Efficient automated quality assessment: Dealing with faulty on-line water quality sensors

Janelcy Alferes
modelEAU, Universite Laval, 1065, Avenue de la Medecine,
Quebec, Canada QC G1V 0A6
E-mail: janelcy@hotmail.com

Peter A. Vanrolleghem
modelEAU, Universite Laval, 1065, Avenue de la Medecine,
Quebec, Canada QC G1V 0A6
E-mail: peter.vanrolleghem@gci.ulaval.ca

Current environmental challenges for water resources include guaranteeing good ecological status of water bodies, promoting sustainable water use and protection of water resources. A key aspect in the achievement of these objectives is the application of a consistent and efficient monitoring strategy. Implementation of continuous water quality measurement systems is allowing to capture the dynamics in water systems for identification of critical events, cause-effect relationships and trends among others. Huge amounts of data are then being generated with uncertain quality. Water quality monitoring networks will only be useful in practice if careful quality assessment, of the data is carried out. With a practical vision, this paper presents a method for automatic data quality assessment extracting information from individual water quality time series from on-line sensors. Data mining techniques based on forecasting models are used to detect and remove unreliable data from the raw data sets. A posterior analysis is applied to remove noise and detect abnormal situations and potential sensor faults. The proposed tool has been successfully tested on water quality time series collected from different water and wastewater systems.

Keywords: Data quality assessment, forecasting techniques, on-line water systems monitoring, fault detection

Introduction

Within the different water legislations existing worldwide, a consistent monitoring strategy is becoming a key component for integrated water resources evaluation. The joint use of on-line monitoring stations and in-situ water quality sensors has become increasingly used in the environmental sector, allowing for the collection of high frequency data to identify and describe pollution dynamics in receiving water bodies, wastewater transport and treatment systems in view of taking remedial actions ([14], [12]). Huge data sets consisting of a large number of physical-chemical parameters are then being generated with those systems, data sets difficult to interpret without automated tools.

Additionally, even though important efforts have been carried out by the sensor manufacturers, due to the tough measurements conditions typically present in the water environments, measurements are still subject to many faults that can reduce the trustworthiness of the data ([13], [24], [15]). Sensors are disturbed by bias, drift, precision degradation or total failure effects that cause the reliability of measurements to decrease ([10], [6]).

The data reliability can suffer, leading to faulty conclusions and to incorrect use of the data. The data being collected will be only beneficial for its intended purpose, being water resources management among others, if it is available, accurate and verifiable in real or near-real time and correctly interpreted ([3], [1]). The previous statement and the sheer size of the data sets point to the importance of an automated systematic methodology for effective collection, management and validation of the data.

The core of an effective data quality assessment tool lays in the proper identification of unreliable data. 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 [20]. However, implementation of traditional data evaluation methods in the water sector is complicated by the properties of the water quality measurements (fast dynamics, nonrandom

AI Communications
ISSN 0921-7126, IOS Press.

noise, etc.) and this tends to make inefficient manual procedures common practice ([2], [23]).

In this paper a novel method, with a practical orientation, is presented that includes two main steps: identification and handling of doubtful values and fault detection. In the first step, forecasting techniques are applied to predict future expected time series data by using the historical information contained in on-line water quality time series. Unreliable data is then identified by comparing a new measured value with its forecasted value and the associated dynamic prediction acceptability interval. In the second step aimed at fault detection, several statistical data features and their acceptability limits are calculated. The proposed method has been successfully applied to assess the quality of water quality time series collected by monitoring stations at different water and wastewater systems. As soon as the data have been validated different applications for the validated data become possible, such as improving process knowledge, building models and in general to promote a more efficient monitoring of water systems. Furthermore, validated signals from on-line sensors can then be used with confidence as part of control strategies and to improve decision-making regarding water system management.

Materials and methods

The developed methodology, illustrated in the Figure 1, is based on statistical and forecasting methods and comprises two main steps: (1) pre-treatment of the data including detection of unreliable data and; (2) fault detection. The final objective is to come up with "quality validated" time series.

The first step is aimed to detect and remove doubtful or unreliable data in the form of outlying data. Outliers are defined as data that appear to be inconsistent or that deviate importantly from the rest of the measurements [11]. A univariate approach based on forecasting methods by using the historical data is implemented to create new "pre-treated" time series that can be used posteriorly for fault detection purposes. In the second step, some statistic data features are calculated over the new pre-treated time series. These data features are aimed to evaluate the goodness of both forecasting model and resulted smoothed data in relation to good historical behavior of the data.

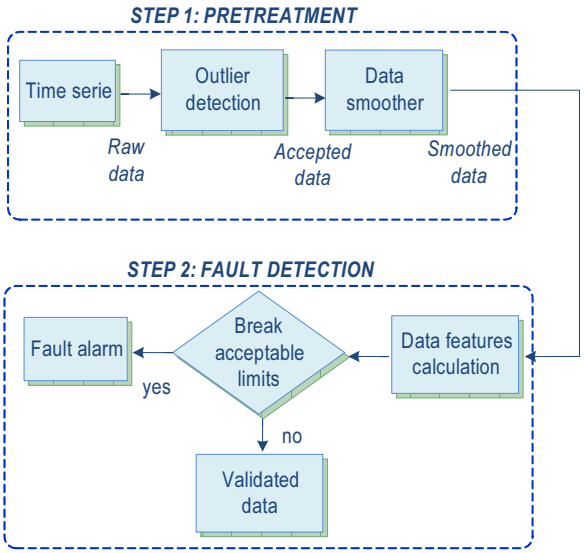

Fig. 1. Univariate time series analysis

Pretreatment: Forecasting of water quality time series for unreliable data removal

The fundamental concept underlying forecasting of time series is to examine past data and then estimate a likely future path for the series based upon the patterns observed in the historical data. The selection of the most suitable method for forecasting the time series depends on the historical behavior of the data. In general a predictive system is required to be robust, accurate, fast and simple to implement. Kalman and extended Kalman based algorithms have been widely used for tracking different electrical and vision systems [9]. Alternatively, exponential smoothing methods have been common in business and economic forecasting and have been successfully applied to time series without a significant trend to average out the irregular components [22]. These predictors have been shown to be simpler to analyze and implement in practical situations and have been used for modeling of time series hydrological data ([7], [8]).

Exponential smoothing models use exponentially decreasing weighted moving averages called "smoothing statistic" to calculate the forecast [21]. The first order statistic S_T can be obtained as $S_T = \alpha x_T + (1-\alpha)S_{T-1}$, where x_T represents the actual value of the data, S_{T-1} the estimated or forecast value for the present time period and α, with values between 0 and 1, a smoothing or weighting factor that controls how fast the weight of the historical data decays. The smoothing constant α, represents the forgetting factor.

It controls the rate of decay and determines the behavior of the forecasting system. Small values of α give more weight to the historical data leading to a slower response. Larger values of α give more weight to the current observation promoting a faster response.

In this way, a new estimate is calculated as the estimate for the present time period plus a fraction of random error. In general, single, double or third exponential smoothing models are used depending on the data characteristics (stationary, trend, seasonality) and coefficients of the model are obtained by using the first three exponentially smoothed statistics $(S_T, S_T^{[2]}, S_T^{[3]})$ that can be calculated as follows:

$$\begin{aligned} S_T &= \alpha x_T + (1-\alpha) S_{T-1} \\ S_T^{[2]} &= \alpha S_T + (1-\alpha) S_{T-1}^{[2]} \\ S_T^{[3]} &= \alpha S_T^{[2]} + (1-\alpha) S_{T-1}^{[3]} \end{aligned} \tag{1}$$

The main advantages of smoothing models include their short-term accuracy, simplicity and low computational cost [19]. However, the core of the forecast efficiency depends on the choice of the smoothing constant along with the process. This also assumes the presence of random error and a low level of autocorrelation. In the case of highly dependent observations, other forecasting techniques should be used.

Based on forecasting of water quality time series and exponential smoothing models an approach is proposed to deal with outliers and unreliable data. Dealing with outliers is a crucial part of any data quality assessment task as those abnormal points can affect any posterior statistical analysis and lead to incorrect conclusions about the behaviour of the system. Two autoregressive models are estimated for predicting the value of the data x and the value of the standard deviation Δ of the prediction error in the next time step, $T+1$. The unknown parameters in the autoregressive models are estimated and then the models are 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.

At time instant T, a third-order exponential smoothing model is used to forecast the value of the data at the next time instant $T+1$ as follows:

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

where $\hat{a}, \hat{b}, \hat{c}$ are the coefficients of the model, computed using the first three exponentially smoothed statistics previously described in (1), named $(S_T, S_T^{[2]}, S_T^{[3]})$. Once the statistics have been calculated, the coefficients of the model are obtained as follows:

$$\begin{aligned} \hat{a} =& 3S_T - 3S_T^{[2]} + S_T^{[3]} \\ \hat{b} =& \frac{\alpha}{2(\alpha-1)^2}[(6-5\alpha)S_T - 2(5-4\alpha)S_T^{[2]} \\ & + (4-3\alpha)S_T^{[3]}] \\ \hat{c} =& \Big(\frac{\alpha}{(\alpha-1)}\Big)^2 (S_T - 2S_T^{[2]} + S_T^{[3]}) \end{aligned} \tag{3}$$

To provide better estimations of the local variance and to quantify the extent by which the actual value differs from the forecast, a first-order exponential smoothing model is used to estimate the standard deviation $\hat{\Delta_T}$ of the forecast error σ_e^2 approximated as $\hat{\sigma}_e = 1.25\hat{\Delta_T}$, assuming a normal distribution for the forecast error [5]. The estimate of the standard deviation $\hat{\Delta_T}$ is calculated as follows:

$$\hat{\Delta_T} = \alpha |e_T(1)| + (1-\alpha)\hat{\Delta}_{T-1} \tag{4}$$

Being $e_T(1)$ the one-step-ahead forecast error calculated as $e_T(1) = x_T - \hat{x}_T$. Outliers are then identified by defining a prediction interval:

$$\begin{aligned} UpperLimit : xlim_T U &= \hat{x}_T + K\hat{\sigma}_{e,T} \\ LowerLimit : xlim_T L &= \hat{x}_T - K\hat{\sigma}_{e,T} \end{aligned} \tag{5}$$

This prediction interval represents the acceptability interval and enables the evaluation of the one-step-ahead forecast error. K represents a proportional constant that can be adjusted to make the limits more or less restrictive. If at time instant T the data falls outside the prediction interval it is considered an outlier and it is replaced by its forecast value. A new time series called accepted data is created where the outliers have been removed.

The core of the forecast and outlier removal efficiency depends on the choice of the smoothing constant α and the proportional constant K along with the nature and dynamics of the variable being estimated [17]. To choose the smoothing constant α, the model is calibrated using a so-called "good data" time series that represents modes of normal behavior (expected variations and system in-control). The root-mean squared error (RMSE) between the observed values and the forecast values within a grid search is used as criteria for selecting an appropriate smoothing con-

stant. K is adjusted according to the "good data" subset in the calibration phase to avoid the risk of rejecting real extreme dynamics as outliers and it may be specific for each time series.

For data assessment purposes the resulting accepted data is subsequently passed through a kernel smoother using the Nadarya-Watson kernel estimator as a locally weighted average [16] to remove noise. This phase allows reducing the corruption of subsequent statistical calculations. At each point x_0 the estimation of the smooth value by using the neighboring points is calculated as follows:

$$\hat{y}_h(x_0) = \sum_{i=1}^{n} W(x_0, x_i; h) y(x_i) \qquad (6)$$

where $\hat{y}_h$ is the estimation of the smooth value of the observed point at x_0 , n is the number of observed points, W is a weighting function, $y(xi)$ are the observations at x_i points and h the bandwidth, is the number of neighboring points around x_0 that are considered to estimate the smooth value. A Gaussian kernel function was implemented for calculation of the weighting function [18] . The bandwidth h controls the smoothness or roughness of the estimates and leads to under-smoothing or over-smoothing situations. For very small values of h the variance of the estimates will be too large while for very large values of h the bias of the estimates will be too large. The sum of squared residuals between the raw data and the smoothed data within a grid search is used as criteria to choose the proper value of h.

A new time series called smoothed data is then created. The whole approach for the pretreatment step is summarized in the Figure 2.

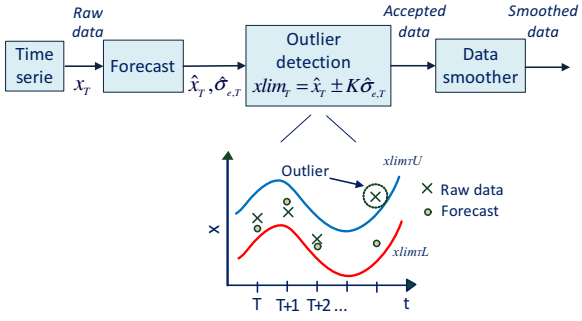

Fig. 2. Outlier detection method

Fault detection: analysis of data features

In the second step, some statistic data features and their confidence limits are calculated over the new treated time series at each time instant T. These data features are aimed to detect potential sensors faults by evaluating whether one, several or all data features break the acceptability limits. According to that, each data point is classified as valid, not valid or doubtful (meaning that posterior analysis must be carried out). Acceptability limits are defined for each data feature according with representative historical "good data".

Data features include the percentage of outliers or data replaced, the rate of change or slope between two data points, the local physical range [max min] expected in a specific location, the residual standard deviation (RSD) and finally the autocorrelation of the residuals. Figure 3 resumes the method proposed. 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. Validated data can then be used to analyze the involved processes and for further actions like understanding of the behaviour of the water system, for modelling and real-time control purposes.

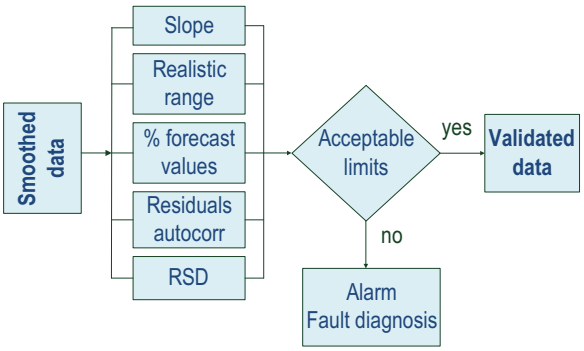

Fig. 3. Data features used for fault detection purposes

Results

The proposed approach has been implemented as part of Primodal Systems RSM30 PrecisionNow software [4] in different water systems including sewers, wastewater treatment plants (WWTP) and river bodies among others. The RSM30 monitoring stations were used to automatically collect in situ real-time water quality data at different locations and high temporal resolution (intervals of 5 to 60 seconds), generating

large and complex information-rich data sets. A systematic calibration and maintenance routine was periodically carried out to achieve the best data quality of the on-line measurements.

The proper tuning of the algorithm for each specific sensor and location is an important factor for the performance of the method. As example, the Figure 4 shows the results of applying the outlier detection method over a short period for conductivity measurements collected at the outlet of a primary clarifier at the Eindhoven wastewater treatment plant (the Netherlands). The dynamic calculation of the prediction limits (blue and red lines) allows for the detection of some outliers around July 20th. Sinusoidal noise due to a malfunctioning of the sensor is observed in the resultant smoothed data (green line Figure 4a). The algorithm can be properly adjusted to remove this special type of noise as illustrated in Figure 4b. Figure 5 shows the filtered data after applying the outlier detection method over a longer period of total suspended solids (TSS) measurements collected in the River Dommel (Eindhoven, The Netherlands). Red points represent the laboratory results from grab samples. In Figure 5a most of the filtered data agree with the laboratory data, although higher divergences are found from July 10th to 18th (shaded section). The results from a less restrictive tuning of the algorithms is shown in Figure 5b and leads to a better fit between on-line and laboratory measurements.

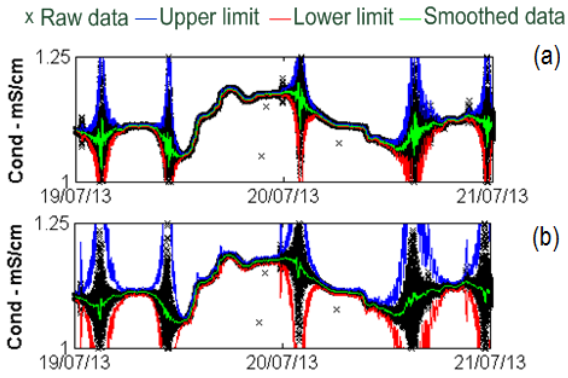

Fig. 4. Effect on noise removal of algorithm adjustments

Next figures show the results of the application of the whole method for different water quality time series from on-line sensors installed at different locations. In all the figures the first subplot represents the results for the first step, "pre-treatment" of the data. Blue and red lines represent the prediction interval

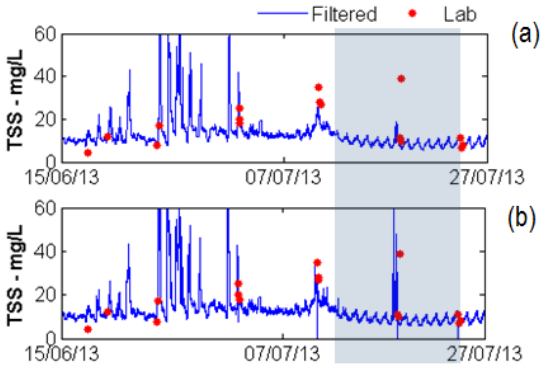

Fig. 5. Effect on agreement with lab measurements of algorithm adjustments

within which normal data should fall and the green line represents the smoothed and outlier-corrected data. The consecutive subplots illustrate the results for the "fault detection" step. The horizontal red lines represent the acceptability limits for each data feature.

The last subplot "Label" summarizes the data assessment process. Once all data features have been assessed for each data point, data is validated according its degree of reliability: 0 - valid (green color, all data quality tests passed), 1 doubtful (orange color, some tests failed), and 2 not valid (red color).

Figure 6 shows the results for a long period of a conductivity time series collected at the inlet of the Eindhoven WWTP, Netherlands. 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 according its degree of reliability (last subplot "Label"), about 65% of the data is considered as valid, comparing well with typical loss rates in such scenarios ([20]. Figure 7 shows the application of the overall method for a Turbidity time series collected at the River Dommel (Eindhoven, The Netherlands). Most of the data fall into the in-control region limited by the dynamic blue and red lines. However, some outliers are identified as indicated by the crosses in the top subplot and by the percentage of data that has been replaced by their forecast value (second subplot). Some abnormal behaviour is also detected by the rest of data features (subplots 3 to 6) and their acceptability limits (red horizontal lines). These limits have been defined by studying a "normal" measurement period.

Fig. 6. Application of the data quality assessment methodology for a 3 days conductivity time series (top subplot), collected at the inlet of the Eindhoven WWTP

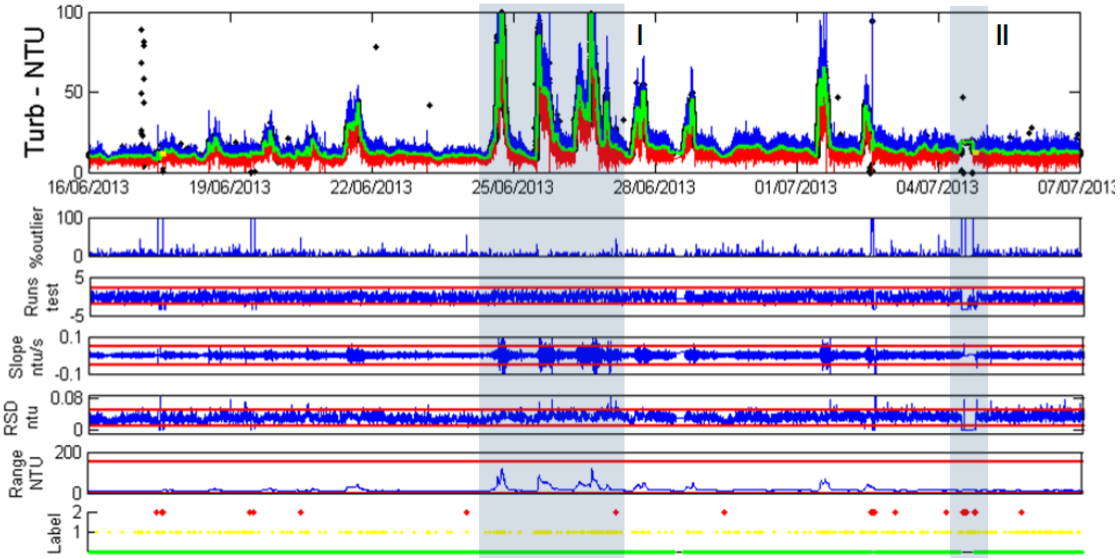

Fig. 7. Application of the proposed method for a three week Turbidity time series (top subplot) collected in the river Dommel (Eindhoven, The Netherlands)

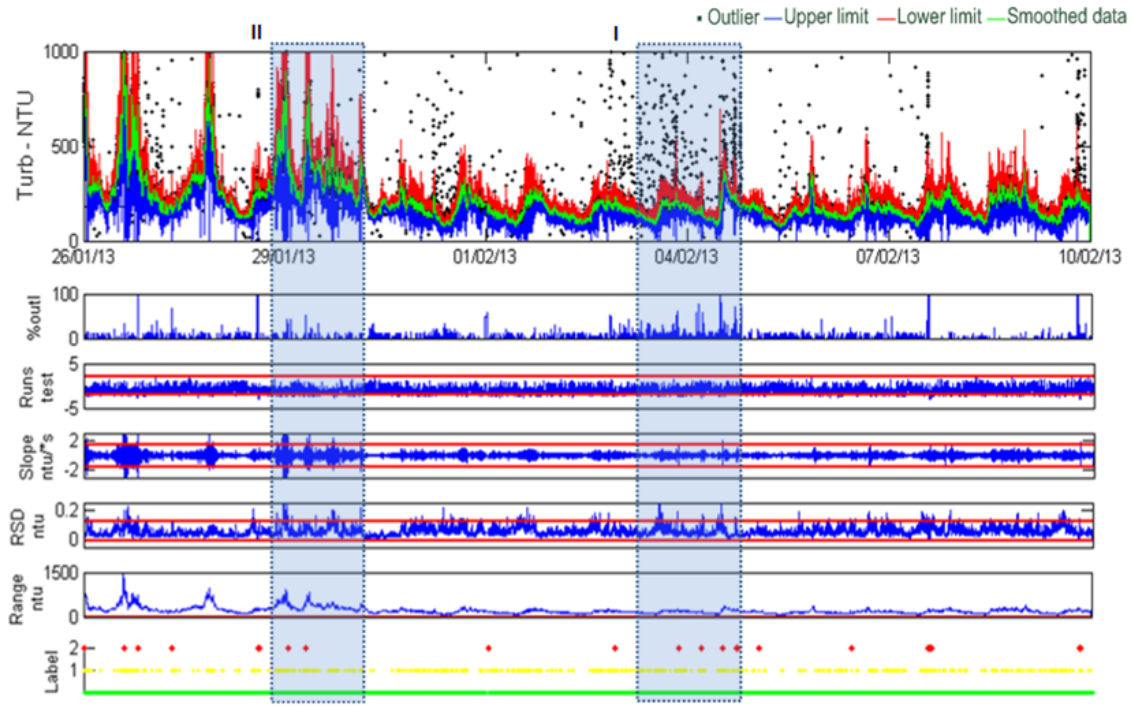

Fig. 8. Application of the proposed method for a three week Turbidity time series (top subplot) collected at the inlet of the Lynette WWTP (Denmark)

For example, in period I between June 25th and June 27th unusual variations in the Turbidity measurements were detected by the slope and RSD values that were higher than normally expected. On the other hand, in period II the runs test (a diagnostic test of the residuals over a moving window (Dochai) also indicated that the forecasting model was not able to describe the data, coinciding with a high percent of outliers and an unusually low variability in the Turbidity data. After evaluation of all the data features the final data quality outcomes are shown in the last subplot (valid, doubtful or not valid data). For the whole period, about 95% of the data was considered valid.

Figure 8 shows the results of applying the method to a turbidity time series collected at the raw sewage of the Lynette Municipal Treatment Plant in Copenhagen, Denmark. The inlet of WWTPs represents an important challenge being the hardest measurement location, especially under wet weather conditions. Special attention must be taken concerning assessing the measurements quality for an effective monitoring. Despite the effort for maintenance, the hard measurement conditions at the measuring sites (raw sewage) caused that data was affected by different kind of faults such as abundance of noise situations.

Some abnormal behaviour is detected for example in period I with RSD values as those expected in normal operating conditions, which account for a high level of noise and variance in the data. That coincides with a high percentage of outliers that have been replaced by their forecast value. In period II abnormal variation in turbidity measurements is detected by both RSD and slope values indicated more important dynamics in the variable that increased by more than three times the normal values. After evaluating all the group of data features and their violation to defined acceptability limits a label is given to each data point (Figure 8, last subplot). For the whole period, data that has passed all the tests and classified as valid represent about the 90% of the data. Even if the raw time series was affected by important noise levels, the application of the method resulted in a clean and validated time series (green line in first subplot) that puts in evidence the dynamics and profile of the turbidity measurements.

Conclusions

New challenges in the management of water systems are requiring a consistent water quality monitor-

ing strategy. Need has also risen to move from discrete measurements to integrated water resources evaluation implying the management and analysis of huge data sets. To fulfill this purpose in practical applications a comprehensive assessment technology for the data being collected needs to be applied to deal with one of the main issues: automated collection of reliable data.

With a practical orientation a method for automatic assessment of water quality time series has been presented. In contrast to traditional laborious manual data evaluation procedures, by using forecasting techniques the method is able in the first place to detect and remove outliers, noise and doubtful data from the raw data and replace these doubtful data with properly calculated values. In the second place, the evaluation of several statistical data features over the clean resultant time series allows the detection of abnormal behavior and potential sensor faults that lead to alarms for proper in-situ remedial actions.

However, the decision about when a value can be considered as valid or not valid is not simple. Several criteria as variable characteristics, type of sensor, final use of the data (modeling, decision making, control strategies, etc.) should be considered. The key for the successful application of the data quality evaluation process lies in the proper tuning of the method and acceptability limits for each specific application. Setting up the method is a crucial task that must be carried out carefully and for each specific sensor and location since each time series has a different dynamic. Expert knowledge about expected data variability and sources of faulty situations should be combined to set the methods parameters for each application. More complex and adaptive models will require estimating more parameters, and generally, it will imply a more computationally intensive calculation, limiting the implementation in on-line practical approaches. Methods considering cross validation through several signals will provide additional information to discern between a change in the sensors properties or in the process variable itself.

Acknowledgements

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. The authors wish to thanks Stefan Weijers, PhD and the collaborators of Waterschap De Dommel for their technical support. Peter Vanrolleghem was Otto Mnsted Guest Professor at the Technical University of Denmark in 2012-2013. Part of this research was co-financed by the Danish Council for Strategic Research (Storm and Wastewater Informatics project, SWI).

References

[1] J. Alferes and P. A. Vanrolleghem. Automated data quality assessment: Dealing with faulty on-line water quality sensors. In *Proceedings 7th International Congress on Environmental Modelling and Software (iEMSs2014)*, San Diego, USA, 2014.

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

[3] L. Clement, O. Thas, J. Ottoy, and P. A. Vanrolleghem. Data management of river water quality data: a semi-automatic procedure for data validation. *Water Resources Research*, 43(8), 2007.

[4] J. Copp, E. Belia, C. Hbner, M. Thron, P. A. Vanrolleghem, and L. Rieger. Towards the automation of water quality monitoring networks. In *Proc. 6th IEEE Conference on Automation Science and Engineering (CASE 2010)*, Toronto, Canada, 2010.

[5] D. Dochain and P. A. Vanrolleghem. *Dynamical Modelling and Estimation in Wastewater Treatment Processes*. IWA Publishing, London, UK, 2001.

[6] D. Garcia, J. Quebedo, V. Puig, and M. A. Cuguero. Sensor data validation and reconstruction in water networks: A methodology and software implementation. In *9th International Conference CRITIS*, Limassol, Cyprus, 2014.

[7] F. Garcia, D. Pedregal, and C. Roberts. Time series methods applied to failure prediction and detection. *Reliability Engineering System Safety*, 95(6):698–703, 2010.

[8] D. J. Hill and B. S. Minsker. Anomaly detection in streaming environmental sensor data: A data-driven modeling approach. *Environmental Modelling Software*, 25.

[9] A. Kiruluta, E. Eizenman, and S. Pasupathy. Predictive head movement tracking using a Kalman filter. *Systems, Man, and Cybernetics, IEEE Transactions*, 27(2):326–331, 1997.

[10] J. L.Bertrand-Krajewski and M. Muste. *Data validation: principles and implementation. Data requirements for Integrated Urban Management*. Unesco and Taylor Francis publishing, Paris, France, 2008.

[11] O. Maimon and L. Rockach. *Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers*. Springer, London, UK, 2010.

[12] M. Metadier and J. L. Bertrand-Krajewski. The use of long-term on-line turbidity measurements for the calculation of urban stormwater pollutant concentrations, loads, pollutographs and intra-event fluxes. *Wat. Res.*, 46.

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

[14] L. Rieger and P. A. Vanrolleghem. monEAU: a platform for water quality monitoring networks. *Water Sci. Technol.*, 57(7):1079–1086, 2008.

[15] S. Sandoval and J. L. Bertrand-Krajewski. Identification of errors in high temporal resolution rainfall time series by model based approaches. In *Proceedings of the 10th UDM - International Conference on Urban Drainage Modelling*, Mont Sainte Anne, Quebec, Canada, 2015.

[16] M. G. Schimek. *Smoothing and Regression, Approaches, Computation and Application*. John Wiley Sons, New York, USA, 2013.

[17] A. Sharma, L. Golubchik, and R. Govindan. Self-organizing systems. In *7th IFIPTC International Workshop IWSOS*, Palma de Mallorca, Spain, May 2013.

[18] T. Takahama and S. Sakai. A comparative study on kernel smoothers in differential evolution with estimated comparison method for reducing function evaluations. In *Proceedings of the Eleventh conference on Congress on Evolutionary Computation*, Trondheim, Norway, May 209.

[19] J. Taylor. Triple Seasonal Methods for Short-Term Electricity Demand Forecasting. *European Journal of Operational Research*, 204:139–152, 2010.

[20] M. Thomann. Quality evaluation methods for wastewater treatment plant data. *Water Sci. Technol*, 57:1601–1609, 10.

[21] B. F. Thomas. *Exponential Smoothing Models. Lecture. Department of Economics Southern Methodist University Dallas*. 2008.

[22] J. L. Viola. Double exponential smoothing: an alternative to kalman filter-based predictive tracking. In *Proc. Workshop on Virtual Environments*, Zurich, Switzerland, May 2003.

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

[24] C. K. Yoo, K. Villez, S. W. V. Hulle, and P. A. Vanrolleghem. Enhanced process monitoring for wastewater treatment systems. *Envirometrics*, 19:602–617, 6.