based on PCA techniques and using the monitoring of some statistical metrics, was shown to be effective for detection and posterior isolation of different sensor faults in view of an on-line practical implementation.

6. ACKNOWLEDGMENTS

Peter Vanrolleghem holds the Canada Research Chair in Water Quality Modelling. The CFI Canada Research Chairs Infrastructure Fund project (202441) provided the monitoring stations. Peter Vanrolleghem was Otto Monsted Guest Professor at the Technical University of Denmark in 2012-2013.

7. REFERENCES

Alcala C. and Qin J. (2003) Unified analysis of diagnosis methods for process monitoring. In: 7th IFAC Symposium on Fault Detection, Supervision and Safety of Technical Processes, Barcelona, Spain.

Alferes J., Poirier P., Lamaire-Chad c., Sharma A.K., Mikkelsen P.S. and Vanrolleguem P.S. (2013a) Data quality assurance in monitoring of wastewater quality: Univariate on-line and off-line methods. In: Proceedings 11th IWA Conference on Instrumentation, Control and Automation (ICA2013). Narbonne, France, September 18-20 2013.

Alferes J., Tik S., Copp J. and Vanrolleghem P.A. (2013b) Advanced monitoring of water systems using in situ measurement stations: Data validation and fault detection. Wat. Sci. Tech., 68, 1022-1030.

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

Caradot N., Sonnenberg H., Riechel M., Heinzmann B., von Seggern D., Matzinger A. and Rouault P. (2011) Application of online water quality sensors for integrated CSO impact assessment in Berlin (Germany). 12th International Conference on Urban Drainage, Porto Alegre/Brazil, 11-16 September 2011

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

Métadier M. (2011) Traitement et analyse de séries chronologiques continues de turbidité pour la formulation et le test de modèles de rejets urbains par temps de pluie. PhD thesis, INSA-Lyon, France. pp. 419. (in French).

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

Schilperoort R. (2011) Monitoring as a tool for the assessment of wastewater quality dynamics. PhD thesis, Technical University of Delft, The Netherlands. pp. 320

Sonnenberg H, Rouault P. and Heinzmann B. (2010) Online monitoring for evaluation of CSO impact on surface water. Proceedings Modelling Monitoring Management Workshop - Application of Integrative Modelling and Monitoring Approaches for River Basin Management evaluation, Luxemburg, 2010.

Thomann M. (2008) Quality evaluation methods for wastewater treatment plant data. Wat. Sci. Technol, 57(10), 1601-1609.

van Bijnen M. and Korving H. (2008) Application and results of automatic validation of sewer monitoring data. In: Proceedings 11th International Conference on Urban Drainage (ICUD2008). Edinburgh, Scotland, UK.

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

Wagner R., Boulger R., Oblinger C. and Smith B. (2006) Guidelines and Standard Procedures for Continuous Water-Quality Monitors: Station Operation, Record Computation, and Data Reporting. U.S. Geological Survey, Reston, Virginia.

Yoo C.K., Villez K., Van Hulle S.W. and Vanrolleghem P.A. (2008) Enhanced process monitoring for wastewater treatment systems. Envirometrics, 19(6), 602-617.