Validating data quality for water quality monitoring: Objective comparison of three data quality assessment approaches

Janelcy Alferes[1], John Copp[2], Stefan Weijers[3] and Peter A. Vanrolleghem[1]

[1]model*EAU*, Université Laval, Département de génie civil et de génie des eaux, Québec, QC G1V 0A6, Canada (Email: *janelcy.alferes@gci.ulaval.ca*)
[2]Primodal Inc., Hamilton, ON L8S 3A4, Canada
[3]Waterschap De Dommel, Boxtel, the Netherlands, 5280 DA

Keywords
Data quality assessment, Fault detection, Process monitoring, Water quality

INTRODUCTION

Well operated continuous water quality measurement systems have the ability to capture dynamics in water and wastewater systems, which allows for the identification of critical events, the evaluation of discharge impacts on receiving water bodies and the identification of cause and effect relationships. For many activities water authorities and water utilities are increasing the number of continuous measurements at different locations within their water and wastewater systems. As a main challenge, water quality sensors are subjected to different kinds of faults, leading to bias, drift, precision degradation or total failure, all causing a deterioration of the reliability of measurements (Mourad and Bertrand-Krajewski, 2002). Common errors observed in the increasingly long raw data series include missing values, NaN values, outliers, measurement values out of range or being constant and varying noise.

To be able to effectively use the information that is in principle present in those continuous measurement data sets, the data should be accurate, verifiable, and properly validated. Given the size of the data sets produced nowadays in online water quality monitoring schemes, automated data validation is the only practicable option (Alferes et al., 2013). However, existing "standard" methods for data validation can only be implemented with difficulty in the water sector due to the properties of the water quality measurements and the water environments they are subjected to. As a result, common practice still today is visual inspection and the application of manual procedures.

In the framework of practical water quality monitoring applications three approaches for automatic data quality assessment developed by different research teams are presented and compared. The three approaches are each intended to detect doubtful and unreliable data by evaluating different features extracted from the data. However, taking the decision whether a value is to be classified as "valid" or "not valid" is not a simple task. Given the differences in methodologies and objectives it is also not evident to objectively evaluate the performance of different data validation approaches in a practical setting.

Thus, in this paper a methodology will be presented to evaluate and compare the usefulness or appropriateness of each approach based on data validation performance indices that are calculated from actually collected water quality time series. The final objective of this collaborative effort

among the three research teams is to come up with better tools and improved approaches for implementing successful automatic data quality validation procedures. All this will be illustrated with extensive data sets collected at different locations in water systems, including wastewater treatment plants and receiving waters.

COMPARISON OF DATA VALIDATION APPROACHES

Three teams have developed actual implementations that are currently being used in measurement systems with data quality validation schemes. In general, different fault types can be detected at different levels with those algorithms: (i) measurements outside physically acceptable limits, (ii) constant value or regular oscillations, (iii) constant difference between points, (iv) abnormal residual standard deviation, (v) difference between sequential points outside an acceptable tolerance, (vi) autocorrelation outside an acceptable tolerance, (vii) abnormal slope values and (viii) high percentage of outliers. Figure 1 illustrates the results of the application of one of the presented methods based on an analysis of the time series and the calculation of several data features (Alferes et al., 2013). In this case the data validation approach is applied to a conductivity time series collected at the inlet of the Eindhoven wastewater treatment plant in the Netherlands. Blue and red lines represent the prediction interval within which normal data should fall and the green line represents the smoothed and outlier-corrected data. Different kinds of faults are detected. For example around September 26th (period I) a high percentage of outliers is identified. Around September 27th (period II) the algorithm indicated excessive slope and residual standard deviation values (RSD). The algorithm also detected a period with constant value around September 28th. Once all data features have been assessed for each data point, data is validated according its degree of reliability (last subplot "Label"): 0 - valid (green color, all data quality tests passed), 1 – doubtful (orange color, some tests failed), and 2 – not valid (red color).

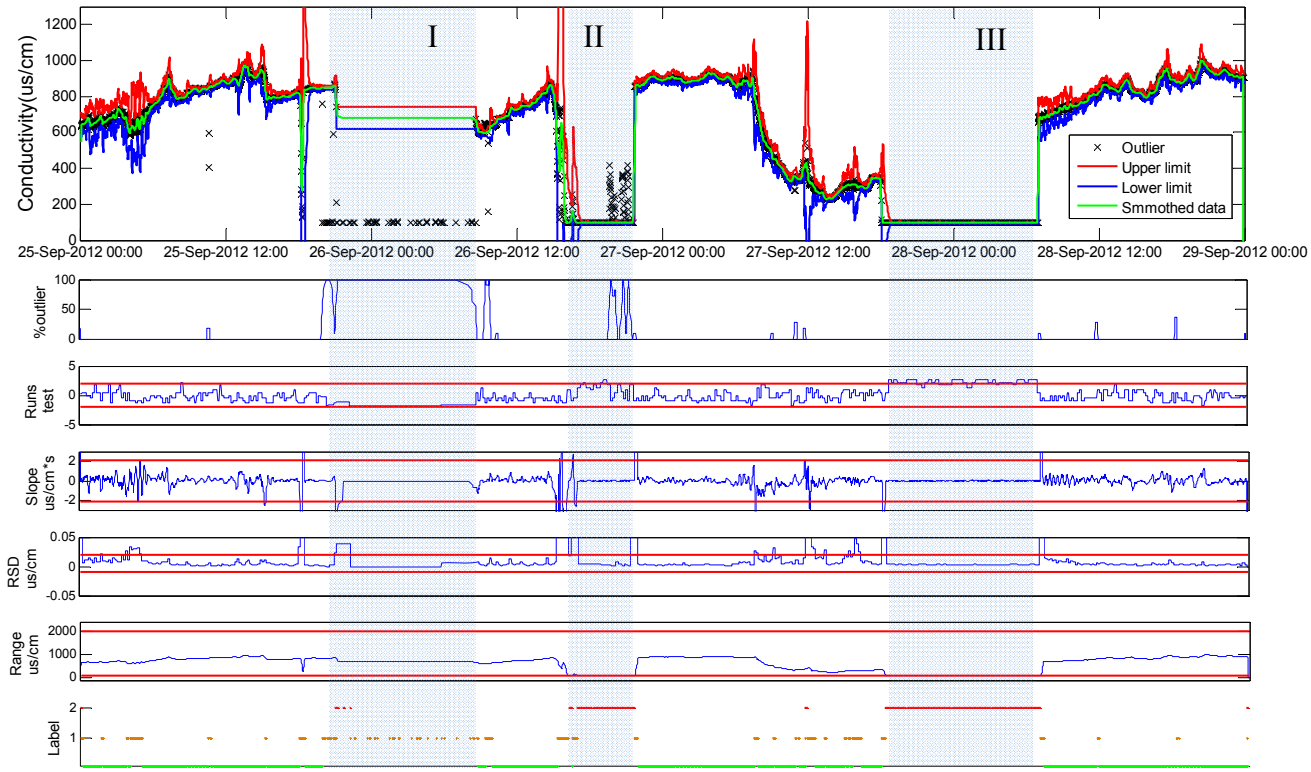

Figure 1. Application of a time series analysis method for data validation purposes

The development of successful on-line data quality validation schemes would benefit greatly from an unbiased comparison of different data validation methods, evaluating their performance in terms of identifying real sensor/process faults. Taking as basis the objective assessment procedure for process monitoring strategies within the benchmark simulation platform BSM1_LT (Corominas et al., 2010), a global monitoring index is proposed that allows comparing data validation approaches on the basis of real data sets. Working with real rather than simulated sensor data is the real challenge of the work. Indeed, in the work of Corominas et al. (2010) a set of given ('real') sensor faults are simulated and the fault sequence given by a validation method can be assessed in a simulation benchmarking context. Here, the assessment is to be done on actual data for which the faults are not known. To replace the real fault sequence, expert evaluation of the time series itself, supported by the results of the applied data validation methods, is used to create a surrogate fault sequence to which the different validation methods can then be compared. Figure 2 presents an example of the construction of the surrogate fault sequence for an online sensor. For each data validation method a value of "1" represents normal behaviour and a value of "2" represents a fault.

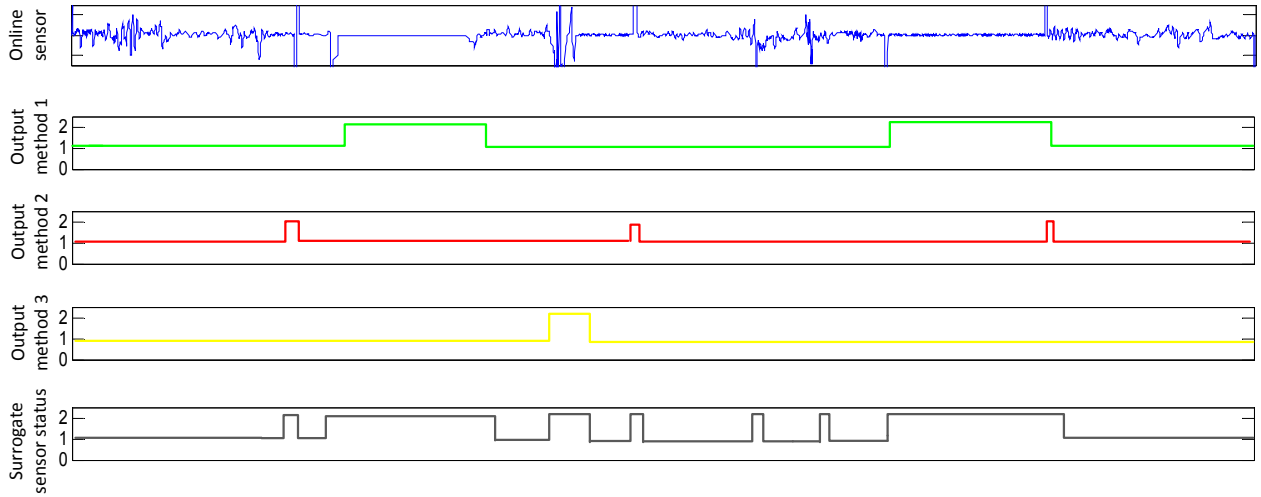

Figure 2. Example of the creation of the surrogate fault sequence for an online sensor

The global monitoring index is calculated in such a way as to assign penalization points for inadequate performance of a data validation approach (providing an estimated fault status of the sensor) against the true (surrogate) fault status of the real sensor. Each data validation method is evaluated over the whole time series giving as output the classification of every sample point to be a faulty or not. Then the global index penalizes for: (1) the indication of a false acceptance, (2) the indication of a false alarm and (3) the intermittent detection of a fault within the duration of a longer fault. Additionally, the objective comparison approach calculates individual indices to evaluate: (1) the average time required to detect a fault event (i.e. speed with which a fault is detected), (2) the number of detections of true faults, (3) the number of not detected faults and (4) the number of false detections.

Currently we are applying the comparison methodology to long time series considered relevant for the partners in their further applications (ammonia, flow and conductivity at the inlet of the WWTP; level measurements in the sewer system; and water velocity, level, dissolved oxygen and conductivity in the receiving water). The results will be discussed in the final version of the paper.

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. Water Sci. Technol., 68(5), 1022-1030. .

Corominas Ll., Villez K., Aguado D., Rieger L., Rosén C. and Vanrolleghem P.A. (2010) Performance evaluation of fault detection methods for wastewater treatment processes. Biotechnol. Bioeng., 108, 333-344.

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