Analysis of Continuous Time Series in Urban Hydrology: Filling Gaps and Data Reconstitution

Aubin Jean-Baptiste* and Bertrand-Krajewski Jean-Luc

Université de Lyon, INSA-Lyon, LGCIE, 20, avenue Albert Einstein, 69921 Villeurbanne Cedex, France; jean-baptiste.aubin@insa-lyon.fr, jean-luc.bertrand-krajewski@insa-lyon.fr
**Corresponding author*

Abstract. *We analyse flow rates and quality (conductivity, temperature, pH, turbidity) of urban discharges during dry and storm weather conditions. Missing data are present in these time series. If the length of a missing interval is short (up to 10 or 15 minutes), then simple techniques (linear interpolation or splines for example) are very effective to fulfill them. When the weather is dry and the missing data interval is longer (some hours or more), we propose a method to rebuild the missing information. First, we cluster the observed -full- dry days to identify their common properties and their variability. Then, we use a linear combination of the centers of the clusters as corner functions to fill the gaps of days with missing data. Numerical results on real dataset show the superiority of the new method with respect to a method based on an empirical clustering.*

Keywords. *Functional Clustering; Continuous Time Series; Urban Hydrology; Data Reconstitution.*

1 General Statements

Since 2004, the LGCIE (Laboratory of Civil and Environmental Engineering) of INSA Lyon, in the OTHU project (see www.othu.org), measures flow rates and quality (conductivity, temperature, pH, turbidity) of urban discharges during dry and storm weather conditions in two sewers of the Greater Lyon (Ecully, residential catchmen of 245 ha with combined sewers and Chassieu, 185 ha industrial area with a separate storm sewer). Models for water and pollutants transfer in the urban sewer systems (see [2]) are established and tested with the measured time series.

For various reasons (technical problems, human errors, etc.), missing data are present in these time series. If the length of a missing interval is short (up to 10 or 15 minutes), then simple techniques (linear interpolation or splines for example) are very effective to fulfill them. When the weather is dry and the missing data interval is longer (some hours or more), then specific methods have to be developed to rebuild the missing information. Analogous methods can be applied in case of storm weather to estimate

discharges of pollutants. This approach allows to separate the respective contributions from rainfall event unoff and dry weather wastewater (see [1]).

In this work, we focus on the filling of missing data for dry weather. First, we cluster the observed -full-dry days to identify their common properties and their variability. Then, we use the centers of the clusters as corner functions to fill the gaps of days with missing data.

After discarding the outliers, we establish a data set of 126 complete dry days between the 1st of January 2007 and the 31st of December 2008 for discharges and turbidity on the sewer located in Ecully. A calendar dry day corresponds to a time series from 12:00 a.m. to 11:58 p.m. . In the following, let's denote $\mathcal{C}_i := (c_{i,1}, \ldots, c_{i,720})$ the 720-vector associated to the i^{th} dry day (with a two minutes time step).

2 The Method

The steps of the analysis of dry days are the following ones:

- elimination of local outliers corresponding to very high gradients,
- smoothing the $\mathcal{C}_i, i = 1, \ldots, 126$ by a 10 minutes-moving average,
- subtraction of 80 % of the minimum value of each $\mathcal{C}_i$ to reduce the effects of a possible drainage of a previous storm event and fluctuations linked to the variations of the water table (infiltration water),
- definition of a distance between $\mathcal{C}_i$ and $\mathcal{C}_j$ for $j = 1, \ldots, 126$ and $i = 1, \ldots, 126$:

$$d(\mathcal{C}_i, \mathcal{C}_j) = \sqrt{\frac{1}{720}\sum_{k=1}^{720}(c_{i,k} - c_{j,k})^2} + \sqrt{\frac{1}{720}\sum_{k=2}^{720}(c_{i,k} - c_{i,k-1} - c_{j,k} + c_{j,k-1})^2}$$

  taking into account both the distance between the values of the vectors and the distance between an approximation of their derivatives. Note that for $i = j$, $d(i, j) = 0$.

- Ascending Hierarchical Classification (AHC) based on $d(.,.)$ leading to 3 clusters,
- for each cluster, determination of the center of the cluster by considering the pointwise mean of the elements of the considered clusters. Note that this center may not be a $\mathcal{C}_i$.
- for each $\mathcal{C}_i$, we artificially create missing data (gaps with duration from 1 hour to 12 hours). Let us denote the artificial gap for a fixed duration $\mathcal{G}_i$ ($\mathcal{G}_i$ is strictly included in $\mathcal{C}_i$).

By ordinary least squares, we determine from which center the data out of the gap $\mathcal{G}_i$ are the nearest ones (by calculating the mean integrated squared error). Then, we fulfill the data of $\mathcal{G}_i$ by a linear function of the nearest center. We calculate the mean integrated squared error between the estimated and the real values of $\mathcal{C}_i$ and, of course, an accurate filling (estimated data close to real ones) is associated to a small mean integrated squared error.

This method is compared to an empiric approach where 3 clusters are defined as follows:

- Cluster 1: days of the week (no holidays),
- Cluster 2: school holidays,
- Cluster 3: week-ends outside school holidays.

Obtained results show the superiority of the new method with respect to the quality index of the filling of "long term" gaps. Moreover, new clusters are more compact.

References

[1] Métadier M., Bertrand-Krajewski J.-L. (2011). Assessing dry weather flow contribution in TSS and COD storm event loads in combined sewer systems. Water Science and Technology, 63(12), 2983-2991.

[2] Métadier M., Bertrand-Krajewski J.-L. (2012). Pollutographs, concentrations, loads and intra-event mass distributions of pollutants in urban wet weather discharges calculated from long term on line turbidity measurements. Water Research, 46(20), 6836-6856.