Data quality assurance in monitoring of wastewater quality: Univariate on-line and off-line methods

J. Alferes[1], P. Poirier[1], C. Lamaire-Chad[1], A.K. Sharma[2], P.S. Mikkelsen[2], P.A. Vanrolleghem[1]

[1]model*EAU*, Université Laval, 1065, Avenue de la Médecine, Québec, Canada QC G1V 0A6
[2]DTU Environment, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark

Abstract: To make water quality monitoring networks useful for practice, the automation of data collection and data validation still represents an important challenge. Efficient monitoring depends on careful quality control and quality assessment. With a practical orientation a data quality assurance procedure is presented that combines univariate off-line and on-line methods to assess water quality sensors and to detect and replace doubtful data. While the off-line concept uses control charts for quality control, the on-line methods aim at outlier and fault detection by using autoregressive models. The proposed tools were successfully tested with data sets collected at the inlet of a primary clarifier, where probably the toughest measurement conditions are found in wastewater treatment plants.

Keywords: Data quality assessment; on-line wastewater monitoring; univariate methods

INTRODUCTION

Thanks to important technological developments regarding on-line water quality sensors, in situ monitoring stations are increasingly being used to identify and describe pollution dynamics in water bodies. Huge data sets consisting of a large number of physical-chemical parameters are then generated with those systems. Since sensors are still subject to functional, technical and operational constraints, and even more under the challenging measuring conditions that prevail in wastewater systems, they are disturbed by bias, drift, precision degradation or total failure effects that cause the reliability of measurements to decrease. Those situations can lead to faulty conclusions and to incorrect use of the data. Consequently, meaningful water quality data will intrinsically depend on the application of quality assessment and quality control practices to ensure that high quality data is being collected.

Different methods have been developed for data quality assessment in different fields, the main goal being the identification of out-of-control situations caused by systematic or gross errors (Thomann, 2008). However, there is still a long way to bring the many academic developments into practice in the water sector, nowadays most of the data assessment process is done by using inefficient and laborious manual procedures. The sheer size of the data sets to be dealt with makes the data assessment process crucial for an effective monitoring strategy. In this paper an automatic data quality assurance procedure with a practical orientation is presented. By combining off-line methods for data quality control and univariate on-line methods for data quality assessment, information from single variables is extracted to assess sensors measuring quality and to identify outliers, noise, and potential sensor faults. Once the individual signals are data quality controlled they can be used for multivariate analysis (Alferes et al., 2013). The developed algorithms are illustrated with automated monitoring systems installed at the inlet of a wastewater treatment plant.

MATERIALS AND METHODS

Case study

Two automated monitoring stations (RSM30, Primodal Systems, Canada) have been installed at the inlet and at the outlet of a primary clarifier line of the 700,000 PE municipal treatment plant Lynetten (Copenhagen, Denmark) to study the inflow dynamics and the performance of the primary clarifier. Both monitoring stations comprise sensors for conventional physical-chemical parameters (temperature, pH, turbidity, conductivity), a UV spectrometer (TSS, CODt, CODf) and ion selective electrodes for ammonia, potassium and chloride. Data were recorded at intervals of 5 to 60 seconds, generating information-rich data sets that can be used among others to provide a better understanding of the behaviour of the WWT during dry and wet weather conditions, for modelling and forecasting influent water quality and real-time control. The particularly hard measurement conditions at the WWTP's inlet represent an important practical challenge to achieve good quality of the on-line measurements.

Data quality assurance procedure

The two principal components of a quality assurance program include quality control and quality assessment. According to the mon*EAU* vision (Rieger and Vanrolleghem, 2008), data quality monitoring encompasses two different tasks: off-line and on-line analysis. While the off-line analysis uses comparative reference measurements to detect systematic errors and poor calibration, the on-line analysis of the time series allows the detection of deviations from a normal state. In the framework of the mon*EAU* vision the proposed quality assurance procedure is presented in **Figure 1**. Concerning the off-line analysis, on-line sensors are normally controlled with grab samples measured with a reference method (Thomann et al., 2002). Control charts are then built with appropriate out-of-control criteria to detect systematic or gross errors. In the presented procedure, to improve the maintenance routines the quality control with respect to fouling and calibration is done by building additional control charts which compare standard solution values with sensor measurements.

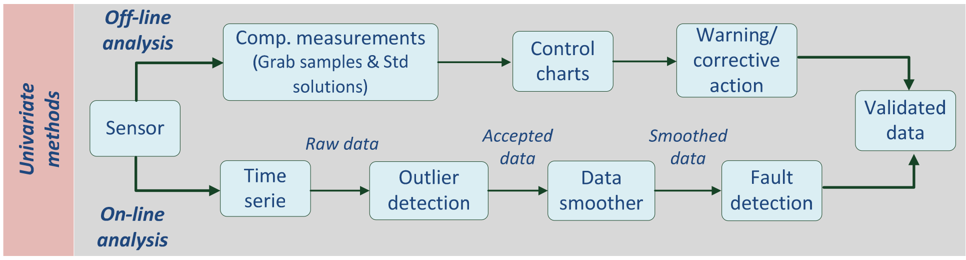

Figure 1 Univariate methods for data quality assurance of water quality data

Concerning the on-line analysis, the proposed tool for automatic data quality assessment is aimed at outlier detection and fault detection in consecutive steps (**Figure 1**). Based on forecasting of time series data by using autoregressive models, the unknown parameters in the autoregressive model are estimated and then the model is projected into the future to obtain a forecast. Outliers are identified by comparing the measured values with the calculated forecast values with their dynamic prediction error interval. Since it is required that the model be a good representation of the observations in any local segment of time close to the present, a trade-off between responsiveness and stability of the forecasting system is key in setting up the outlier detection algorithm. A third-order exponential smoothing model was chosen due to its simplicity, its computational efficiency, the ease of adjusting its responsiveness to changes in the process and its adequate accuracy (Taylor, 2010). At time *T*, the forecast value of the data *x* in the next time unit, *T+1* is calculated as follows:

$$\hat{x}_{T+1} = \hat{a}_T + \hat{b}_T + \frac{1}{2}\hat{c}_T \qquad (1)$$

where $\hat{a}, \hat{b}, \hat{c}$ are the coefficients of the model, computed using the first three exponentially smoothed statistics ($S_T, S_T^{[2]}, S_T^{[3]}$). The smoothed statistic S_T given by

$S_T = \alpha x_T + (1-\alpha)S_{T-1}$, constitutes a geometrically weighted average of past observations. The smoothing constant α determines the behaviour of the forecast system. Small values of α give more weight to the historical data promoting a slow response. With large α values more weight is placed on the current observation leading to faster response. Once the forecast value is obtained, outliers are identified by analysing the one-step-ahead forecast error $e_T(1)$ calculated as $e_T(1) = x_T - \hat{x}_T$. To provide better estimations of the local variance and to quantify the extent by which the actual value differs from the forecast, a simple exponential smoothing model is also chosen to estimate the variance of the forecast error σ_e^2 through the estimation of the mean absolute standard deviation, Δ. At time *T*, supposing that the forecast error is normally distributed, the estimate of σ_e^2 is obtained as $\hat{\sigma}_e = 1.25\hat{\Delta}_T$, with $\hat{\sigma}_e$ the estimate of the standard deviation and $\hat{\Delta}_T$ calculated as follows:

$$\hat{\Delta}_T = \alpha_{std}\left|e_T(1)\right| + (1-\alpha)\hat{\Delta}_{T-1} \qquad (2)$$

The goodness of the two models depends on the right choice of the smoothing constants (α, α_{std}) that minimize the residuals between the model and a representative set of calibration data. The prediction interval $xlim$ is defined based on a probability statement about the forecast error as $xlim_T = \hat{x}_T \pm K\hat{\sigma}_{e,T}$, with K being a proportional constant that can be adjusted to make the limits more or less restrictive. If x_T breaks the prediction interval it is considered an outlier and it is replaced by the forecast value. A new *accepted data* series is then generated. **Figure 2** gives a graphical representation.

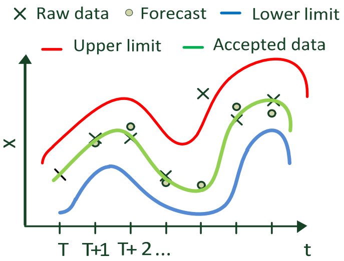

Figure 2 Outlier detection and method

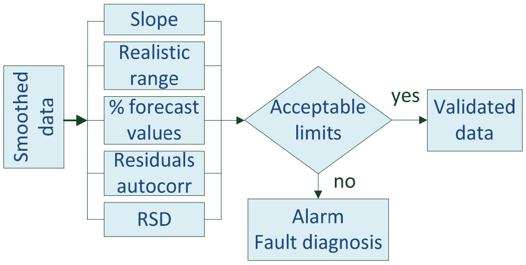

Figure 3 Fault detection method

Potential sensor faults are identified by calculating some data features and their acceptability limits, as shown in **Figure 3**. To avoid corruption of the calculations by signal noise, the features are computed over the *smoothed data* (**Figure 1**) that are obtained by using a kernel smoother. The calculated data features comprise the slope, the locally realistic range, the fraction forecast values that have replaced the raw data, the autocorrelation of the residuals and finally the residuals' standard deviation. After a fault is detected, an alarm is generated and posterior analysis is carried out to identify the fault and to apply the required corrective action in the field.

RESULTS AND DISCUSSION

The univariate methods were successfully applied to the on-line TSS and conductivity time series collected at the inlet of the primary clarifier. In **Figure 4**, even if most of the TSS data fall into the prediction interval (blue and red zones), a large number of outliers is identified. Once the outliers have been replaced by the corresponding forecast values and smoothing is applied (green line) the daily TSS-pattern becomes apparent. Notably the developed and tuned method maintains its outlier removing performance under the unusually wet weather conditions around December 24th and

27th 2012. In the last step of the on-line analysis, fault detection (Figure 1 and 3), the calculation of the different features on the smoothed data did not reveal significant faults. The smoothed TSS time series can thus be properly used for further analysis.

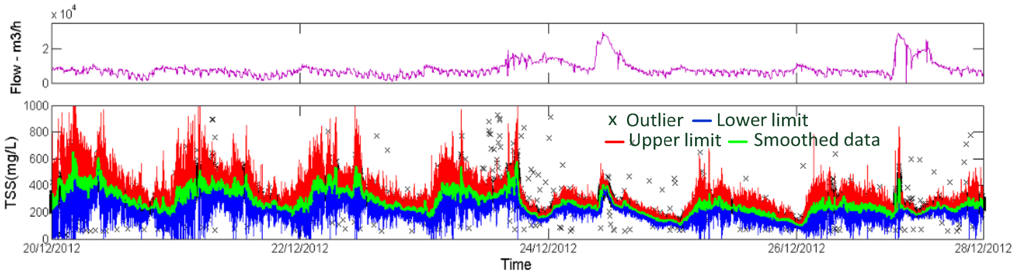

Figure 4 On-line outlier detection procedure applied to an actual TSS time series from Lynetten WWTP

For the conductivity on-line measurements in **Figure 5** some abnormal behaviour was detected. Acceptance limits for the data features are shown as red lines. Diagnosing the residuals (by carrying out a runs test on a moving window) allows checking whether the forecasting model is adequately describing the raw data. Most of the data pass this test. A snow-melt road salts event is clearly observed around December 15th 2012 when the conductivity values increased twofold from the normal values. Larger slope values for that period demonstrate the more important dynamics in the variable. Around December 17th the residuals' standard deviation (RSD) exceeds the typical measurement standard deviation. This coincides with a high percent of forecast values that replace the raw data and larger slope values suggesting an atypical variation of the data. This was due to the sensor being out of the water.

Due to space limitation no illustration is given of the off-line analysis but its application has been useful for detecting for example drift of the ammonia and chloride sensor and missed calibration steps in the pH sensor and UV spectrometer.

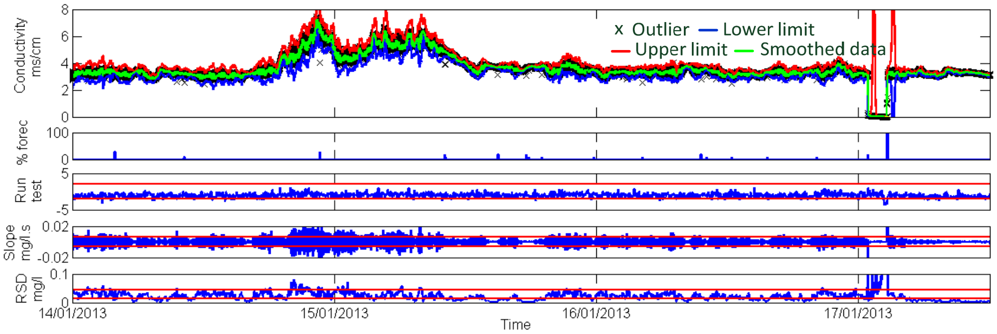

Figure 5 On-line outlier and fault detection procedure applied to an actual conductivity time series

REFERENCES

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

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

Taylor J. (2010) Triple seasonal methods for short-term electricity demand forecasting. *Eur. J. Operational Res.*, 204, 139-152.

Thomann, M., Rieger L., Frommhold S., Siegrist H. and Gujer W. (2002) An efficient monitoring concept with control charts for on-line sensors. *Water Sci. Tech.*, 46(4-5), 107-11.

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