A nonparametric framework for water consumption data cleansing: an application to a smart water network in Naples (Italy)

Roberta Padulano and Giuseppe Del Giudice

ABSTRACT

Remote monitoring and collection of water consumption has gained pivotal importance in the field of demand understanding, modelling and prediction. However, most of the analyses that can be performed on such databases could be jeopardized by inconsistencies due to technological or behavioural issues causing significant amounts of missing or anomalous values. In the present paper, a nonparametric, unsupervised approach is presented to investigate the reliability of a consumption database, applied to the dataset of a district metering area in Naples (Italy) and focused on the detection of suspicious amounts of zero or outlying data. Results showed that the methodology is effective in identifying criticalities both in terms of unreliable time series, namely time series having huge amounts of invalid data, and in terms of unreliable data, namely data values suspiciously different from some suitable central parameters, irrespective of the source causing the anomaly. As such, the proposed approach is suitable for large databases when no prior information is known about the underlying probability distribution of data, and it can also be coupled with other nonparametric, pattern-based methods in order to guarantee that the database to be analysed is homogeneous in terms of water uses.

Key words | missing data, nonparametric methods, outlier detection, Smart Water Grid, time series, water consumption data

Roberta Padulano (corresponding author)
Fondazione CMCC (Centro Euromediterraneo sui Cambiamenti Climatici),
Via Maiorise, Capua, CE,
Italy
E-mail: *roberta.padulano@cmcc.it*

Giuseppe Del Giudice
Department of Civil, Architectural and Environmental Engineering,
Università degli Studi di Napoli Federico II,
Via Claudio 21, 80125 Naples,
Italy

INTRODUCTION

Water demand modelling and forecast is a key issue in modern approaches to an efficient water management (Padulano & Del Giudice 2018). A comprehensive knowledge of water consumption allows for correct planning of water supply (Firat *et al.* 2010), for the estimate of leakages in the water distribution networks (Froukh 2001; Buchberger & Nadimpalli 2004; Bragalli *et al.* 2019) and for the development of innovative approaches and attractive plans to consumers (Jain *et al.* 2001).

The increasing interest towards water systems efficiency has led to the implementation of 'Smart Water Grids' within urban areas, with significant portions of customers connected to a telemetry system for flow data reading and collection. Smart grids allow for the collection of large amounts of data, usually on an hourly basis or less (Gargano *et al.* 2016) that water companies can utilize to calibrate bills in the short term (Cheifetz *et al.* 2017), and to perform research to increase efficiency in the long term (Cominola *et al.* 2018). Understanding consumption drivers at the customer scale can be a challenging task in a complex urban environment because of the extreme variability in the characteristics of households, such as the number of individuals served by each flow meter, water usage, which can be related to either residential or commercial activities

doi: 10.2166/hydro.2020.133

(Padulano & Del Giudice 2019), and different life habits of the end users (Brentan *et al.* 2018). One common approach to solve this problem is the profiling (Wright 2009), namely a detection of demand patterns based on a large amount of data; this is a typical approach in the electricity sector (Rasanen *et al.* 2010; López *et al.* 2011; Ferreira *et al.* 2013; Zhou *et al.* 2013; Macedo *et al.* 2015), with a few applications for water demand modelling (McKenna *et al.* 2014; Avni *et al.* 2015). Profiling of consumption data is typically performed to catch differences in the customers' behaviour, with particular focus on the weekdays/weekends distinction (Padulano & Del Giudice 2019), especially when no previous information is known.

MOTIVATION

Thanks to recent advances in technology, the remote monitoring and collection of water consumption has gained pivotal importance in the field of demand understanding, modelling and prediction (Loureiro *et al.* 2016b). The amount of information that can be extracted, with different degrees of complexity, from a significantly large database of water consumption is considerably vast and diverse, including basilar statistics (mean, median, modal values, standard deviations among others), temporal and spatial correlation, aggregation and disaggregation, pattern detection, extreme value analysis, that can be performed by means of parametric or nonparametric approaches.

However, most of these statistics and analyses could be invalidated by inconsistencies in the database due, for example, to the following:

- Technological issues, such as data transmission interruptions or flow meter malfunctioning; this usually leads to gaps in the recorded data series of a single connection or of a group of connections relying on the same hub. Those gaps are usually represented by missing data, but according to the type of malfunctioning could be translated in long sequences of null (namely zero) values; in any case, doubts can be cast about the reliability of those flow meters with significantly long anomalous sequences.
- Behavioural issues, such as a change in the household served by a connection (e.g. a temporary or permanent vacation, a relocation); this can lead to gaps or long sequences of zero data if the user connection is not properly or promptly deactivated, or in changes in the mean consumptions, producing stepped annual patterns. Again, doubts can be cast about the reliability of those flow meters with significantly long gaps.
- Hydraulic issues, such as leakages, causing bursts in the consumption values recorded by flow meters placed along supply lines (Loureiro *et al.* 2016a), or acting as unexpected consumption values (Buchberger & Nadimpalli 2004), especially at night-time (Mazzolani *et al.* 2016).
- Anomalous recorded data associated with the randomness of human behaviour in terms of water consumption, producing 'outliers', namely observations that significantly differ from the others (Johnson & Wichern 1992; Barnett & Lewis 1994). Strictly speaking, outliers cannot be considered an inconsistency since they are possible values, although associated with low probabilities; however, a massive presence of outliers in a series could affect analysis and could also cast doubts about the reliability of the time series as a whole.

As seen, different sources of inconsistency produce similar effects in terms of sequences of missing or null values and of anomalous data. Particular attention should be paid to null data, since they are not necessarily related to any malfunctioning in the system but can be the realization of a non-consumption event (Gargano *et al.* 2016). However, the occurrence of a large amount or of a long series of zero data could mask a technological or behavioural issue, especially for coarse/medium time resolutions (e.g. hourly records) (House-Peters & Chang 2011). A failure in detecting such inconsistencies could cause model misspecification, biased parameters estimation and incorrect results (Ben-Gal 2005).

Procedures aimed at detecting inconsistencies usually rely on validation (Loureiro *et al.* 2016a), which is only possible when the sources producing anomalous or inconsistent values are known. However, this seldom occurs for water consumption databases, and especially in large water districts, where remote monitoring and control of the involved variables are usually not an operational standard. In these cases, unsupervised techniques must be

adopted, possibly relying on automated procedures due to the typically large size of the collected databases, which is, in turn, a function of the sample frequency.

In the present paper, a nonparametric framework is proposed for the pre-processing of water consumption datasets, particularly focusing on the following:

- Reliability of time series 'as a whole', providing a strategy for identifying gaps in data and checking whether they imply the rejection of the 'suspicious' connections, according to the specific purpose of the overall analysis;
- Reliability of single data, providing a strategy for identifying outliers and checking whether they imply the rejection of the 'suspicious' connections, according to the specific purpose of the overall analysis.

MATERIALS AND METHODS

Detection of anomalies

Whatever the purpose of the analysis, there are no objective criteria other than common sense to tell whether anomalous data sequences of missing and zero data are so long and/or frequent that they provide significant alterations. In this context, it is useful to define the 'completeness' and the 'continuity' of a time series (Braca *et al.* 2013). Completeness C_1 represents the amount of valid data with respect to the maximum possible number of data in a time series; continuity C_2 describes in which measure the time series is interrupted by invalid data. Completeness and continuity are numerically defined as follows:

$$C_1 = \frac{N_{val}}{N} \tag{1}$$

$$C_2 = 1 - 2 \cdot \frac{n_{inv}}{N} \tag{2}$$

where N is the expected maximum number of data in the time series, N_{val} is the number of valid data in the time series and n_{inv} is the number of intervals of invalid data in the time series. The concepts of continuity and completeness of a time series are not new; for instance, completeness can be considered a common sense metric for time series quality evaluations. However, the present research shows that C_1 and C_2 are mutually dependent since the number of valid data limits the maximum and minimum possible number of intervals of invalid data. As a consequence, all the possible combinations of C_1 and C_2 in a time series must be comprised within a fixed triangle (henceforth defined as 'Completeness/Continuity Triangle' CCT), shown in Figure 1, delimited by three lines describing the boundary correlations between completeness and continuity.

Maximum continuity $C_{2\max}$ is obtained when data are organized in one valid data interval followed by one invalid data interval, or vice versa; in this case, $n_{inv} = 1$ and $C_2 = 1–2/N$, independent of C_1 (vertical line in Figure 1). For $C_1 < 0.5$ (namely the number of invalid data N_{inv} is higher than the number of valid data N_{val}), minimum continuity is obtained when the valid data intervals are made up of just one data, and n_{inv} is equal to $N_{val} + 2$. For $C_1 > 0.5$, minimum continuity is obtained when the invalid data

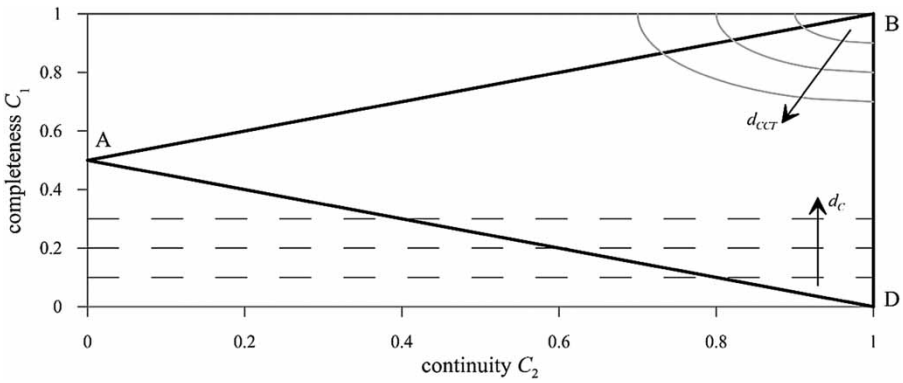

Figure 1 | An example of Completeness/Continuity Triangle (CCT) for negligible 4/*N* (vertices, completeness thresholds and compound thresholds).

intervals are made up of just one data, and n_{inv} is bound to be equal to N_{inv}. The equations describing minimum continuity for completeness values lower than 50%, and minimum continuity for completeness values higher than 50% (lines in Figure 1) can be obtained by assuming $C_1 = 0.5$ in Equation (1) and $n_{inv} = N_{val} + 2$, for the lower line, and $n_{inv} = N_{inv}$, for the upper line, in Equation (2):

$$\begin{cases} C_{2\,\max} = 1 - \dfrac{2}{N} & \\ C_{2\,\min} = 1 - 2C_1 - \dfrac{4}{N} & \text{for } C_1 < 0.5 \\ C_{2\,\min} = 2C_1 - 1 & \text{for } C_1 > 0.5 \end{cases} \tag{3}$$

For $C_1 = 0.5$, the minimum continuity condition is achieved when the time series is made up of an alternation of one valid and one invalid data, so that $n_{inv} = N_{inv} = 0.5\,N$ resulting in $C_2 = 0$ (vertex 'A' in Figure 1). Vertices 'B' and 'D' in Figure 1 both have maximum continuity $C_2 = 1$, but only B is concretely possible because it corresponds to a perfect time series with all valid data, whereas D corresponds to a time series with no valid data. For large N values, the terms $2/N$ and $4/N$ in Equation (3) are negligible with respect to the remaining terms.

There are no universally accepted thresholds for C_1 or C_2; however, their mutual correlation makes it necessary to set a threshold for just one of them, reducing subjectivity. If the C_1 threshold is fixed to 0.5, C_2 has its larger span, ranging between 0 and $C_{2\max}$; if the C_1 threshold is increased, the accepted time series will be more and more continuous (and therefore reliable) the higher the threshold. In turn, a low although adequate C_1 threshold could result in an unacceptably discontinuous time series, depending on the purpose of the analysis; to overcome this issue, compound thresholds could be preferred accounting for both C_1 and C_2 with a single metric. In the present paper, a simple compound distance metric was adopted defined as:

$$d_{CCT} = \sqrt{(1 - C_1)^2 + (1 - C_2)^2} \tag{4}$$

Such a metric gives $d_{CCT} = 0$ for time series having completeness and continuity equal to one, and $d_{CCT} = 1$ for time series having zero completeness and unitary continuity, namely the worst condition, represented by vertex D (Figure 1). A ranking of time series provided by d_{CCT} can be adopted to infer about the reliability of user connections, and, consequently, as a decision support tool for different applications. For instance, time series with d_{CCT} values exceeding a user-defined threshold could be considered unreliable and rejected. The definition of such a threshold could be driven by expert knowledge (e.g. an assumption about the expected amount of zero values) or could be driven by data (e.g. an assumption about the amount of time series that should be retained to provide significant results). In the paper, for anomaly detection purposes, comparisons will be made between results obtained by adopting the compound threshold d_{CCT} and those obtained by assuming a simple completeness-based threshold d_C (Figure 1). It should be noticed that the two thresholds are correlated since the first includes the second, but it is more restrictive because it also gives limitations for continuity. For example, if the completeness threshold is set to a specific value d_C^*, a number N_C of flow meters will be retained having $C_1 \geq d_C^*$; if the compound threshold is set to a value that guarantees the same minimum completeness, namely $d_{CCT}^* = 1 - d_C^*$, a number $N_{CCT} < N_C$ will be retained.

Outlier detection

Outlier detection methods can be classified as parametric and nonparametric: parametric methods assume a known underlying distribution of the observations, whereas outliers are supposed to have a different distribution (Schwertman *et al.* 2004); nonparametric methods do not rely on any assumption about underlying probability distribution. Parametric methods classify as outliers those data that deviate from the probability model assumptions; they are often unsuitable for large high-dimensional datasets and without prior knowledge of the underlying data distribution (Papadimitriou *et al.* 2003; Ben-Gal 2005). Under these conditions, instead, nonparametric methods find their most suitable application (Ben-Gal 2005); examples of nonparametric methods are the distance-based techniques, measuring the distance of data from some significant central values (Knorr & Ng 1997) and the clustering methods, based on the concept that a cluster with only a few elements in it could represent a set of outlying data (Shekhar *et al.* 2002; Padulano & Del Giudice 2018; Padulano *et al.* 2018).

Distance-based methods are usually based on local distance measures (Fawcett & Provost 1997; Knorr & Ng 1997; Williams & Huang 1997; Mouchel & Schonlau 1998; Breunig *et al.* 2000; Knorr *et al.* 2000, 2001; Jin *et al.* 2001; Hawkins *et al.* 2002; Bay & Schwabacher 2003). When using distance-based techniques, the median and median-based parameters are usually preferred to the mean because they are considered more robust, since the median is very insensitive to the presence of outliers (Cousineau & Chartier 2010; Leys *et al.* 2013). One of the most used distance metrics for outlier detection is the Median Absolute Deviation (*MAD*), defined as follows:

$$MAD = \text{median}(x_i - \text{median}(x_i)) \tag{5}$$

with x_i being the data in the dataset. Outlier detection by *MAD* consists of eliminating data that satisfy the following condition:

$$x_i - \text{median}(x_i) > |\delta \cdot c \cdot MAD| \tag{6}$$

which implies an underlying symmetry in the distribution of data. According to Equation (6), outliers are detected as those data that exceed a limit distance from the median represented by *MAD* multiplied by a scaling factor c that is needed to make the estimator consistent for the parameter of interest. In case of Gaussian distribution of data, disregarding the abnormality induced by outliers, c is set to 1.4826 (Rousseeuw & Croux 1993); if another underlying distribution is assumed, this value changes to $c = 1/Q_{0.75}$, where $Q_{0.75}$ is the 75% quantile of that underlying distribution (Leys *et al.* 2013). The parameter δ represents the rejection criterion, whose value is subjectively decided according to the research purpose; Miller (1991) proposes a value of 3 (very conservative, namely most data are preserved), 2.5 (moderately conservative) or 2 (poorly conservative, namely a large amount of data could be identified as outliers).

As previously mentioned, in water distribution networks outliers are usually associated with abnormal household water uses, changes in network system operation, breaks in pipelines or service connections and flow meters or telemetry malfunctions (Loureiro *et al.* 2016a). In each case outliers can show themselves as isolated anomalous values or as long periods of significantly different data with altered base values (for example, a long series of constant consumption). In the first case, once the outliers have been detected, they can be either considered as missing data (if the remaining sample size is not significantly reduced) or replaced with values deriving from some theoretical considerations, such as the maximum likelihood criterion or probabilistic methods (Rustum & Adeloye 2007).

In practical applications, outlier detection and treatment are undertaken with very little prior knowledge of both the underlying probability distribution of data and the processes that generated the outliers. A drawback of such a condition is that the usual criteria for outlier detection, such as Equation (6), could be incorrect, either because of the implicit assumption of symmetrical distribution of data or for the choice of the multiplicative coefficient c. To overcome this issue, which is fundamental in order to apply any outlier detection methodology, Cousineau & Chartier (2010) suggest performing a data transformation so that the transformed distribution becomes Gaussian. Such a transformation is a modification of the classical square root transformation:

$$x_{iT} = \sqrt{\frac{x_i - x_{\min}}{x_{\max} - x_{\min}}} \tag{7}$$

in which $x_{\min}$ and $x_{\max}$ are the smallest and largest values of the dataset, respectively. Dividing by ($x_{\max}$–$x_{\min}$), transformed data are forced to range between 0 and 1; then, the square root operator enlarges the smallest data, pushing the lower part of the distribution towards a more central location. As a consequence, the probability distribution of transformed data is more symmetrical than the original one; however, resulting symmetry could not be sufficient if the original distribution is too skewed or if outliers concentrate within only one side of the function. In case of deeply skewed distributions, the problem of outlier detection remains unsolved (Cousineau & Chartier 2010).

Data description

The district metering area (DMA) which is the subject of the study (Figure 2) is located in the northwestern part of the City of Naples (Italy). This area was chosen as a pilot area for a Smart Water Grid implementation, with particular focus on the remote monitoring of flow meters, as part of a co-operation between the University of Naples and ABC – Napoli, which is the local water company. The DMA is provided with 4,254 customer connections whose flow meters were completely replaced during the last three years. There are 3,701 (87% of the total number) residential flow meters, whereas the remaining 553 (13%) corresponds to commercial flow meters, offices and public buildings, consistently with the consideration that the neighbourhood mainly has a residential purpose.

The present paper focuses on single-household flow meters, that constitute 76% of the residential flow meters ($N = 2808$); for each flow meter, 12 months of hourly consumption measurements are available, dating January 1st, 2016 to December 31st, 2016, each data representing the cumulative water volume consumed by the household in the hour preceding the record. In the present paper, data quality control, cleansing and pre-processing procedures preceding the analyses presented in Padulano & Del Giudice (2018, 2019) and Padulano *et al.* (2018) are shown, which were only mentioned in those papers but not shown in detail.

RESULTS AND DISCUSSION

The approaches proposed in the methodological section were applied to the flow meter database of the DMA of Soccavo ($N = 2808$). In detail, the analysis is structured in three consecutive steps:

1. Preliminary check. This first step consists of the visual inspection of a consumption map (namely a simple, graphical representation of the database) in order to obtain general information about features and criticalities of the consumption database (e.g. inconsistencies due to outlying values or to anomalous sequences of null or missing data), driving subsequent analysis.
2. Anomaly detection. In this second step, the reliability of time series as a whole is questioned by means of the CCT, and different metrics are proposed both to obtain a reliability ranking of flow meters and to remove low-quality time series.
3. Outlier detection. In this third step, outlier detection is performed by means of a cascading approach complying with different temporal aggregation levels; again, a metric is suggested to improve results and to remove insignificant time series.

Preliminary analysis

A preliminary analysis of the database showed that, apart from a small number of flow meters whose telemetry

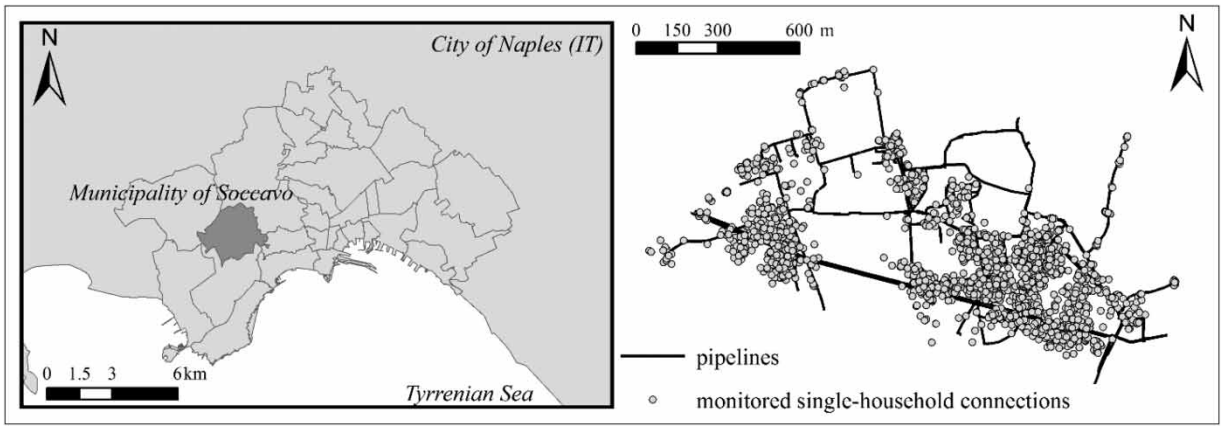

Figure 2 | The DMA of Soccavo (Naples, Italy) with focus on single-household flow meters.

never became operational, no significant missing data issues occurred for any of the user connections in the DMA. For this reason, a quality check was only performed in relation to zero data, substituting the occasional missing data with null values. This approach was considered valuable since the proposed procedure only aims at identifying unreliable flow meters, irrespective of the specific anomaly source causing invalid data, coping with both zero and missing values with one single computation. On the other hand, only positive consumption values are used for outlier detection purposes, so that an 'artificial' inflation of null values does not affect those computations.

As a preliminary control, Figure 3 shows the whole database in a synthetic but effective fashion, named 'consumption map'. In the map, each recorded data is represented as a pixel coloured according to its consumption value (Figure 3 shows hourly recorded data Q_h, in L/h), and associated to the time of recording (abscissa of Figure 3) and to the flow meter label (ordinate of Figure 3). This kind of representation allows for the immediate identification of features and criticalities, driving the following steps of quality control. The following information can be extracted by a simple visual inspection of Figure 3:

1. The adopted colour bar depicts high values of positive consumption in the shades of yellow; in Figure 3(a), yellow data are only visible as either horizontal lines or recurring patterns, regularly alternated with dark blue values (low consumption). This suggests that high values mostly reproduce high consumption dynamics, occurring at specific hours of the day, and sparse outliers are possible (although not easily detectable in Figure 3) but do not necessarily entail reliability issues, although this will be verified with a tailored analysis whose results are shown in the following sections. Lined yellow data, instead, refer to flow meters with particularly high mean consumption, whose representativeness of single-household consumption will be questioned, again, with a tailored procedure.
2. Zooming to a restricted temporal range and to a small group of flow meters (Figure 3(b)), the occurrence of zero data during night-time is particularly evident, suggesting that, despite the quite coarse time resolution

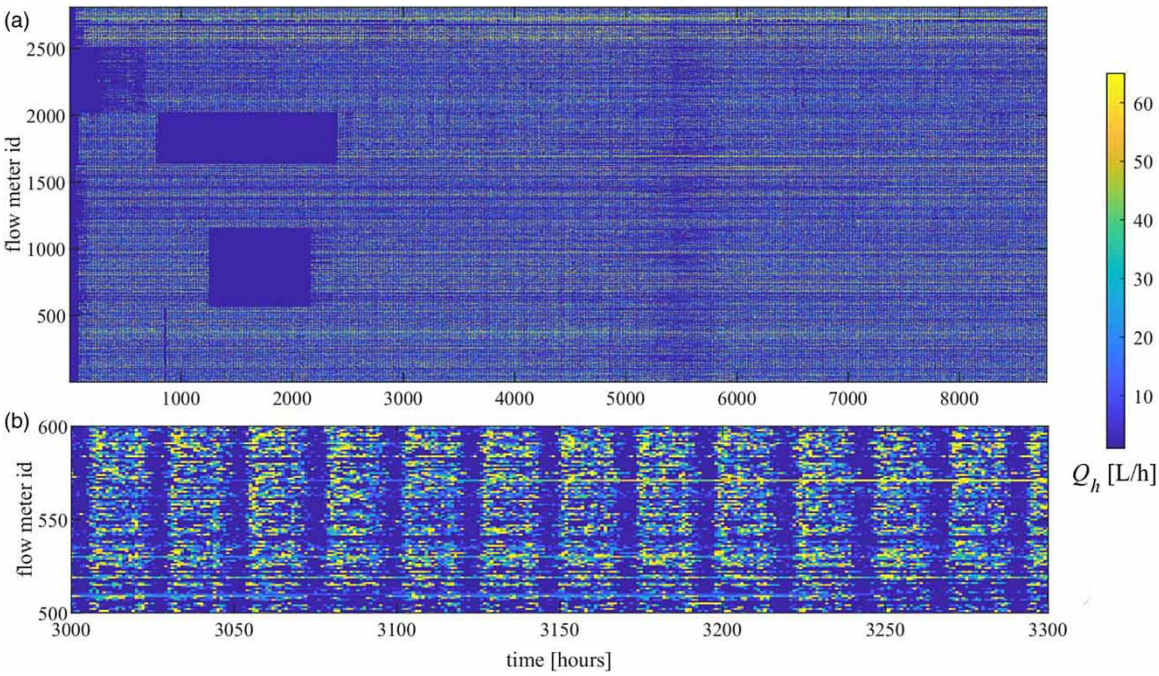

Figure 3 | Map of hourly raw consumption data for the DMA (a) and a zoom to a selection of flow meters and hours (b). Please refer to the online version of this paper to see this figure in colour: http://dx.doi.org/10.2166/hydro.2020.133.

of recordings, night consumption is generally very low and can attain null values, providing preliminary information about possible leakages in the network, often revealed by unexpected, systematic positive night consumption values (Mazzolani *et al.* 2016).

3. Zero values concentrate in several areas of Figure 3(a), and specifically:
   - In two 'voids' or clusters of null data; the regularity of those voids in time (abscissa) and space (ordinate) suggests that those data are not a realization of a non-consumption event, but they represent a malfunctioning of the system. In this case, the labelling of flow meters was performed in such a way that sequent labels correspond to spatially contiguous flow meters (e.g. in the same building or block). As a consequence, those voids are presumably due to a failed connection to the transmission hub.
   - In the first days of the year; indeed, a malfunctioning was reported during the Christmas 2015 holidays but was only repaired after New Year's Day, with a small delay for connections 2000–2500.
   - During the month of August (hours 5112–5856); in this case, zero values represent the typical residential behaviour of spending summer holidays out of town; this is corroborated by the observation that in the month of August zero values seem to be randomly scattered across the connections.

Anomaly detection

Visual inspection of Figure 3 gives qualitative information about the presence of null values in the time series, highlighting that some consumption time series could not be reliable or representative for the analysis of single-household water consumption. To reduce subjectivity, reliability is tested by means of the CCT, shown in Figure 4, based on Equations (1) and (2) applied to hourly consumption data, where each flow meter is represented by a point in the CCT. It is particularly evident that points gather along a parabolic curve with the equation:

$$C_2 = 1 - C_1 + C_1^2 \tag{8}$$

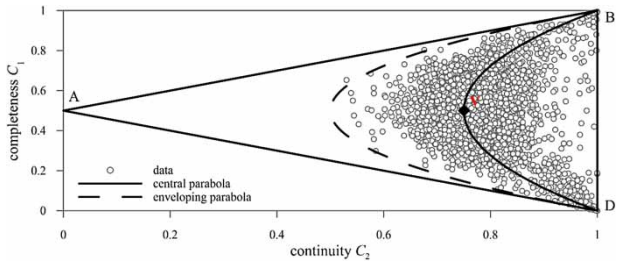

Figure 4 | Completeness/Continuity Triangle for raw hourly data (total number of data points is $N = 2808$).

and they are enveloped by a parabolic curve having the equation:

$$C_2 = 1 - 2 \cdot C_1 + 2 \cdot C_1^2. \tag{9}$$

which is tangent to the $C_{2\,\min}$ lines in vertices B and D. It should be noted that polynomial coefficients in Equation (8) were obtained by a standard Ordinary Least Square regression; indeed, Equation (8) can be regarded as the regression curve providing the best prediction of continuity for each possible completeness value, so describing the 'average' characteristics of the datasets in terms of completeness/continuity correlation. Instead, coefficients in Equation (9) were only assumed as those providing the limiting envelope for data points; Equation (9) can be regarded as the lower limit in the variation range of continuity for each completeness value, being the upper limit represented by the vertical bound describing $C_{2\max}$. Moreover, although Equation (9) limits the space that can be potentially occupied by the time series, namely the triangle delimited by Equation (3), this limitation should be considered peculiar of the database of interest and there is no evidence that this can be considered a general result.

A different perspective (Figure 5(a)) shows that points are not uniformly distributed along Equation (8) but gather at vertex V (having $C_1 = 50\%$ and $C_2 = 75\%$), implying that the completeness/continuity features of the DMA database occur with a frequency that can be interpreted by a bivariate probability distribution, with modal value represented by the coordinates of vertex V of Equation (8) and marginal distributions shown in Figure 5(b) and 5(c). Moreover, probability mass spikes can be observed at vertices B and D, showing the typical behaviour of an upper and lower limited probability distribution (Buchberger & Nadimpalli 2004).

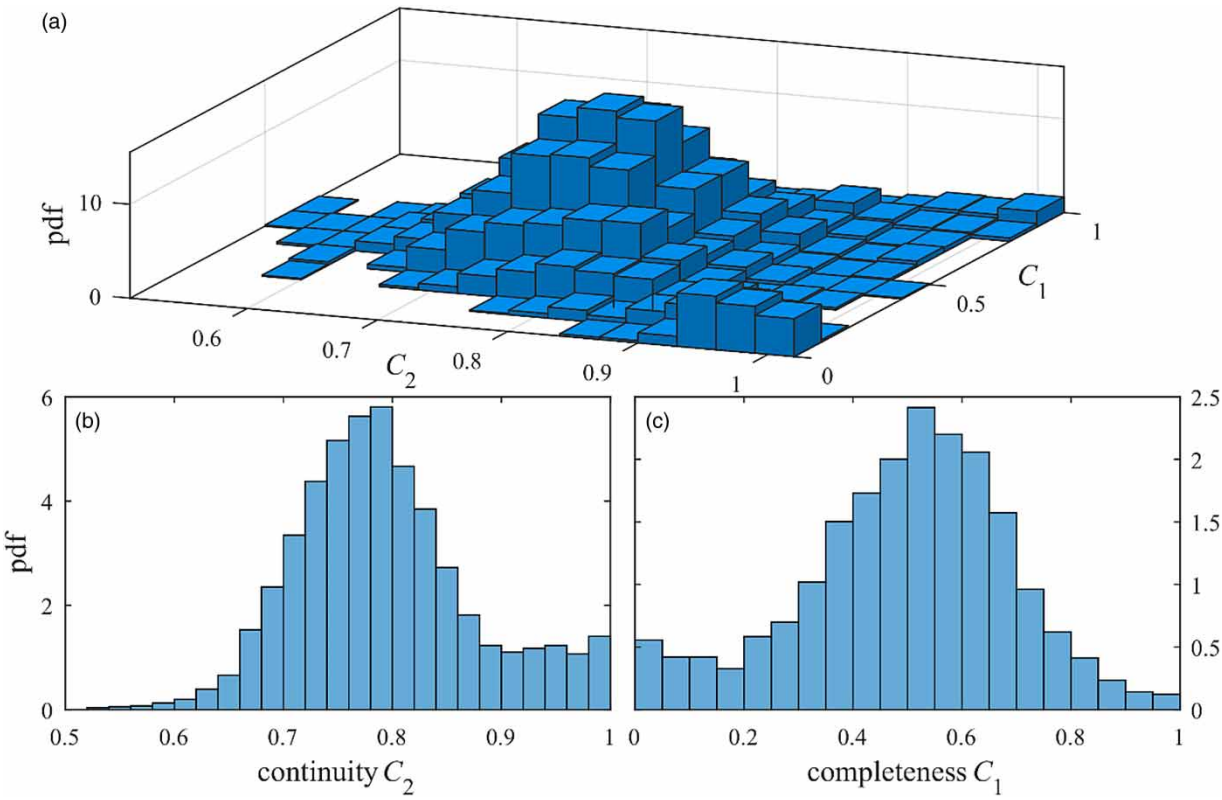

Figure 5 | Joint occurrence frequency of C_1 and C_2 values in the database (a) and marginal frequencies for C_2 (b) and C_1 (c) (total number of data points is $N = 2808$). Please refer to the online version of this paper to see this figure in colour: http://dx.doi.org/10.2166/hydro.2020.133.

As Figure 5(a) shows, the distribution of completeness/continuity data is roughly symmetrical, implying that the mean, median and modal values can be considered coincident and corresponding to the coordinates of vertex V. This, in turn, implies that the average characteristics of the database are a completeness of around 0.5 and a continuity of around 0.75, against the optimal value of 1 for both variables. Under this premise, it can be useful to establish a ranking of time series that accounts for the distance of each point from the optimal time series, represented by vertex B.

Figure 6 shows a possible ranking of the time series in five classes defined by fixing four increasing values for the threshold in Equation (4), which can be represented in the

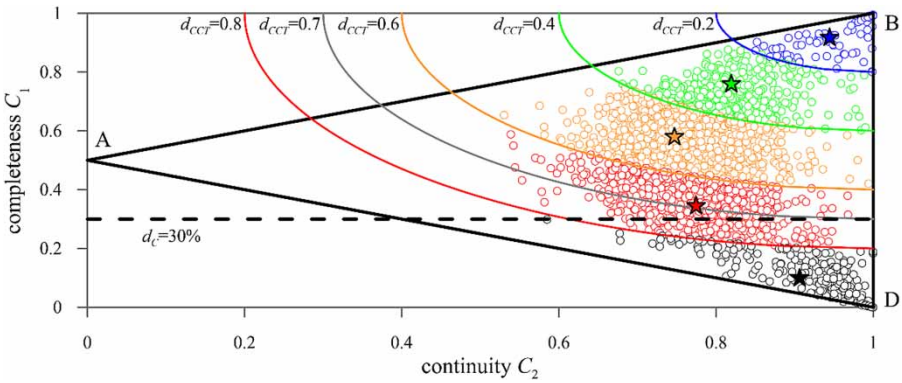

Figure 6 | Time series ranking according to compound threshold d_{CCT} and completeness threshold d_C (total number of data points is $N = 2808$).

CCT as circumferences centred at vertex B. The five classes are class 1 ($d_{CCT} \leq 0.2$), class 2 ($0.2 < d_{CCT} \leq 0.4$), class 3 ($0.4 < d_{CCT} \leq 0.6$), class 4 ($0.6 < d_{CCT} \leq 0.8$) and class 5 ($d_{CCT} > 0.8$). In other words, class 1 is made up of time series close to the optimum; as such, they can be considered particularly suitable to be up-scaled to weekly, monthly and annual resolutions. Class 5 is made up of the less reliable time series close to vertex D; applications in this case should be limited and up-scaling should be avoided. Finally, classes 2–4 are made up of increasingly unreliable time series. In any case, applications should also account for the probability distribution shown in Figure 5(a), which affects the size of the classes ($N_1 = 73$, $N_2 = 400$, $N_3 = 1286$, $N_4 = 776$, $N_5 = 273$); for instance, in the database of interest limiting the in-depth analysis to class 1 implies discarding most of the time series, possibly neglecting significant information content (e.g. spatial correlation).

Figure 7 shows the number of retained time series as a function of two different thresholds, one specifically accounting for completeness only (d_C), and the other accounting for the compound metric d_{CCT}. To better visualize the reciprocal connection of the two threshold metrics, in Figure 7 for each threshold d_C^* for completeness, a consistent compound threshold $d_{CCT}^* = 1 - d_C^*$ was set that includes d_C^*, being more restrictive: for instance, a reference value of 0.3 in the vertical axis of Figure 7 implies that all the flow meters in the database will be retained having $C_1 \geq 0.3$, if the completeness threshold is considered (blue line in Figure 7), whereas all the flow meters will be retained having $d_{CCT} \leq 0.7$ if the compound threshold is considered (red line in Figure 7). In the former case, the retained time series are 2386; in the latter case, among these time series the ones with low values of continuity will be also rejected with a final number of 2274 (these thresholds are also visible in Figure 6). It is worth noticing that $d_C^* = 0.3$ is roughly consistent with the assumption that a null consumption for 8 hours per day, on average, is admissible; this is corroborated by the observation that $d_C^* = 0.3$ and $d_{CCT}^* = 0.7$ both correspond to an elbow in the respective curves.

Figure 6 points to five time series that can be considered representative of the five classes, chosen along the parabolic curve in Equation (8) for intermediate completeness values; those time series are also shown in Figure 8, where differences across the classes are highlighted. It is worth noticing that low values of completeness and continuity (Figure 8(e)) hinder the adoption of that time series for further analysis. In turn, high values of completeness and continuity are not a guarantee of reliability and/or representativeness, as shown by Figure 8(a) where no zero values occur for two-thirds of the series, possibly revealing some kind of leakage or network malfunctioning. Moreover, it is interesting to note that the time series in Figure 8(b) corresponds to one of those flow meters characterized by a gap (the shortest one) in Figure 3(a); however, the overall quality of the series in the remaining days is so high that the corresponding flow meter belongs to one of the most reliable classes.

In the present paper, it was decided to discard flow meters belonging to class 5, namely those having $d_{CCT} > 0.8$; the remaining 2535 were passed to the outlier detection procedure.

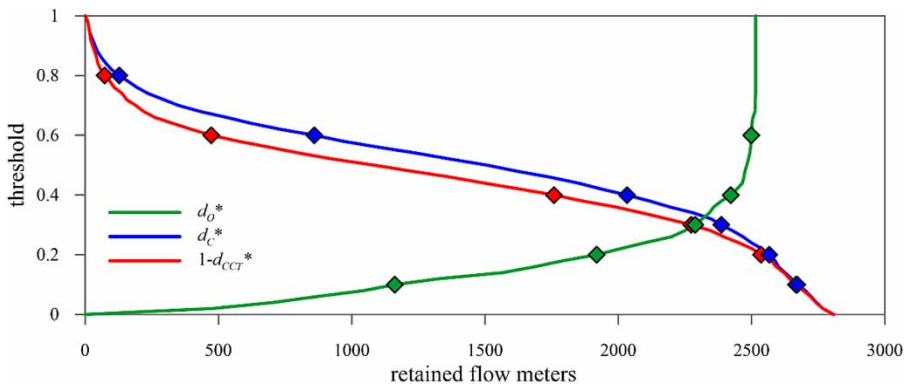

Figure 7 | Acceptance thresholds for hourly and daily data. Please refer to the online version of this paper to see this figure in colour: http://dx.doi.org/10.2166/hydro.2020.133.

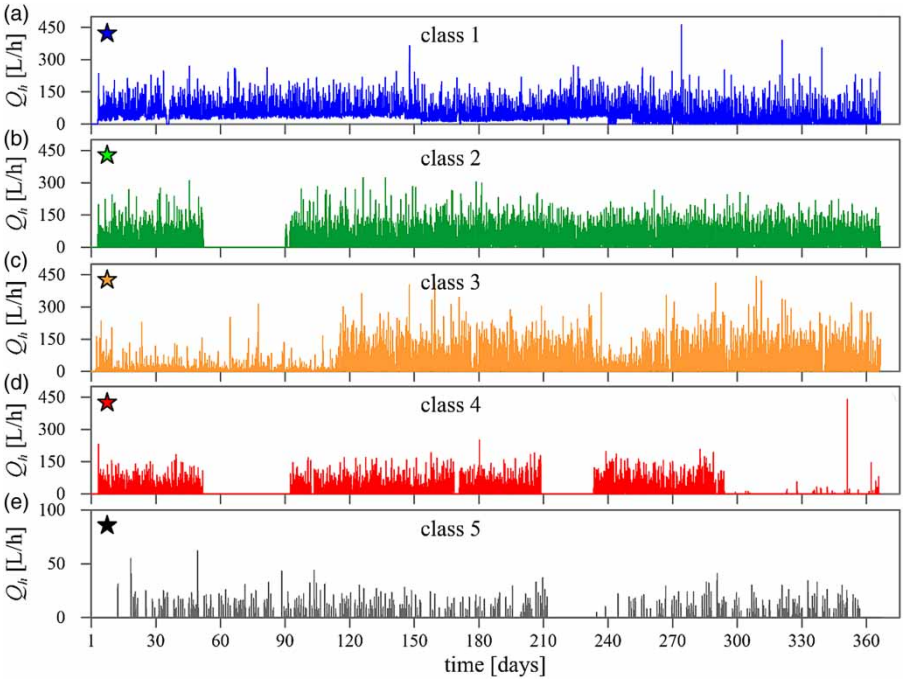

Figure 8 | Representative time series for the five d_{CCT} classes.

Outlier detection

The outlier detection criteria in Equations (5)–(7) were applied following a top-down approach; in other words, outliers were not identified based on the analysis of raw hourly data, but different temporal aggregation levels were considered in a cascade. This approach was preferred because including in the outlier detection procedure hourly data of outlying flow meters (e.g. wrongly labelled connections) could significantly affect the analysis, resulting in excessively large thresholds for outlier removal. In this respect, it is worth noticing that the definition of an outlier threshold is questionable since extreme data are associated with a low but positive frequency. However, probability distribution of consumption values may differ according to the typology of consumers connected to the system (residential, single-household, multiple-household, commercial activities, public buildings, among others); as a consequence, frequent values for a group of consumers could be outliers for a different group, and vice versa. If the composition of the database in terms of consumers is not known, a frequency analysis of a distinctive variable could reveal multiple populations in the sample. For instance, a histogram of mean or cumulative annual consumption usually allows for discriminating flow meters serving multiple households (e.g. a building) from flow meters connected to a single flat, since significantly higher values are expected in the former case. In the present paper, residential connections in the DMA were labelled according to the number of households served and the database of interest is limited to single-household connections. However, the occurrence of wrong labels was tested by applying the outlier detection criteria in Equations (5)–(7) to the annual mean of positive consumption for each time series in classes 1, 2, 3 and 4.

Figure 9(a) shows the frequency of annual mean positive consumption Q_a, computed for each time series and normalized according to Equation (7). The *MAD* criterion in Equation (6) provides a median of 28.4 L/h and an upper limit of 87.8 L/h; given the considerable asymmetry in the data distribution (suggested by the residual asymmetry in normalized data), an acceptance threshold $c = 4$ was set and no lower limit was considered. The 20 time series with outlying annual mean discharges were removed from the initial dataset with the assumption that they were incorrectly labelled as representative of single households

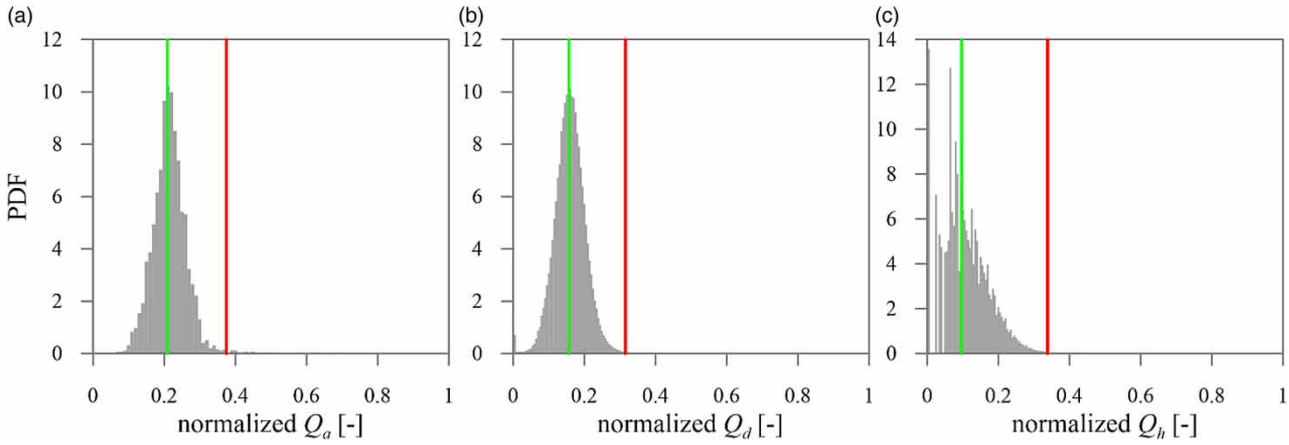

Figure 9 | Occurrence frequency of normalized annual (a), daily (b) and hourly (c) positive consumption data, with median (green line) and upper limit (red line) for outlier removal. Please refer to the online version of this paper to see this figure in colour: http://dx.doi.org/10.2166/hydro.2020.133.

or, as Figure 8(a) shows, affected by an anomalous baseline. For the 2515 retained time series, daily aggregated data Q_d were considered as realizations of a unique random variable, whose normalized distribution has only a little residual skewness (Figure 9(b)). The *MAD* criterion in Equation (6) provides a median of 27.2 L/h and an upper limit of 106.7 L/h; again, an acceptance threshold $c = 4$ was set and no lower limit was considered. Across the whole database 13% (122,510 against $2515 \times 366 =$ 920,490 total daily values) of daily data higher than the upper limit were labelled as outliers and removed from the original time series, along with the corresponding hourly data. At this stage, reliability was questioned for those time series having a considerable amount of invalid daily data (namely missing daily data or daily outliers, now removed); to solve this issue, a threshold d_O^* was set to exclude from possible following analyses those time series whose percentage of invalid daily data (namely, the number of invalid daily data divided by the maximum possible number of daily data, equal to 366 for year 2016) exceeds that threshold. Figure 7 shows the number of retained flow meters as a function of d_O^* (green line in Figure 5); again, $d_O^* = 0.3$ (namely, the rejection of time series having more than 110 days with invalid daily data) represents an elbow in the corresponding curve. However, a threshold of 0.3 could be too large since it implies more than three cumulative missing months; in turn, setting a threshold of 0.1, roughly corresponding to 36 missing days of measurements, implies rejecting more than half the database (Figure 7), with a number of retained flow meters equal to 1162. Of course, d_O^* can be a function of the specific purpose of the analysis; however, a wise strategy could involve the retention of the most reliable time series (e.g. those having a percentage of daily outliers lower than 10%) to calibrate subsequent procedures, such as hourly outlier detection, and, successively, the reconsideration of those time series where daily outliers were caused by sparse extreme hourly outliers.

For the remaining 1162 flow meters, all non-zero data Q_h were considered as realizations of the same random variable representing water consumption in Soccavo, whose distribution is so skewed that a significant asymmetry persists in the normalized data provided by Equation (7) (Figure 9(c)). The *MAD* criterion in Equation (6) provides a median of 19.00 L/h and an upper limit of 222.4 L/h; given the deep asymmetry in the data distribution, an acceptance threshold $c = 4$ was set and no lower limit was considered.

Raw data higher than the upper limit were labelled as outliers and removed from the original time series; however, all the 1162 time series were retained because the maximum percentage of removed data was very low (about 4%). It is worth noticing that if daily outliers are not removed, but only the flow meters with an excessive number of daily outliers, the median and upper limit values do not change, implying a very robust distribution in hourly data. However, the maximum percentage of hourly outliers for each time series doubles to around 9%, suggesting that the occurrence

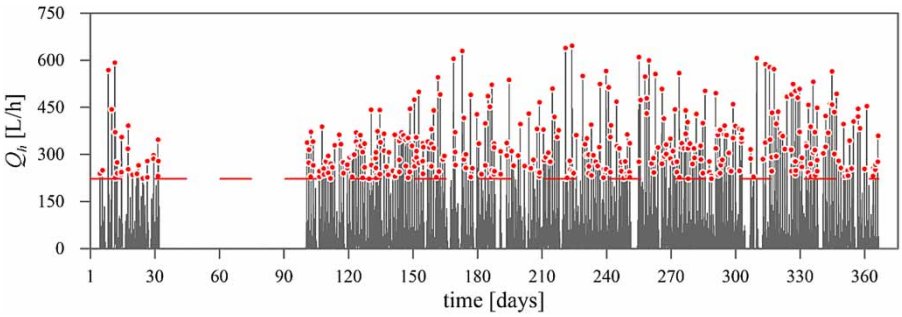

Figure 10 | Lowest-quality time series after outlier detection: valid hourly data (grey line), upper limit for hourly outliers (dashed red line) and hourly outliers (red circles). Please refer to the online version of this paper to see this figure in colour: http://dx.doi.org/10.2166/hydro.2020.133.

of daily outliers cannot be imputed to either the presence of a small number of very high hourly outliers or to a large number of high, but regular, hourly data, but possibly to a combination of the two causes.

Finally, Figure 10 shows the worst residual condition, namely the retained time series with the highest number (about 4%) of hourly outliers once the cascading outlier detection procedure has been performed; this series is affected by a variety of inconsistencies in terms of gaps, zero data and anomalous values, but in a magnitude that does not hinder any kind of quantitative analysis.

CONCLUSIONS

In the present paper, a procedure is proposed aiming to control the overall quality of time series of water consumption and solving possible reliability issues concerning both the time series as a whole and single data. The procedure is tested on the consumption database of the DMA of Soccavo (Naples, Italy), where hourly volumes of consumed water were measured by 2808 flow meters, referring to residential single-family households and connected to a telemetry system, covering the period January 1st, 2016 to December 31st, 2016. The analysis is structured in three consecutive steps, consisting of: (1) a preliminary visual inspection of the consumption map, (2) anomaly detection and (3) outlier detection.

The procedure has a general validity and can be potentially applied to datasets collecting other variables than water consumption, as long as the database to be analysed is homogeneous in terms of data typology; the added value lies in a number of issues:

- It works as a guideline driving the users through consecutive stages of data understanding, starting from a comprehensive, bird's-eye perspective of the whole database and progressively zooming to the interpretation of a single value.
- The approach is nonparametric and unsupervised, so it does not rely on any assumption about prior probability distribution of data at any spatial or temporal aggregation level.
- The proposed metrics for the use of rejection thresholds at every stage are flexible, meaning that they can be tailored to fit the specific purpose of the analysis. For instance, if the main goal is to investigate annual consumption patterns, only the top-quality classes could be analysed, or alternatively d_O^* could be lowered in order to retain only outlier-free time series. On the other hand, if the focus is on hourly consumption, removing daily outliers or even time series with too many daily outliers may not be required.

As shown in the paper, the proposed procedure only targets inconsistencies producing gaps, null values (keeping in mind that zero values can be the realization of a non-consumption event, and do not necessarily refer to malfunctioning) and isolated outliers. It is not aimed at detecting inconsistencies producing anomalous patterns, such as anomalous baselines, changes in the mean or periods of constant consumption, that can stem from either behavioural or technological issues. To solve these issues, nonparametric approaches can be adopted as well,

such as pattern detection procedures relying on statistics- or machine learning-based clustering techniques.

REFERENCES

Avni, N., Fishbain, B. & Shamir, U. 2015 Water consumption patterns as a basis for water demand modeling. *Water Resour. Res.* **51**, 8165–8181.

Barnett, V. & Lewis, T. 1994 *Outliers in Statistical Data*. Wiley Series in Probability and Statistics. Wiley & Sons, New York.

Bay, S. D. & Schwabacher, M. 2003 Mining distance-based outliers in near linear time with randomization and a simple pruning rule. In: *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Washington, DC, USA, pp. 29–38.

Ben-Gal, I. 2005 Outlier detection. In: *Data Mining and Knowledge Discovery Handbook* (O. Maimon & L. Rokach, eds). Springer, Boston, MA, pp. 131–146.

Braca, G., Bussettini, M., Lastoria, B. & Mariani, S. 2013 *Linee guida per l'analisi e l'elaborazione statistica di base delle serie storiche di dati idrologici (Guidelines for the analysis and basic statistical processing of hydrological time series)*. ISPRA Manuals and Guidelines, 84.

Bragalli, C., Neri, M. & Toth, E. 2019 Effectiveness of smart meter-based urban water loss assessment in a real time network with synchronous and incomplete readings. *Environ. Model. Softw.* **112**, 128–142.

Brentan, B. M., Meirelles, G. L., Manzi, D. & Luvizotto, E. 2018 Water demand time series generation for distribution network modeling and water demand forecasting. *Urban Water J.* **15** (2), 150–158.

Breunig, M. M., Kriegel, H. P., Ng, R. T. & Sander, J. 2000 LOF: Identifying density-based local outliers. *ACM SIGMOD Record* **29** (2), 93–104.

Buchberger, S. G. & Nadimpalli, G. 2004 Leak estimation in water distribution systems by statistical analysis of flow readings. *J. Water Resour. Plan. Manage.* **130** (4), 321–329.

Cheifetz, N., Noumir, Z., Samé, A., Sandraz, A., Féliers, C. & Heim, V. 2017 Modeling and clustering water demand patterns from real-world smart meter data. *Drinking Water* **2** (10), 75–82.

Cominola, A., Giuliani, M., Castelletti, A., Rosenberg, D. E. & Abdallah, A. M. 2018 Implications of data sampling resolution on water use simulation, end-use disaggregation, and demand management. *Environ. Model. Softw.* **102**, 199–212.

Cousineau, D. & Chartier, S. 2010 Outliers detection and treatment: a review. *Int. J. Psychol. Res.* **3** (1), 58–67.

Fawcett, T. & Provost, F. 1997 Adaptive fraud detection. *Data Min. Knowl. Discov.* **1** (3), 291–316.

Ferreira, A. M. S., Cavalcante, C. A. M. T., Fontes, C. H. O. & Marambio, J. E. S. 2013 A new method for pattern recognition in load profiles to support decision-making in the management of the electric sector. *Electr. Power Energy Syst.* **53**, 821–831.

Firat, M., Turan, M. E. & Yurdusev, M. A. 2010 Comparative analysis of neural network techniques for predicting water consumption time series. *J. Hydrol.* **384** (1–2), 46–51.

Froukh, M. L. 2001 Decision-support system for domestic water demand forecasting and management. *Water Resour. Manage.* **15** (6), 363–382.

Gargano, R., Tricarico, C., Del Giudice, G. & Granata, F. 2016 A stochastic model for daily residential water demand. *Water Sci. Technol.: Water Supply* **16** (6), 1753–1767.

Hawkins, S., He, H. X., Williams, G. J. & Baxter, R. A. 2002 Outlier detection using replicator neural networks. In: *Proceedings of the 5th International Conference and Data Warehousing and Knowledge Discovery*, Aix-en-Provence, France.

House-Peters, L. A. & Chang, H. 2011 Urban water demand modeling: review of concepts, methods, and organizing principles. *Water Resour. Res.* **47** (5), W05401.

Jain, A., Varshney, A. K. & Joshi, U. C. 2001 Short-term water demand forecasting modelling at IIT Kanpur using artificial neural networks. *Water Resour. Manage.* **15** (5), 299–321.

Jin, W., Tung, A. & Han, J. 2001 Mining top-n local outliers in large databases. In: *Proceedings of the 7th International Conference on Knowledge Discovery and Data-Mining*, San Francisco, CA.

Johnson, R. & Wichern, D. W. 1992 *Applied Multivariate Statistical Analysis*. Prentice Hall, Englewood Cliffs, NJ.

Knorr, E. M. & Ng, R. T. 1997 A unified approach for mining outliers. In: *Proc. Knowl. Discov. KDD*, pp. 219–222.

Knorr, E. M., Ng, R. T. & Tucakov, V. 2000 Distance-based outliers: algorithms and applications. *Int. J. Very Large Databases* **8** (3–4), 237–253.

Knorr, E. M., Ng, R. T. & Zamar, R. H. 2001 Robust space transformations for distance-based operations. In: *Proceedings of the 7th International Conference on Knowledge Discovery and Data-Mining*, San Francisco, pp. 126–135.

Leys, C., Ley, C., Klein, O., Bernard, P. & Licata, L. 2013 Detecting outliers: do not use standard deviation around the mean, use absolute deviation around the median. *J. Exp. Soc. Psychol.* **49** (4), 764–766.

López, J. J., Aguado, J. A., Martín, F., Munoz, F., Rodríguez, A. & Ruiz, J. E. 2011 Hopfield-K-Means clustering algorithm: a proposal for the segmentation of electricity customers. *Electr. Power Syst. Res.* **81**, 716–724.

Loureiro, D., Amado, C., Martins, A., Vitorino, D., Mamade, A. & Teixeira Coelho, S. 2016a Water distribution systems flow monitoring and anomalous event detection: a practical approach. *Urban Water J.* **13** (3), 242–252.

Loureiro, D., Mamade, A., Cabral, M., Amado, C. & Covas, D. 2016b A comprehensive approach for spatial and temporal water demand profiling to improve management in network areas. *Water Resour. Manage.* **30**, 3443–3457.

Macedo, M. N. Q., Galo, J. J. M., deAlmeida, L. A. L. & de C. Lima, A. C. 2015 Demand side management using artificial

neural networks in a smart grid environment. *Renew. Sustain. Energy Rev.* **41**, 128–133.

Mazzolani, G., Berardi, L., Laucelli, D., Martino, R., Simone, A. & Giustolisi, O. 2016 A methodology to estimate leakages in water distribution networks based on inlet flow data analysis. *Proc. Eng.* **162**, 411–418.

McKenna, S. A., Fusco, F. & Eck, B. J. 2014 Water demand pattern classification from smart meter data. *Proc. Eng.* **70**, 1121–1130.

Miller, J. 1991 Reaction time analysis with outlier exclusion: bias varies with sample size. *Q. J. Exp. Psychol.* **43** (4), 907–912.

Mouchel, W. & Schonlau, M. 1998 A fast computer intrusion detection algorithm based on hypothesis testing of command transition probabilities. In: *Proceedings of the 4th International Conference on Knowledge Discovery and Data-Mining*, pp. 189–193.

Padulano, R. & Del Giudice, G. 2018 A mixed strategy based on self-organizing map for water demand pattern profiling of large-size smart water grid data. *Water Resour. Manage.* **32**, 3671–3685.

Padulano, R. & Del Giudice, G. 2019 Pattern detection and scaling laws of daily water demand by SOM: an application to the WDN of Naples, Italy. *Water Resour. Manage.* **33**, 739–755.

Padulano, R., Del Giudice, G., Giugni, M., Fontana, N. & Sorgenti degli Uberti, G. 2018 Identification of annual water demand patterns in the City of Naples. *Multidiscip. Dig. Publ. Inst. Proc.* **2** (11), 587–595.

Papadimitriou, S., Kitawaga, H., Gibbons, P. G. & Faloutsos, C. 2003 LOCI: Fast outlier detection using the local correlation integral. In: *Proceedings of the 19th International Conference on Data Engineering* (Cat. No. 03CH37405), pp. 315–326.

Räsänen, T., Voukantsis, D., Niska, H., Karatzas, K. & Kolehmainen, M. 2010 Data-based method for creating electricity use load profiles using large amount of customer-specific hourly measured electricity use data. *Appl. Energy* **87**, 3538–3545.

Rousseeuw, P. J. & Croux, C. 1993 Alternatives to the median absolute deviation. *J. Am. Stat. Assoc.* **88** (424), 1273–1283.

Rustum, R. & Adeloye, A. J. 2007 Replacing outliers and missing values from activated sludge data using Kohonen self-organizing map. *J. Environ. Eng.* **133** (1), 909–916.

Schwertman, N. C., Owens, M. A. & Adnan, R. 2004 A simple more general boxplot method for identifying outliers. *Comput. Stat. Data Anal.* **47**, 165–174.

Shekhar, S., Lu, C.-T. & Zhang, P. 2002 Detecting graph-based spatial outliers. *Intell. Data Anal.* **36** (5), 451–468.

Williams, G. J. & Huang, Z. 1997 Mining the knowledge mine. In: *Australian Joint Conference on Artificial Intelligence*, Springer, Berlin, Heidelberg, pp. 340–348.

Wright, D. 2009 Profiling the European citizen: cross-disciplinary perspectives. *Info* **11** (1), 96–98.

Zhou, K., Yang, S. & Shen, C. 2013 A review of electric load classification in smart grid environment. *Renew. Sustain. Energy Rev.* **24**, 103–110.

First received 12 July 2019; accepted in revised form 24 January 2020. Available online 6 March 2020