Gap-filling of dry weather flow rate and water quality measurements in urban catchments by a time series modelling approach

Comblement de lacunes de mesures de débit et de qualité de l'eau de temps sec dans des bassins versants urbains par modélisation de séries chronologiques

Santiago Sandoval*, Luca Vezzaro**, Jean-Luc Bertrand-Krajewski*

*Université de Lyon, INSA Lyon, DEEP, 34 avenue des Arts, F-69621 Villeurbanne cedex, France
** DTU ENVIRONMENT, Department of Environmental Engineering, Technical University of Denmark, Building 115, 2800 Kgs. Lyngby, Denmark.

RÉSUMÉ

Les séries chronologiques de débit et de qualité des eaux par temps sec dans les systèmes d'assainissement unitaires peuvent contenir une quantité importante de données manquantes, ceci pour de multiples raisons, telles que les défaillances de fonctionnement des capteurs ou des contributions additionnelles par temps de pluie. Par conséquent, l'approche proposée cherche à évaluer le potentiel de la méthode Singular Spectrum Analysis (SSA), une méthode de modélisation et de comblement de données manquantes, pour combler des séries chronologiques de temps sec. La méthode SSA est testée en reconstruisant 1000 séries chronologiques discontinues artificielles, construites aléatoirement à partir de séries réelles de débit et matières en suspension (MES) (année 2007, pas de temps de 2 minutes, système unitaire, Ecully, Lyon, France). Les résultats montrent la capacité de la méthode à combler des lacunes de données supérieures à 0.5 jour, surtout entre 0.5 et 1 jour (NSE moyen < 0.6) dans les séries chronologiques de débit. Les résultats sur les MES ne sont pas encore satisfaisants. Plusieurs analyses à différentes échelles temporelles sont envisagées.

ABSTRACT

Flow rate and water quality dry weather time series in combined sewer systems might contain an important amount of missing data due to several reasons, such as failures related to the operation of the sensor or additional contributions during rainfall events. Therefore, the approach hereby proposed seeks to evaluate the potential of the Singular Spectrum Analysis (SSA), a time-series modelling/gap-filling method, to complete dry weather time series. The SSA method is tested by reconstructing 1000 artificial discontinuous time series, randomly generated from real flow rate and total suspended solids (TSS) online measurements (year 2007, 2 minutes time-step, combined system, Ecully, Lyon, France). Results show up the potential of the method to fill gaps longer than 0.5 days, especially between 0.5 days and 1 day (mean NSE > 0.6) in the flow rate time series. TSS results still perform very poorly. Further analysis at different temporal scales might be needed.

KEYWORDS

Data validation, dry weather, gap filling, metrology, online monitoring, time series

1 INTRODUCTION

Flow rate and water quality time series measured during dry weather periods at different locations in urban drainage systems (*e.g.* sewer system, WWTP, gully pots, detention basins) can be useful for several purposes (*e.g.* modelling, real time control, water management…). However, long-term dry weather time series may contain an important amount of unregistered or invalidated data due to failures related to the operation of the sensors, errors in measurement devices, maintenance and cleaning activities or disturbing contributions during rainfall events (wet weather period). These data gaps might vary from 1 or 2 minutes to days, weeks or even months.

Previous data-driven experiences sought to estimate the dry weather signal by the use of simplified periodic equations (*e.g.* Rodriguez *et al.*, 2013) or by filling gaps with data corresponding to similar dry weather periods (*e.g.* Métadier and Bertrand-Krajewski, 2011). However, these approaches do not consider the continuity of the real and long-term dry weather time series, dismissing possible frequency-variable, non-stationary and seasonal behaviors. These simplifications might bring up inconsistent results such as overestimations of the dry weather contributions during rainfall events, or mismatches between the beginning/ending of the gap with the beginning/ending of the signal to be fitted, especially for longer gaps (beyond hourly scale) (adapted from Métadier, 2011).

Singular Spectrum Analysis (SSA) is a modern non-parametric method for the analysis of time series and digital images (Korobeynikov, 2010). The SSA method has been applied for filling gaps in long-term and non-linear time series from analogue environmental contexts, reporting encouraging results (Musial *et al.*, 2011). The aim of this study is to assess the potential of the SSA method to estimate periods of missing data (from 6 minutes to 4.3 days), which might be useful for several additional applications, such as assessing the dry weather behavior during rainfall events.

2 MATERIALS AND METHODS

The method is tested with a one year flow rate and a TSS time series of the Ecully catchment (combined system, Lyon, France). The raw data includes 261 477 measurements (year 2007, 2 min time-step), with duration of gaps ranging from 2 min to 4.3 days for flow rate and 8.29 days for TSS, throughout the whole year (3.6 % and 28 % of the year respectively). Three data processing steps are applied to the raw data:

- Removing flow rate values during dry weather greater than the 95 percentile of the flow rates measured during the preceding storm event, which are about 70 L/s (dry weather outliers). For the case of TSS, values over 590 mg/L are considered as outliers from preliminary analyses.

- Removing the wet weather periods for both flow rate and TSS series (event durations ranged from 50 minutes to 39 hours, giving a total duration of events of about 21 days of additional data to be removed, even if they are already missing values).

- Filling redundant short-gaps in both series (with durations from 2 to 6 minutes), by linear interpolation, as the purpose is to explore especially longer gaps.

After applying the above raw data processing, the total percentage of gaps in the time series increased to 13 % (flow rate) and 34 % (TSS) of the year, with gaps from 6 minutes to 4.3 days for flow rate and 6 minutes to 8.3 days for TSS. The influence of beginning and ending of rainfall events over the flow rate and TSS time series (rainfall event) is identified from previous studies in this data set (Métadier, 2011).

The SSA method is applied to fill the gaps in both flow rate and TSS time series with the function “gapfill”, from the “Rssa” package (Korobeynikov, 2010), implemented in R software (R Development Core Team, 2015). The function “gapfill” fills the missed entries in the series by performing forecast from both sides of the gap and taking an average in order to reduce the forecast error (see details: SSA sequential gap-filling method in Golyandina and Osipov, 2007). With the purpose of evaluating the performance of the SSA method in terms of predictability, a validation strategy based on the Monte Carlo method is hereby proposed. 1000 artificial discontinuous time series are generated by introducing gaps with random durations (uniformly distributed random numbers from 6 minutes to 4 days) over random parts of the original time series, with a check to guarantee a uniform distribution of gaps along the series. The additional percentage of gaps for each of the artificial discontinuous time series was set between 5 % and 30 % of the total duration of the time series (one year). The artificial time series are completed (gap filling) by the SSA method and compared to the original time series using the Nash-Sutcliffe model Efficiency (NSE). The NSE is chosen as the performance measure as

it compares the performance of the method to a model that only uses the mean of the observed data (simplest prediction method) (from Bennett *et al.*, 2013). The variability of the NSE value against gaps of different duration is analyzed as well.

3 RESULTS AND DISCUSSION

For illustrative purposes, the reconstruction obtained by the SSA method for an artificial gap (from 20/10/2007 16:23 to 23/10/2007 08:35) in the flow rate time series is compared to the original measured values, reporting a NSE value of 0.5 (Figure 1 a; line: reconstruction, dots: measurements). The NSE value is calculated between all time series fragments reconstructed by the SSA method in each of the 1000 artificial discontinuous time series and the corresponding fragments in the original flow rate and TSS time series. Regarding flow rate, the NSE values are greater than 0.6 for all reconstructed fragments in half of the 1000 artificial discontinuous time series (Figure 1 b). The cases in which the NSE values show a poor performance of the SSA method can be attributed to the complexity and the large amount of data in the series. This trend is stronger in the case of TSS time series, in which 75 % of the NSE values are lower than zero (which is the NSE value corresponding to filling the gaps with the mean of the series).

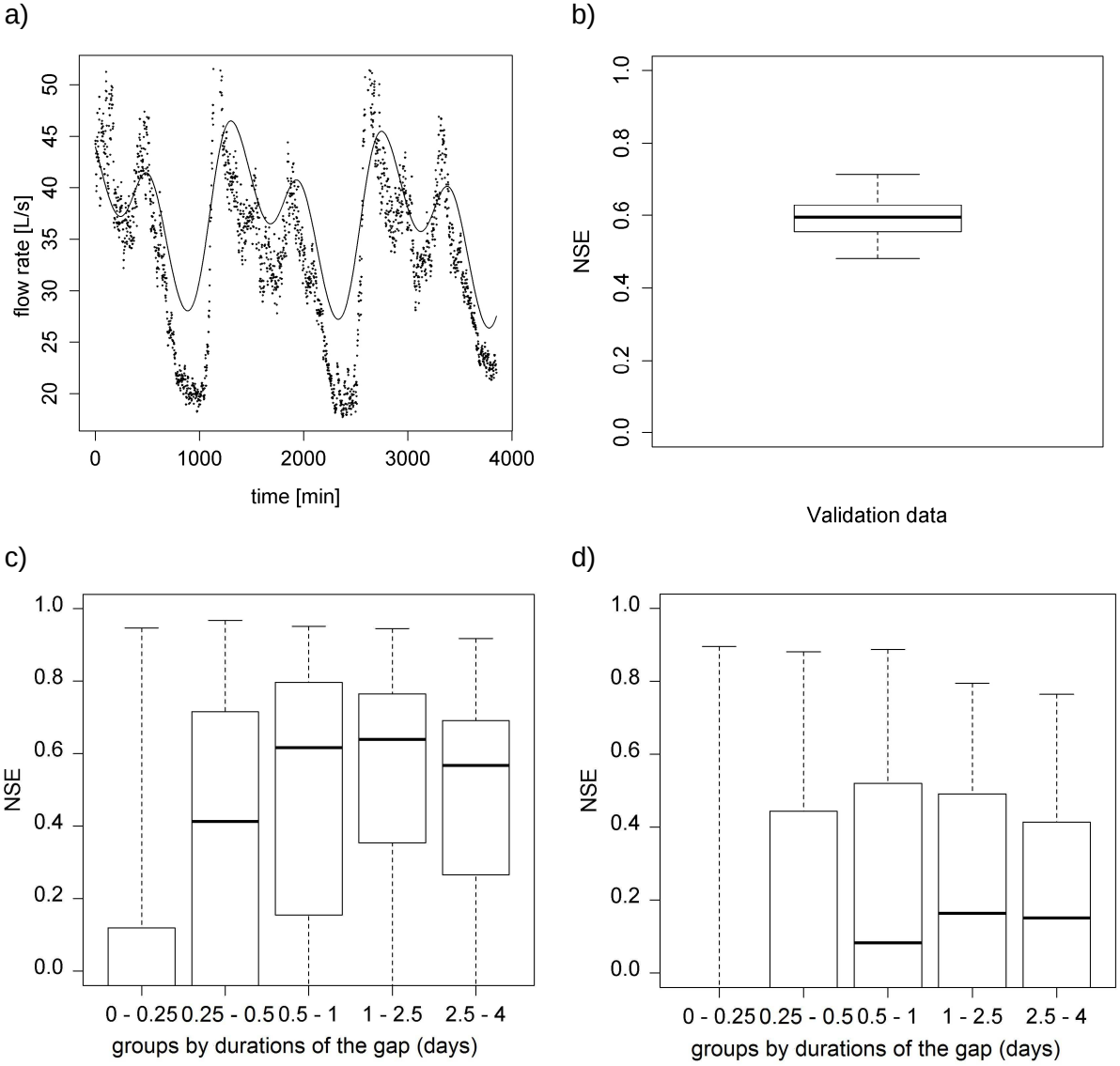

Figure 1. a) reconstruction obtained (line) for an artificial gap compared to the original flow rate values (dots), b) NSE values for the reconstruction of the 1000 artificial discontinuous flow rate, c) NSE values for gaps of different durations for flow rate and d) NSE values for gaps of different durations for TSS.

The performance of the SSA method is also analyzed by grouping the NSE values obtained for the

reconstruction of gaps of different durations. For the case of flow rate, the SSA method shows a better performance for filling long-term gaps longer than 0.5 days (Figure 1 c). Specifically, the best performances are obtained for the reconstruction of gaps with durations between 0.5 day and 1 day (Figure 1 c). For shorter gaps, the results are poorer. This can be expected, as for this case, the SSA method includes long-term (weekly to monthly scales) components that are not related with the short-term (sub-daily scale) behaviors. Therefore, the SSA method at low temporal scales (*e.g.* daily or hourly scales) might have some potential adaptability by considering exclusively a certain amount of data adjacent to the gaps consistent with the temporal scale of analysis. However, filling gaps shorter than 0.5 day with other methods that do not consider long-term patterns (*e.g.* mean values, typical dry weather daily curve or linear interpolations) might also be a suitable strategy.

The results for TSS show the same trend as for the flow rate series but with a significantly lower performance (more complex behaviors at all temporal scales) (Figure 1 d). Previous analyses highlighted the importance of finding an appropriate approach for representing the different long-term and short-term behaviors, aimed at modelling flow rate and TSS dry weather time series.

ACKNOWLEDGEMENTS

The data used in this work have been collected and made available by the OTHU project in Lyon, France (see www.othu.org). Santiago Sandoval is grateful to COLCIENCIAS (Colombian Institute for the Develop of Science and Technology) for funding his PhD studies in France.

REFERENCES

Bennett, N.D., Croke, B.F.W., Guariso, G., Guillaume, J.H.A., Hamilton, S.H., Jakeman, A.J., Marsili-Libelli, S., Newham, L.T.H., Norton, J.P., Perrin, C., Pierce, S.A., Robson, B., Seppelt, R., Voinov, A.A., Fath, B.D., and Andreassian, V. (2013). Characterising performance of environmental models. *Environmental Modelling & Software*, 40, 1-20.

Golyandina, N. and Osipov E. (2007). The “Caterpillar”-SSA method for analysis of time series with missing values. *Journal of Statistical Planning and Inference*, 137, 2642-2653.

Korobeynikov, A. (2010). Computation- and space-efficient implementation of SSA. *Statistics and Its Interface*, 3(3), 257-368.

Musial, J.P., Verstraete, M.M. and Gobron, N. (2011). Technical Note: Comparing the effectiveness of recent algorithms to fill and smooth incomplete and noisy time series. *Atmospheric Chemistry and Physics*, 11(15), 7905-7923.

Métadier, M. (2011). *Traitement et analyse de séries chronologiques continues de turbidité pour la formulation et le test de modèles des rejets urbains par temps de pluie*. PhD Thesis, Institut National des Sciences Appliquées de Lyon, Lyon, France.

Métadier, M. and Bertrand-Krajewski, J.-L. (2011). From mess to mass: a methodology for calculating storm event pollutant loads with their uncertainties, from continuous raw data time series. *Water Science and Technology*, 63(3), 369-376.

Rodríguez, J.P, McIntyre, N., Díaz-Granados, M., Achleitner, S., Hochedlinger, M. and Maksimovic, C. (2013). Generating time-series of dry weather loads to sewers. *Environmental Modelling & Software*, 43, 133-143.

R Core Team (2015). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/