8-1-2014

Forecasting Techniques Applied To Water Quality Time Series In View Of Data Quality Assessment

Janelcy Alferes

John Copp

Peter A. Vanrolleghem

11th International Conference on Hydroinformatics
HIC 2014, New York City, USA

FORECASTING TECHNIQUES APPLIED TO WATER QUALITY TIME SERIES IN VIEW OF WATER QUALITY ASSESSMENT

JANELCY ALFERES (1), JOHN B. COPP (2), PETER A. VANROLLEGHEM (1)
(1): modelEAU, Université Laval, Québec, QC G1V 0A6, Canada
(2): Primodal Inc., Hamilton, ON L8S 3A4, Canada

The main advantage of continuous water quality measurement systems is the ability to capture dynamics in water and wastewater systems, which allows for the identification of critical events, the evaluation of impacts on receiving water bodies, the identification of cause and effect relationships and the ability to discern trends. However, the challenge associated with automatic monitoring systems is the collection of data with sufficient quality for the intended application. That is, useful monitoring is dependent on cautious data quality assessment. With particular attention to its practical implementation, this paper presents a method for data quality assessment that attempts to extract useful information from individual water quality measurement time series. Based on forecasting techniques that make use of the historical behavior of the data, raw measurements are evaluated for the detection of doubtful data and outliers. Posterior treatment is then applied to remove noise and detect potential sensor faults. The proposed tool has been successfully tested on water quality time series collected from different water and wastewater systems.

INTRODUCTION

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, and wastewater transport and treatment systems. Massive data sets are being generated. Even though important efforts have been carried out by the sensors 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 (Mourad and Bertrand-Kralewski, 2002). The data being collected is only beneficial for its intended purpose if it is available, accurate and verifiable in real or near-real time. The previous statement points 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. Implementation of traditional data evaluation methods in the water sector is complicated by the properties of the water quality measurements (fast dynamics, non-random noise, etc.) and this tends to make inefficient manual procedures common practice (Wagner, 2006). In this paper a novel method, with a practical orientation, is presented that includes three

main steps: identification of doubtful data, handling of doubtful values and fault detection. Using historical information contained in the water quality time series, forecasting techniques are applied to predict future expected time series data. Unreliable data is then identified by comparing a new measured value with its forecasted value that comes with a dynamic prediction acceptability interval. For fault detection, several statistical data features are calculated. The proposed method has been successfully applied to assess the quality of water quality time series collected by monitoring stations at different real-life water and wastewater systems. Validated data have then been used to improve process knowledge, build models and improve decision-making regarding water system management.

MATERIALS AND METHODS

Forecasting of water quality time series

Time series are defined as a sequence of observations of a particular variable at regular time intervals. 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 (Kiruluta, 1997). 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 (La Viola, 2003). These predictors have been shown to be simpler to analyze and implement in practical situations.

Exponential smoothing models use exponentially declining weighted moving averages called “smoothing statistics” to calculate the forecast. The first order static 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. 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:

$$S_T = \alpha x_T + (1-\alpha) S_{T-1}; \quad S_T^{[2]} = \alpha S_T + (1-\alpha) S_{T-1}^{[2]}; \quad S_T^{[3]} = \alpha S_T^{[2]} + (1-\alpha) S_{T-1}^{[3]} \tag{1}$$

The main advantages of smoothing models include their short-term accuracy, simplicity and low computational cost (Taylor, 2010). 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.

Water quality assessment

Based on forecasting of water quality time series data by exponentially smoothed models, a tool is proposed to first deal with outliers and then identify faults or abnormal values in the 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. At time instant T, while a third-order exponential smoothing model is used to forecast the value of the data at the next time instant T+1, 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. Outliers are then identified by defining a prediction interval $x_{lim_T} = \hat{x}_T \pm K\hat{\sigma}_{e,T}$ that enables the evaluation of the one-step-ahead forecast error. *K* represents a proportional constant. 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. For data assessment purposes the resultant data is subsequently passed through a kernel smoother to remove noise. Then data features and their confidence limits are calculated over the smoothed data at each time instant T. 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 1 resumes the method proposed.

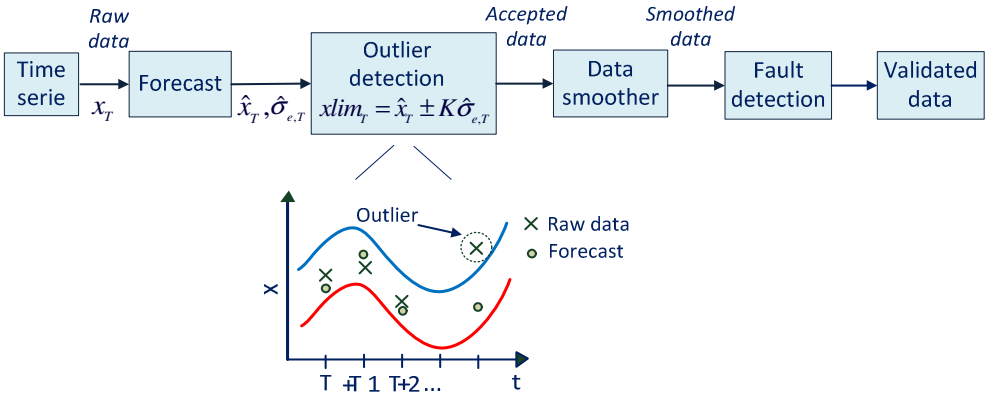

Figure 1. Time series analysis by using forecasting techniques

RESULTS

The proposed approach has been implemented as part of Primodal Systems' RSM30 PrecisionNow software (Copp et al., 2010). The RSM30 monitoring stations were used to automatically collect *in situ* real-time water quality data at different locations and high temporal resolution. A systematic calibration and maintenance routine was periodically carried out to achieve the best data quality of the on-line measurements.

Critical for the performance of the method is the proper tuning of the algorithm for each specific sensor and location. Figure 2 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 is observed in the resultant smoothed data (green line – Figure 2a). The algorithm can be properly adjusted to remove this special type of noise as illustrated in Figure 2b. Figure 3 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 3a 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 3b and leads to a better fit between on-line and laboratory measurements.

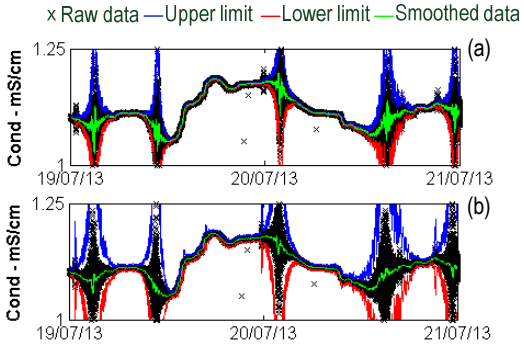

Figure 2. Effect on noise removal of algorithm adjustments

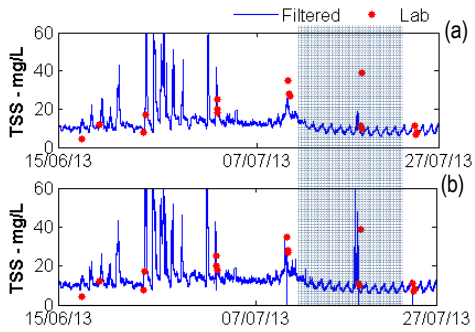

Figure 3. Effect on agreement with lab measurements of algorithm adjustments

Figure 4 shows the application of the overall method for a Turbidity time series collected at the River Dommel. 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.

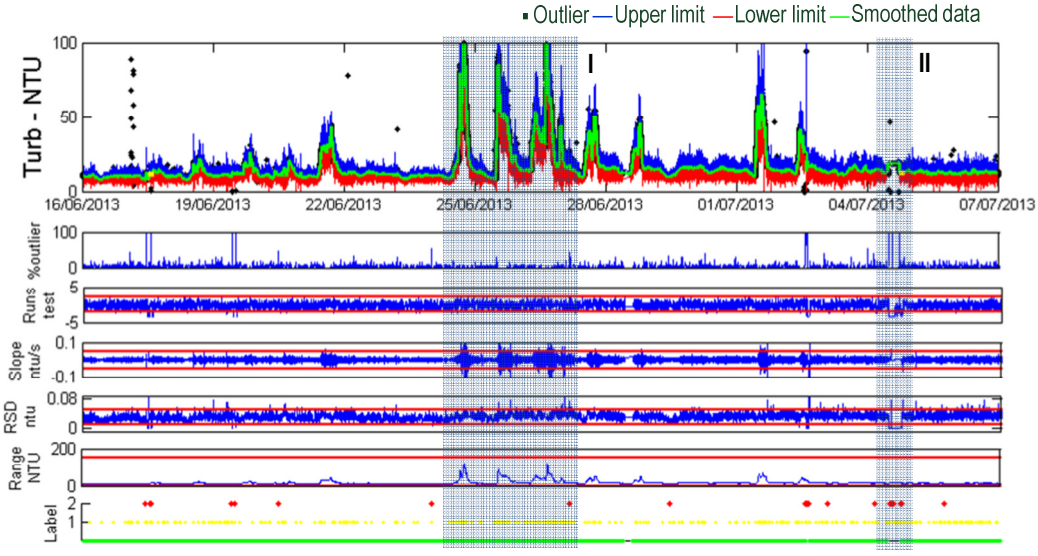

Figure 4. Application of the proposed method for a three week Turbidity time series (top subplot) collected in the river Dommel (Eindhoven, The Netherlands). Data quality assessment results for the shaded zones (other subplots) are discussed in the text

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 (Dochain and Vanrolleghem, 2001)) 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. Once all the data features are evaluated for each data point, the data quality is identified with a certain mark (0 – valid, 1 – doubtful, 2 – not valid), depending on its degree of trustworthiness. Final data quality outcomes are shown in the last subplot. For the whole period, about 95% of the data was considered valid.

CONCLUSIONS

In contrast to traditional laborious manual data evaluation procedures, a method for automatic water quality assessment of water quality time series has been presented. By using forecasting techniques, the method detects and removes outliers and noise from the raw data. A posterior analysis based on the evaluation of several data features over the filtered resultant time series is then carried out for abnormal behavior detection purposes. The method has been successfully applied to different water quality time series collected from different water systems. 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.

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. The authors wish to thanks Stefan Weijers, PhD and the collaborators of Waterschap De Dommel for their technical support during the measurement campaign of 2013.

REFERENCES

[1] Copp J., Belia E., Hübner C., Thron M., Vanrolleghem P.A. and Rieger L. (2010) Towards the automation of water quality monitoring networks. In: Proc. 6th IEEE Conference on Automation Science and Engineering (CASE 2010). Toronto, Canada, August 21-24, 2010.

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

[3] La Viola J. (2003) Double exponential smoothing: an alternative to kalman filter-based predictive tracking. In: Proc. Workshop on Virtual Environments. Zurich, Switzerland, May 22-23.

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

[5] Mourad M. and Bertrand-Kralewski 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.

[6] Taylor J. (2010) Triple Seasonal Methods for Short-Term Electricity Demand Forecasting. European Journal of Operational Research, 204, 139-152.

[7] 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.