AGU ADVANCING EARTH AND SPACE SCIENCE

Water Resources Research

RESEARCH ARTICLE
10.1029/2022WR032393

Flowrate Time Series Processing in Engineering Tools for Water Distribution Networks

Bruno Ferreira[1], Nelson Carriço[1], Raquel Barreira[1], Tiago Dias[1], and Dídia Covas[2]

[1]INCITE, Barreiro School of Technology, Polytechnic Institute of Setúbal, Setúbal, Portugal, [2]CERIS, Instituto Superior Técnico, Universidade de Lisboa, Lisbon, Portugal

Key Points:

- Processing flowrate time series is necessary before any use in engineering tools for water distribution networks
- A comprehensive methodology is developed for the processing of unevenly (and evenly) spaced flowrate time series
- Tests are carried out with three real case studies with distinct characteristics to demonstrate the application of the proposed methodology

Correspondence to:

B. Ferreira,
bruno.s.ferreira@estbarreiro.ips.pt

Citation:

Ferreira, B., Carriço, N., Barreira, R., Dias, T., & Covas, D. (2022). Flowrate time series processing in engineering tools for water distribution networks. *Water Resources Research*, *58*, e2022WR032393. https://doi.org/10.1029/2022WR032393

Received 21 MAR 2022
Accepted 16 MAY 2022

Abstract The current paper presents a comprehensive methodology for processing unevenly (and evenly) spaced flowrate time series for subsequent use in engineering tools, such as the calibration of hydraulic models or the detection and location of leaks and bursts. The methodology is a four-step procedure: (a) anomaly identification and removal, (b) short-duration gap reconstruction, (c) time step normalization, and (d) long-duration gap reconstruction. The time step normalization is carried out by a numerical procedure prior to the reconstruction process. This reconstruction process uses a pattern model coupled with regression techniques (i.e., autoregressive integrated moving average and exponential smoothing). The methodology is calibrated using Monte Carlo simulations applied to a water utility flowrate time series and validated with two additional time series from different water utilities. Obtained results demonstrate that the proposed methodology can process flowrate time series from water supply systems with different characteristics (e.g., consumption pattern, data acquisition system, transmission settings) both for normal operating conditions and during the occurrence of abnormal events (e.g., pipe bursts). This methodology is a very useful tool for the daily management of water utilities, preparing the time series to be used in different engineering tools, namely, hydraulic simulation, model calibration or online burst detection.

1. Introduction

Many water distribution networks (WDNs) are continuously monitored through installed sensors that measure hydraulic parameters (e.g., pressure, flowrate, users' consumption) and water quality parameters (e.g., chlorine concentration, pH, temperature) (Kara et al., 2016). This continuous monitoring allows to collect raw data to be used in multiple engineering applications, being flowrate and pressure data the most widely used time series by water utilities in different engineering applications, such as: the calculation of water balances (Meseguer & Quevedo, 2017); the development and calibration of hydraulic models in terms of nodal demands and pipe roughness coefficients (Do et al., 2016; Zhang et al., 2018; Zhou et al., 2018); the application of burst detection and location techniques by inverse analysis (Blocher et al., 2020; Moasheri & Jalili-Ghazizadeh, 2020; Sophocleous et al., 2019), by using classifier approaches (Capelo et al., 2021; Fereidooni et al., 2021; Hu et al., 2021), or by using transient-based techniques (Capponi et al., 2017; Covas & Ramos, 2010; Covas et al., 2004; Duan, 2017). Fiorillo et al. (2020) presented a methodology to reconstruct the total demand of a district metered area starting from a small number of monitored users, directly using records of water consumption by users. Cominola et al. (2019), Creaco et al. (2021), and Kossieris et al. (2019) focused on using smart metering technologies to minimize water losses and energy consumption also using users' measurements. The success of most engineering tools strongly depends on the existence of well-processed, reliable, and synchronized time series for flowrate and pressure.

Traditionally, flowrate data are collected and stored at a regular pre-defined time step, usually between 5 and 15 min (Barrela et al., 2017; Chen & Boccelli, 2018; Huang et al., 2020). However, some sensors may not collect and transmit data at a regular time step, which is the case of impulse flowrate meters that register data when a fixed water volume passes in the meter (e.g., one pulse per cubic meter) (Boyle et al., 2013; Clifford et al., 2018). In these cases, unevenly spaced flowrate time series are obtained, with an irregular interval between measurements. Most engineering tools for WDNs require evenly spaced time series. Thus, a time step normalization is usually required depending on the acquisition and transmission settings.

Regardless of the acquisition and transmission settings, the obtained raw flowrate time series may contain measurement errors, such as missing, repetitive or even false readings (Mounce et al., 2010; Xenochristou et al., 2020).

These errors can have their origin in sensor or logger malfunctioning, faulty transmission system due to battery failure, inadequate acquisition range (e.g., above or below meter range or bidirectional flow) and data storage limitations (Loureiro et al., 2016; Machell et al., 2014; Xu et al., 2020). These measurement errors, typically known as outliers or anomalous values (Kirstein et al., 2019), should be detected and corrected before they can be used in engineering applications (e.g., hydraulic modeling, calibration, leak detection) (Romano et al., 2014). Gaps of non-validated measurements must be reconstructed with estimated data, being different methods used to this purpose (Ascensão et al., 2021). Note that flowrate time series usually present daily and weekly patterns which must be considered by the used reconstruction techniques (Barrela et al., 2017; Kirstein et al., 2019; Quevedo et al., 2010).

Processing raw flowrate time series (prior to any use by engineering tools) is essential to guarantee reliable data that do not compromise the success of the multiple engineering applications. Data processing aims at identifying and removing outliers, normalizing the time step and data filling gaps with estimated values. Often, these time series processing is manually carried out, ultimately limiting the amount of data that can be simultaneously treated (Quevedo et al., 2017).

Several techniques have been developed for processing flowrate time series in WDNs. Quevedo et al. (2010) compared the raw flowrate time series (assumed to be evenly spaced) with estimations obtained by using a daily model based on autoregressive integrated moving average (ARIMA) and a second model based on distributing the daily flowrate using a 10-min demand pattern. Flowrate values were considered validated when the difference between the estimated values and the raw values are lower than a given threshold; otherwise, these raw values were removed and replaced with estimations. This methodology was further improved by Cugueró-Escofet et al. (2016) and Quevedo et al. (2016, 2017) with a combination of "low-level" and "high-level" tests. The former checked elementary signal properties and the latter relied on the use of models to check the consistency of the sensor data. Note that, during the occurrence of an abnormal event (e.g., pipe burst), the sensors in the WDN may collect inconsistent values that may not be reflected in the model estimations. Loureiro et al. (2016) compared the raw flowrate time series (assumed to be unevenly spaced) with predefined sensor minimum and maximum threshold values; raw flowrate values below or above these threshold values were not validated and, therefore, were removed; a time step normalization process was then carried, and the gaps (with up to 1 hr of duration) were filled using linear interpolation. In a different approach, Kirstein et al. (2019) defined a set of tests to be carried in the raw flowrate time series (assumed to be evenly-spaced) to ensure that outliers due to the transmission and acquisition errors were removed from the raw flowrate time series. The existing literature focuses mostly on the processing of evenly-spaced flowrate time series, being the used techniques often inadequate to process unevenly spaced time series, such as, data acquired by some sensors (e.g., impulse meters). Also, most existing techniques rely on estimations to validate the raw flowrate time series. This may limit the ability to correctly process time series when abrupt changes in consumption behavior occur (e.g., in touristic coastal areas or lockdown measures due to COVID-19).

The current paper proposes a novel comprehensive methodology for processing unevenly (or evenly) spaced flowrate time series for future use in engineering tools, such as those in the calibration of hydraulic models, in the detection and location of leaks and bursts, and by data-mining techniques using smart meter data. The methodology includes four main steps, namely, the automatic identification of anomalous values, time series reconstruction in short duration gaps, the time step normalization, and time series reconstruction in long duration gaps. The first step uses six tests to automatically categorize the main anomalies in the time series. In this way, the abnormal behavior associated with pipe bursts or with other unexpected events can be preserved in the processed time series. This step generates gaps of non-validated data that can be divided into short and long-duration gaps. The second step is the reconstruction of short-duration gaps, being carried out prior to the time-step normalization which is the third step. The fourth and last step is the reconstruction of long duration gaps using the technique presented in Quevedo et al. (2010). This technique accounts for the daily and weekly cycles of flowrate time series, being able to correctly reconstruct gaps during and weekends and failing when reconstructing holidays that occur in weekdays (Ascensão et al., 2021). This paper presents improvements to be carried to the original technique when reconstructing holidays using a pattern model coupled with simple exponential smoothing and by using a subset of past holidays and Sundays. The proposed flowrate time series processing approach is parametrized through the application to a real case study representative of most Portuguese water utilities monitoring systems and consumption patterns. Two other data series collected in systems with distinct characteristics (e.g.,

Table 1
Descriptive Statistics of 3 Days of Raw Flowrate Time Series for the Three Case Studies

	CS1		CS2		CS3	
	Spacing (s)	Flowrate (m^3/hr)	Spacing (s)	Flowrate (m^3/hr)	Spacing (s)	Flowrate (m^3/hr)
Average	24.4	43.4	315.2	22.9	73.6	186.4
P25	9.0	33.8	300.0	14.4	62.0	134.7
P50	16.0	43.6	300.1	24.0	66.0	172.7
P75	29.0	54.7	300.1	30.0	75.0	230.8
IQR	20.0	20.9	0.1	15.6	13	96.1

consumption pattern, data acquisition equipment, transmission processes) are used to validate the developed methodology. Finally, a comparison of the reconstruction of a national holiday using both the original and the improved technique is carried out.

The proposed methodology can be used in the context of water resources conservation, for instance, in early warning systems against failures or WDN modeling for leakage management actions, including identification of anomalies due to new pipe burst, amongst other uses. The main novel contributions are: (a) the development of a comprehensive methodology for the processing of unevenly (and evenly) spaced flowrate time series (b) the development of tests for the automatic categorization of the main anomalies in unevenly spaced flowrate time series, including a procedure for the calibration of required parameters; and (c) the improvement of the time series reconstruction method, originally developed by Quevedo et al. (2010), in order to account for holidays occurring in weekdays. The overall methodology is encoded in Python programming language into a software tool that will be made available to researchers and to water utilities dealing with observed time series of flowrate.

2. Case Study Description

Three distinct real case studies are considered in this study. The first case study, CS1, is a WDN located in Lisbon metropolitan area, supplying a population of about 3,300 inhabitants. The area is quite homogeneous in terms of the typology of buildings and types of demands, mostly composed of single-family dwellings and some residential buildings with two floors, as well as a few shops and an elementary school. The WDN has an approximate length of 48 km with about 1,800 service connections. It is supplied by one storage tank and has a pumping station located at the upstream end. An impulse flowrate meter is installed at the inlet of the network area (specifically downstream of the pumping station). The network characteristics, monitoring sensors and transmission system are similar to those used by most Portuguese supply systems, being representative of the national reality.

The second case study, CS2, is a small network sector located in the countryside of Portugal, in the southern region, that supplies around 110 inhabitants and an airport that is currently out of operation. The network is supplied by one storage tank with a flowrate meter which acquires the inlet flowrate with a fixed acquisition time step of 5 min.

The third case study, CS3, is a WDN located in a touristic area of Portugal southernmost region with high demand seasonality. The residents are about 3,000 inhabitants, which characterize consumption in the winter season, whereas the summer population reaches 14,000 inhabitants. The supplied area is mainly composed of villas with large and irrigated gardens. The network is directly supplied by one storage tank connected to a pumping station. An impulse flowrate meter is installed downstream of the pumping station.

The three case studies present seasonal variability, with the average daily flowrate greatly varying between winter and summer seasons. The average daily flowrate in winter is around 20, 15, and 70 m^3/hr for CS1, CS2, and CS3, respectively. Higher values of 65, 40, and 200 m^3/hr can be found in the summer for the three case studies.

The CS1 is used for methodology parameterization (as described in Section 4 with results presented in Section 5.1). To this end, a raw flowrate time series with a month's duration (related to September of 2018) is considered. It contains a total of 96,408 flowrate measurements, with the average spacing between measurements and flowrate equal to, respectively, 27 s and 54 m^3/hr.

The three distinct case studies are used to test and to validate the proposed methodology (results in Section 5.2). Three days (Friday, Saturday and Sunday) of raw flowrate data are considered for each case study. These specific days of the week were selected, since they usually present very distinct consumption patterns. For CS1, CS2, and CS3, these periods relate to, respectively, 1–3 June 2018, 5–7 April 2019, and 1–3 September 2017. The descriptive statistics of flowrate values and of spacing between records for each three-day time series are presented in Table 1, highlighting the main differences in both data acquisition characteristics and magnitude of the consumption of the three case studies. CS1 presents the highest sampling rate, with 10,633 records during the 3 days,

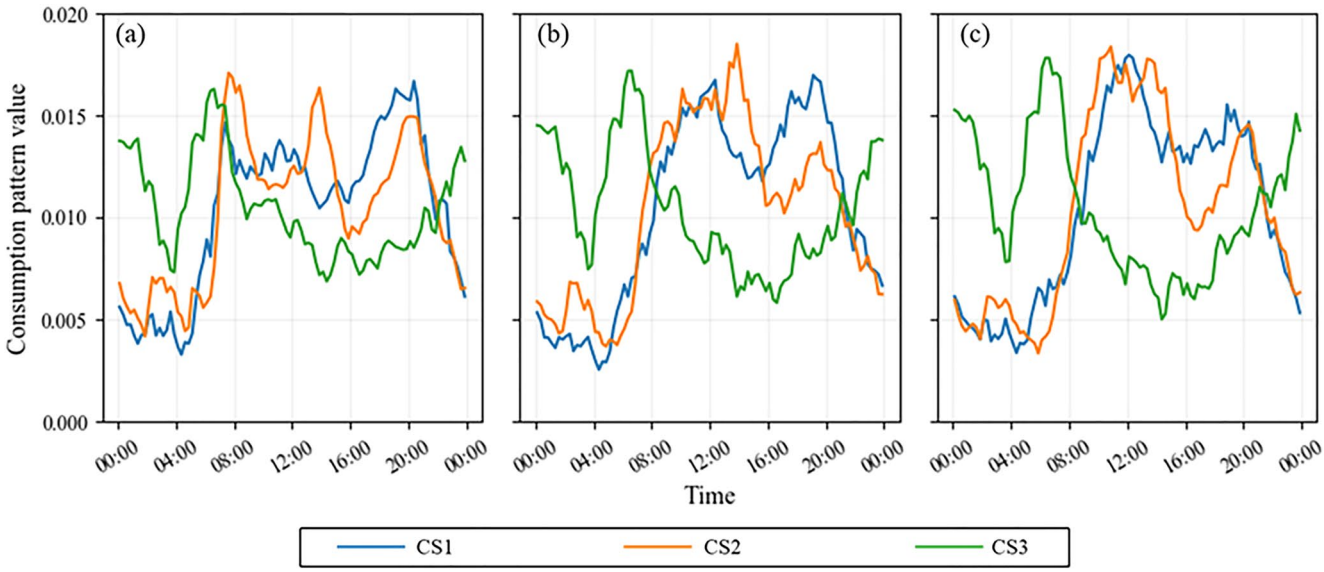

Figure 1. Consumption patterns for: (a) weekdays; (b) Saturdays; (c) Sundays.

resulting in an average interval between measurement of 24.4 s and an interquartile range (IQR) of 20 s (the highest variability in sampling rate). CS2 has the lowest sampling rate, with 822 records with an average interval of 315.2 s and IQR of just 0.1 s (the lowest variability in sampling rate). CS3 presents 3,520 records, with an average interval of 73.6 s and an IQR of 13 s. Regarding consumption magnitude, CS3 has the highest average flowrate, with 186.4 m^3/hr, as well as the highest variability, with an IQR of 96.1 m^3/hr. CS3 has the highest population in addition to the few golf courses being supplied. CS2, with the smallest population of the three case studies, presents the lowest average flowrate, with 22.9 m^3/hr, as well as the lowest variability in flowrate value with an IQR of 15.6 m^3/hr.

Figure 1 presents the consumption patterns of weekdays, Saturdays, and Sundays for the three case studies. These patterns are calculated by considering the already processed flowrate time series of the same sensor with a duration of a month and a time step of 15 min, resulting in 96 pattern values for each type of day (weekdays, Saturdays, and Sundays). These already processed flowrate time series relate to May 2018, March 2019, and August 2017 for CS1, CS2, and CS3, respectively. Both CS1 and CS2 have similar consumption patterns, typical of residential areas, characterized by a minimum night flow and consumption increase during the day. CS3 directly supplies numerous irrigation system from the small villas, being the consumption pattern strongly constrained by the irrigation time. Most consumption occurs during the night and the minimum flowrate occurs in the afternoon around 16:00.

The two most distinct case studies (CS2 and CS3) are used in the comparison of reconstruction methods to estimate the flowrate time series of a holiday that occurred during a weekday (results in Section 5.3). The consumption in the Portuguese national holiday of 25th of April of 2018 and 2017, that occurred in a weekday, is estimated, respectively, for CS2 and CS3 using both the original reconstruction method introduced by Quevedo et al. (2010) and the improved reconstruction method proposed in Section 3.4. One month of already processed flowrate data of the same sensor is used for each case study (from 24 March to 24 April).

3. Methodology for Flowrate Time Series Processing

3.1. The General Approach

The proposed methodology comprises four main steps that should be sequentially applied to the raw flowrate time series (Figure 2): (a) Anomaly identification and removal; (b) Short-duration gap reconstruction; (c) Time step normalization; (d) Long-duration gap reconstruction.

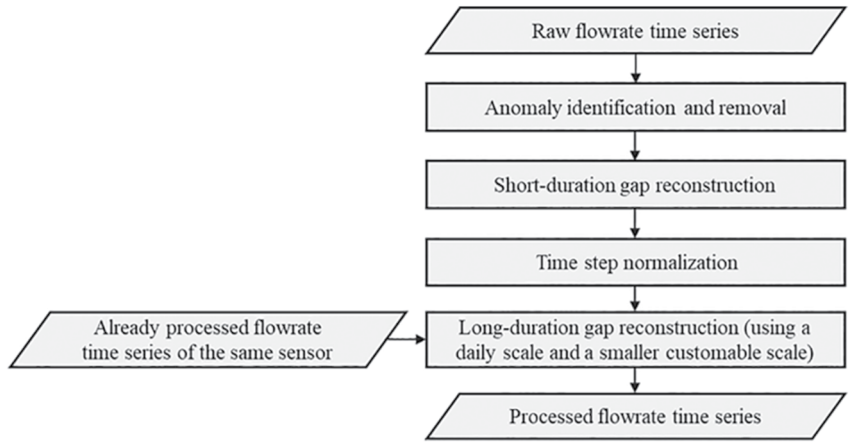

Figure 2. Flowchart of the proposed methodology for flowrate time series processing.

The application of this methodology requires the definition of specific parameter values which are used at different steps. Table 2 presents the description of these parameters, including the analyzed values and the best value found after the calibration procedure (see Section 4 for the parameter calibration procedure and Section 5.1 for parameter calibration results). Next sections describe in detail each step, as well as the specific use of each parameter.

3.2. Anomaly Identification and Removal

Most common anomalies found in raw flowrate time series are associated with measurement duplication, negative values, abnormally low or high values, periods of abnormally low variation in flowrate and long periods without measurements, caused by acquisition, transmission, and storage problems. Some examples of these anomalies are depicted in Figure 3, with a generic flowrate time series of residential areas, characterized by a minimum night flow and an increased consumption during the day, including a peak in consumption during lunch time, in the afternoon and at dinner time. Each flowrate measurement is related to the average flowrate over that period.

Table 2
Description of the Methodology Parameters, Including the Analyzed Values and the Best Value Found After the Calibration Procedure

Test/procedure	Parameter	Unit	Analyzed values
Abnormally high values	Maximum duration, *p1*	Second	1, 3[a], and 5 times the 50th percentile of the time step intervals
	Maximum acceptable flowrate rate of change, *p2*	Flowrate units[c] per second	80th, 90th, and 97th[a] percentile of the rate of change for successive flowrate records
Abnormally low values	Maximum duration, *p3*	Second	1, 3[a], and 5 times the 50th percentile of the time step intervals
	Maximum acceptable flowrate rate of change, *p4*	Flowrate units[c] per second	80th, 90th, and 97th[a] percentile of the rate of change for successive flowrate records
Flat lines	Minimum duration of flat line, *p5*	Second	Max (600 s, 2.5 time the 50th percentile of the time step intervals)
	Maximum acceptable flowrate variation, *p6*	Flowrate units[c] per second	0.5%–9.5% of the standard deviation of the flowrate time series[b]
Long period without measurements	Minimum duration of a long period without measurements, *p7*	Second	Equal to the desired time step after normalization
Short and long-duration gaps	Maximum duration of a short-duration gap, *p8*		

[a]Best values found by the calibration procedure. [b]Best found values of 2.5%, 3%, and 3.5% of the standard deviation of the flowrate time series. [c]Different flowrate units can be considered, for instance, L/s, m^3/h, m^3/s, amongst others.

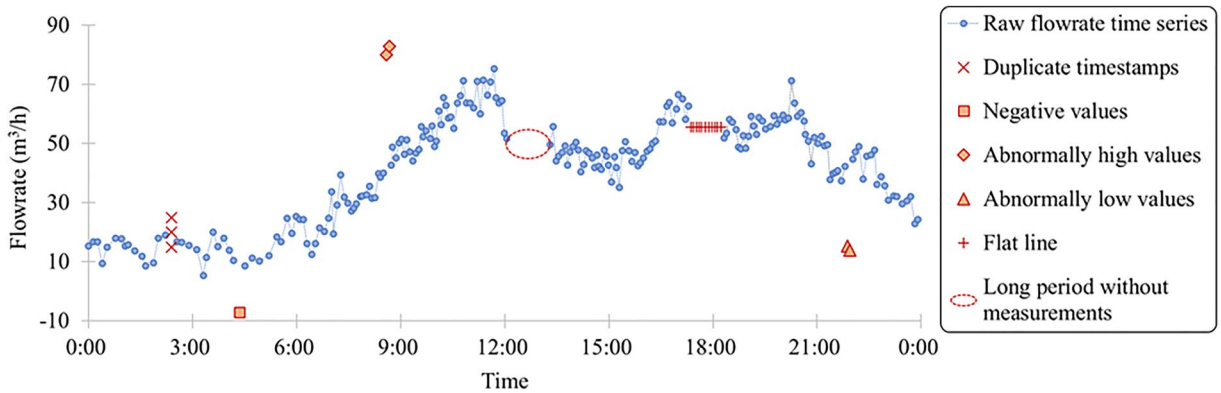

Figure 3. Example of identified anomalies in raw flowrate time series.

Each of these anomalies is identified using a specific test, which should be carried out sequentially for the raw flowrate data set. A total of six tests are presented herein. In each test, every raw flowrate measurement is either classified as non-validated or validated, varying if the measurement is, or is not, part of the anomaly being tested. Measurements that are not validated have their values set as null. The different types of anomalies and their automatic identification tests are presented in the following paragraphs.

The raw flowrate time series may have *duplicate timestamps* that must be identified. The problem is that duplicate timestamps can have their flowrate values either equal or different. Equal values can be solved by eliminating the duplicated measurements. Different values can be solved by eliminating all duplicated timestamps except one, whilst not validating its value (i.e., setting it as null).

The data set should not have *negative values*, since the flowrate sensors are installed at the inlet of a networks sector measuring only positive values. Nonetheless, due to reverse flow occurring during water hammer, negative values can be generated. These negative values do not reflect the network actual behavior (in terms of water consumption) and should be removed so that they do not affect the performance of engineering applications (e.g., demand forecasting techniques). Therefore, every measurement of the raw flowrate time series with a negative value is not validated. This test does not apply to data collected in network sections that may have reverse flow, in which negative flowrate values can occur and negative values represent flow occurring in the opposite way. The processing of such time series is out of the scope of this paper.

An *abnormally high value* (also known as peak or spike) is described as an abrupt increase of the flowrate value followed by a sudden sharp value drop, occurring during a limited period. The identification and removal of these measurements is of upmost importance, as they can affect the performance of the later engineering applications (for instance, they can potentially lead to false positive alarms in burst detection techniques). A novel test is developed and proposed herein to detect abnormally high values in the raw flowrate time series by assessing the rate of change for successive flowrate measurements. The rate of change is the ratio between the difference of two subsequent measurements and their respective differences in time. For each measurement, a time window is created covering all rates of change within a defined time interval p_1. An abnormally high value exists, in the time window p_1, when a positive rate of change is followed by a negative rate of change, both exceeding a predefined threshold p_2. Both measurements associated to the flagged rate of change are not validated, as well as those in between. A specific procedure for the calibration of parameters p_1 and p_2 is described in Section 4.2. The application and results are presented in Section 5.1; the best values found by the calibration procedure (or calibrated values) for parameters p_1 and p_2 were 3 times the 50th percentile of the time step intervals and the 97th percentile of the rate of change for successive raw flowrate measurements, respectively.

Contrarily, an *abnormally low value* (also known as a negative peak) is described as a sharp decrease in flowrate value followed by an abrupt increase, during a certain time interval p_3. These abnormally low values equally compromise the success of many applications for leak and burst detection. The automatic identification test is similar to the one carried for positive peaks, being the main difference searching for a negative rate of change followed by a positive one, both exceeding a predefined threshold p_4. A procedure for the calibration of parameters p_3 and p_4 is described in Section 4.2. The application and results are presented in Section 5.1; the best values

found by the calibration procedure for parameters p_3 and p_4 were 3 times the 50th percentile of the time step intervals and the 97th percentile of the rate of change for successive raw flowrate measurements, respectively.

A *flat line* is usually deemed as consecutives measurements with identical values (Kirstein et al., 2019), yet small variations among consecutive measurements could also be classified as a flat line. Such variations are linked to specific features of the raw flowrate series (e.g., acquisition method and time step, flowrate range or sensor sensitivity) and may greatly vary between utilities. So, a *flat line* is herein defined as a sequence of measurements, covering more than a given time interval p_5, whose values are within a certain range. This range is defined by the value of the first measurement of the sequence, plus or minus a given threshold p_6. Measurements that lay inside this range are not validated and have their value set as null. The value for parameter p_5 is proposed as the maximum between 5 min and 2.5 times the 50th percentile of time step intervals. This rule considers that at least three points should be required to form a flat line. Nonetheless, in flowrate time series with a small time step (e.g., 30 s), this could result in flat lines of short duration (e.g., 90 s). As such, the 5 min value aims at establishing a minimum absolute duration of a flat line. A procedure for the calibration of parameter p_6 is described in Section 4.3 and results presented in Section 5.1. The best (or calibrated) values were found equal to 2.5%, 3%, and 3.5% of the standard deviation of the raw flowrate time series.

A *long period without measurements* is defined as a gap between two measurements that exceeds a given duration p_7. Such periods should be detected and filled with estimated data. However, the exact timestamp of the next measurement in unevenly spaced time series is difficult to predict, since it depends on the sensor characteristics and the transmission system settings. A threshold p_7 is considered for successive time step intervals in the raw flowrate time series above which the period is flagged. The minimum duration of a long period without measurements, p_7, is considered equal to the desired time step after normalization (e.g., 15 min). These flagged periods are later reconstructed using the long-duration gap reconstruction technique, as described in Section 3.4.

3.3. Short-Duration Gap Reconstruction

As a result of the anomaly identification process, the flowrate time series contain gaps of non-validated measurements that must be reconstructed with estimated data. These gaps can be divided into short and long-duration gaps. The former is treated at this stage, whereas the latter requires prior time step normalization.

A short-duration gap can be defined as a single, or a set of, non-validated measurements, whose gap between the surrounding "left" and "right" validated measurements is shorter than a given threshold p_8. The maximum duration of a short-duration gap, p_8, is considered equal to the desired time step after normalization (e.g., 15 min). Different and relatively simple interpolation methods can be used to reconstruct short duration gaps of unevenly spaced flowrate, namely, Nearest-Neighbor, Linear or Polynomial Interpolation (Lepot et al., 2017). In the Nearest-Neighbor, the value of the closest validated measurements is assigned to the missing value. Although easy to use, this method often produces the worst results, as flat lines are inevitably generated. In linear interpolation, the estimated values are assumed to lie on the line joining the nearest validated measurements to the "left" and "right" of the short-duration gap. As such, estimated values are bounded between the "left" and "right" validated values. Polynomial interpolation could be achieved by fitting a polynomial of the lowest possible degree that passes through the points of the data set. Nonetheless, they may estimate values outside of the observed range of data. As such, the recommendation is to use linear interpolation, since it yields the best compromise between easiness-to-use and adequate results for the reconstruction of short-duration gaps.

3.4. Time Step Normalization

Most engineering computational applications require a fixed acquisition time step for the flowrate time series. Additionally, if the gap exceeds the threshold p_8 defined in Section 3.3 (and thus becoming a long-duration gap), simple reconstruction techniques might not be suitable, and more complex techniques are required. These reconstruction techniques usually require a fixed acquisition time step. As such, the normalization to a regular time step is carried out between both long-duration gaps and long periods without measurements. The normalization to a regular time step can be performed by using a trapezoidal rule of numerical integration to compute the mean flowrate within the desired time step (e.g., 1 hr, 15 min, 5 min) (Loureiro et al., 2016).

3.5. Long-Duration Gap Reconstruction

Once the time step normalization process is carried out, a technique based on Quevedo et al. (2010) methodology is used for the reconstruction of the remaining missing values associated with long-duration gaps. The reconstruction process is carried out once for each day with at least one non-validated record. Note that the duration of the raw flowrate time series being processed can be variable, with just a few hours up to a week or a complete month.

The basic idea behind the reconstruction technique is to work in two different timescales, namely, a daily scale and a smaller customizable scale (e.g., 1 hr, 15 min, 5 min). In each scale, a specific model is constructed. Already processed flow rate time series of the same sensor is used in both timescales.

On the daily scale, an ARIMA model is used to predict the total volume of the day being reconstructed. For this purpose, an already processed flow rate time series of the same sensor is required. The total volume of each day of the already processed flow rate time series of the same sensor, y, should be calculated. The value y_{k-1} is the total volume for 1 day before the day being reconstructed. The prediction for the total volume of the day being reconstructed, $\hat{y}_k$, is derived from three main components:

1. A 1-week-period oscillating polynomial to account for cyclic deterministic behavior:

$$\hat{y}_k = 2\cos\left(\frac{2\pi}{7}\right) y_{k-1} - y_{k-2} \tag{1}$$

2. An integrator that takes into account possible trends:

$$\hat{y}_k = y_{k-1} \tag{2}$$

3. An autoregressive component that accounts for the influence of previous values within a week:

$$\hat{y}_k = -a_1 y_{k-1} - a_2 y_{k-2} - a_3 y_{k-3} - a_4 y_{k-4} \tag{3}$$

in which a_1, a_2, a_3, and a_4 are the autoregressive model parameters.

The structure of the total volume of the day being reconstructed, $\hat{y}_k$, is obtained by combining the three components (see Quevedo et al., 2010 for further details), leading to the following equation:

$$\hat{y}_k = -\sum_{i=1}^{7} [b_i y_{k-i}] \tag{4}$$

in which b_i are auxiliary model functions defined as follows:

$$\begin{aligned}
b_1 &= a_1 - \left[2\cos\left(\frac{2\pi}{7}\right) + 1\right], \\
b_2 &= a_2 - \left[2\cos\left(\frac{2\pi}{7}\right) + 1\right] a_1 + \left[2\cos\left(\frac{2\pi}{7}\right) + 1\right], \\
b_3 &= a_3 - \left[2\cos\left(\frac{2\pi}{7}\right) + 1\right] a_2 + \left[2\cos\left(\frac{2\pi}{7}\right) + 1\right] a_1 - 1, \\
b_4 &= a_4 - \left[2\cos\left(\frac{2\pi}{7}\right) + 1\right] a_3 + \left[2\cos\left(\frac{2\pi}{7}\right) + 1\right] a_2 - a_1, \\
b_5 &= -\left[2\cos\left(\frac{2\pi}{7}\right) + 1\right] a_4 + \left[2\cos\left(\frac{2\pi}{7}\right) + 1\right] a_3 - a_2, \\
b_6 &= \left[2\cos\left(\frac{2\pi}{7}\right) + 1\right] a_4 - a_3, \\
b_7 &= -a_4.
\end{aligned}$$

The autoregressive model parameters a_1, a_2, a_3, and a_4 can be estimated using the Least Squares Method by minimizing the root mean squared error (RMSE) of the residuals between the total daily volumes of already processed flow rate time series and the corresponding predictions.

Once an estimation of the total volume of the day being reconstructed has been obtained, it can be distributed along the customizable scale (e.g., 1 hr, 15 min, 5 min) using a pattern model. Different consumption patterns should be used for weekdays, Saturdays, and Sundays/holidays to attend the weekly consumption variation

(Saludes et al., 2017). Each consumption pattern is calculated by considering the average values in each time step for all the days that belong to that type of consumption pattern in the already processed flow rate time series of the same sensor. In the end, each pattern value is divided by the total daily average volume for all the days that belong to that type of consumption pattern (i.e., weekdays, Saturdays, and Sundays/holidays). The duration of the already processed flow rate time series of the same sensor is variable and should be selected depending on the characteristics of the supplied area (i.e., rural, urban, touristic) and on the month being studied. For instance, the water consumption in rural and urban areas may be affected by seasonal and weather variability, which can only be observed during the course of a few months. On the other hand, touristic coastal areas can have abrupt changes in water consumption (and in measured flowrate data) during the course of 1 or 2 weeks. Furthermore, lockdown measures due to COVID-19 can heavily affect the consumption patterns from one week to another. So, a reference duration of 1 month for the already processed flow rate time series of the same sensor is recommended and used herein. This duration typically accounts for 22 Weekdays, 4 Saturdays, and 4 Sundays.

Estimations for the customizable scale can be achieved by multiplying the estimated total daily volume (previously obtained in the daily model) by each pattern value (of the corresponding type of pattern). When using this technique, the day being reconstructed is completely estimated at the customizable scale. Finally, the missing values can be reconstructed by assigning the corresponding estimations.

Nonetheless, problems may arise when attempting to use the ARIMA model to predict the total volume of a holiday occurring during a weekday (Ascensão et al., 2021). As such, improvements to the original method are carried in the daily scale model to account for this possibility. The objective is to estimate the total daily volume of the day being reconstructed (which is a holiday) based solely on past Sundays and holidays. A data set composed of the dates of holidays is required. When initializing the daily model, a verification is carried out to assess if the day being reconstructed is or not a holiday. If it is not a holiday, the ARIMA model runs as previously presented; otherwise, the total daily volume is estimated with a simple exponential smoothing model. The input is a subset of the total daily volumes of past Sundays and holidays, s, representing the value s_{k-2} the total volume for two Sundays or holidays before the holiday being reconstructed. The structure of the total volume of the holiday being reconstructed, $\hat{s}_k$, can be defined as follows:

$$\hat{s}_k = \alpha s_{k-1} + \alpha(1-\alpha)s_{k-2} + \alpha(1-\alpha)^2 s_{k-3} + \cdots \tag{5}$$

in which α is the smoothing parameter that controls the relative importance of past observations compared to more recent observations. This parameter can be estimated using the Least Squares Method by minimizing the RMSE of the residuals between the total daily volumes of the already processed flow rate time series (past Sundays and holidays) and the corresponding predictions.

4. Parameter Calibration Procedure

4.1. General Overview of the Calibration Procedure

From the six types of anomalies identified in Section 3.2, duplicate and negative values are straightforwardly identified and do not require parameter calibration. The remaining anomaly identification tests described in Section 3.2 require the preliminary calibration of parameters, namely, the abnormally high values (duration p_1 and rate of change p_2), the abnormally low values (duration p_3 and rate of change p_4) and the flat lines (duration p_5 and flowrate p_6). This section presents a procedure for the assessment and calibration of different parameter values in order to obtain a robust and reliable set of parameters for each test. As a result of the calibration procedure, a set of calibrated values (See Table 1) is obtained for the distinct parameters. Note that these calibrated values are in fact statistical properties that should be calculated for the raw flowrate timeseries being processed (for instance, the 97th percentile of the rate of change for successive raw flowrate records for parameter p_2).

4.2. Abnormally High and Low Values

The proposed test for identifying abnormally high values classifies each raw flowrate measurement (point) as validated or non-validated. For this test, a validated measurement means that the flow rate value is not abnormally high (absence of anomaly). On the other hand, a non-validated measurement means that the measurement is

abnormally high (abnormally high flowrate value). After classification of the raw flowrate time series using this test, each measurement falls into one of the four categories:

1. True Positive (TP): Non-validated measurement (abnormally high flowrate value) classified as non-validated (abnormally high flowrate value).
2. False Negative (FN): Non-validated measurement (abnormally high flowrate value) classified as validated (absence of anomaly).
3. True Negative (TN): Validated measurement (absence of anomaly) classified as validated (absence of anomaly).
4. False Positive (FP): Validated measurement (absence of anomaly) classified as non-validated (abnormally high flowrate value).

This test contains two specific parameters, namely, the maximum duration p_1 (s) and the maximum acceptable flowrates rate of change p_2 (flowrate unit per second) (see Section 3.2 and Table 1). The objective of this test is to correctly classify all measurements as validated and non-validated, that is, to reduce both FNs and positives (FN and FP). However, multiple combinations of parameters can be generated for this test by varying the p_1 and p_2 parameters. These combinations of parameters produce different results when applied to the same raw flowrate time series. Different class-specific metrics are used to assess the performance of distinct combinations of parameters, namely, the Recall, the Precision, and the F-measure.

The *Recall*, also known as sensitivity or TP rate, represents the fraction of the non-validated measurement (abnormally high flowrate value) correctly identified, is described as follows:

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \tag{6}$$

where True Positive refers to the number of measurements classified as TP and False Negative to the number of measurements classified as FN.

Selecting a combination of parameters that maximizes the Recall guarantees that the number of non-validated measurements (abnormally high flowrate value) classified as validated (absence of anomaly) is minimized (i.e., FN is minimized). However, parameters solely selected based on the Recall may lead to a misclassification of validated measurement (absence of anomaly) as non-validated (abnormally high flowrate value), increasing the number of FP.

The Precision, also called positive predictive value, shows the fraction of the non-validated measurement (abnormally high flowrate value) correctly identified among all measurements classified as non-validated measurement (abnormally high flowrate value), is calculated as follows:

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \tag{7}$$

where False Positive refers to the number of measurements classified as FP.

Selecting a combination of parameters that maximizes the Precision guarantees that the number of validated measurements (absence of anomaly) classified as non-validated (abnormally high flowrate value) is minimized (i.e., FP is minimized). Nonetheless, parameters selected solely based on the Precision may lead to a misclassification of non-validated measurement (abnormally high flowrate value) as validated (absence of anomaly), ultimately increasing the number of FN.

Both Recall and Precision metrics are complementary to each other in the analysis, having an inverse relationship (i.e., an increase in Recall leads to a decrease in Precision), and, as referred, should be both maximized.

The *F*-measure combines Precision and Recall metrics in a single parameter. Thus, it is used herein to assess the performance of each combination of parameters. This measure corresponds to the harmonic mean of the Precision and Recall and can be defined as follows (Chinchor, 1992):

$$F = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{8}$$

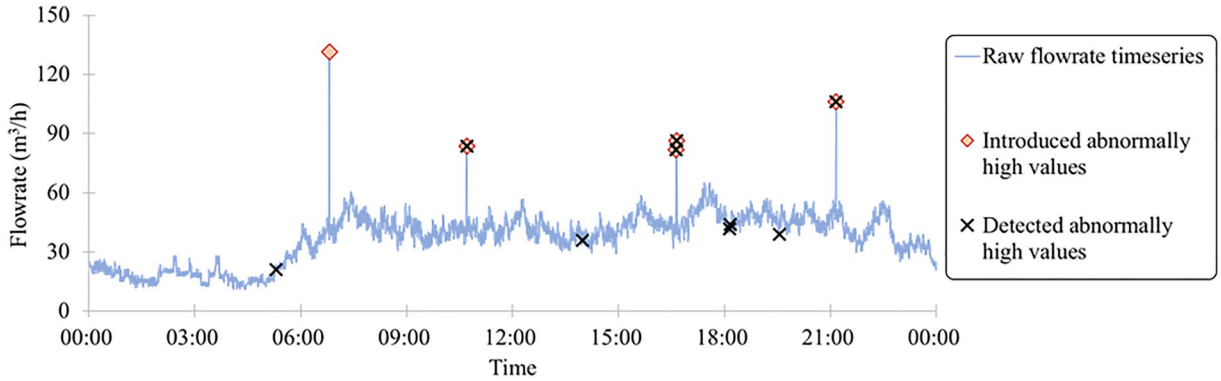

Figure 4. Example of a Monte Carlo simulation by introducing abnormally high values.

The maximum value of the *F*-measure is 1, which corresponds to both Recall and Precision being equal to 1, that is, no false negatives (FN) and false positives (FP) are identified, which is attained by the ideal (perfect) classification algorithm.

Abnormally high flowrate values may occur randomly and sparsely through time. As such, the p_1 and p_2 parameters must be calibrated taking into account this randomness. To this end, a procedure based on Monte Carlo (MC) simulations is proposed to determine the best combination of parameters p_1 and p_2 (out of multiple combinations of parameters) and by using the *F*-measure (Chinchor, 1992). For each possible combination of parameters, a set of 10,000 MC simulations is run. The following process is carried out in each MC simulation to take into account the randomness of occurrence of abnormally high values in raw flowrate time series: (a) a raw flowrate time series with a month's duration is considered, and data from one random day are considered; (b) five abnormally high values are randomly generated in the previously selected day by selecting five measurements (with up to two consecutive ones) and by artificially incrementing their values by a factor between two and four (see Figure 4); (c) Abnormally high values are identified using the proposed test with the combination of parameters being assessed and the *F-measure* is calculated. At the end of the 10,000 MC simulations, the overall performance of the combination of parameters is assessed by computing the average *F-measure*. The process is repeated for the remaining combinations of parameters. The best combination of parameters is selected as the one that presents the highest average *F*-measure within the 10,000 simulations, ultimately calibrating this test specific parameters p_1 and p_2.

As referred in Section 3.2, the identification test for abnormally low values uses two specific parameters, namely, the maximum duration p_3 (s) and the maximum acceptable rate of change p_4 (flowrate units per second). A similar calibration process based on 10,000 MC simulations per combination of parameters can be carried to parameters p_3 and p_4, with the difference that abnormally low values are introduced in the time series by artificially reducing the real value with a factor ranging from two to four.

4.3. Flat Lines

The value of parameter p_6 must be carefully selected since it has a direct effect on the time series processing, specifically on the number of validated and non-validated measurements. An unreasonably high value will lead to an excessive number of periods being considered as flat lines. Conversely, an extremely low value can lead to no period being classified in the flat lines category. Ultimately, the true percentage of points in flat lines is unknown and, hence, the optimal value of p_6 is subjective. An extensive sensitivity analysis of different values for the parameter p_6, according to the percentage of points detected as flat lines, is proposed. It is recommended that the value for the p_6 parameter is proportional to the standard deviation of the raw flowrate values in order to account for utility flowrate range and variability (e.g., between 1% and 5% of the standard deviation of the raw flowrate values). Finally, the value parameter p_6 should correspond to an adequate percentage of points detected as flat lines (e.g., 0.5%).

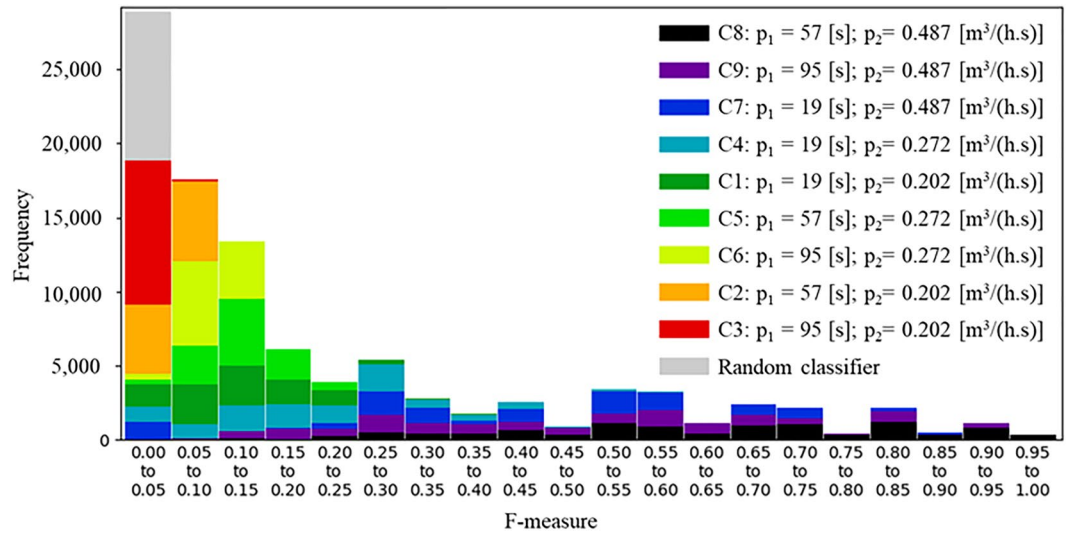

Figure 5. Distribution of *F*-measure for 10,000 MC simulations for each of the nine combinations of parameters p_1 and p_2.

5. Application to Case Studies and Discussion

5.1. Parameter Calibration

The procedure for parameter calibration described in Section 4 is applied to the case study 1 (CS1). A raw flowrate time series with a month duration (September of 2018) is considered.

Concerning abnormally high values, three values are analyzed for the maximum duration p_1, namely, 1, 3, and 5 times the 50th percentile of the time step intervals of the raw flowrate time series. These values consider that, on average, 1, 3, and 5 points are required to form an abnormally high value. Similarly, three values are analyzed for the maximum acceptable flowrates rate of change p_2, namely, the 80th, 90th, and 97th percentile of the rate of change for successive raw flowrate measurements. These values consider that 20%, 10%, and 3% of flowrate measurements are possibly abnormally high values, at least, from an excessive rate of change point of view. Both the 50th percentile of the time step intervals of the raw flowrate time series and the 80th, 90th, and 97th percentile of the rate of change for successive raw flowrate measurements are calculated for the complete month of September 2018 of the raw flowrate time series. This leads to a total of nine combinations of parameters for this test.

Figure 5 shows the distribution of the *F*-measure across the 10,000 simulations for all the nine combinations plus a random classifier. Parameter p_2 (the maximum acceptable rate of change) has a more direct impact on the classification results when compared to the p_1 parameter (the maximum duration). Note that by considering a higher value for p_2 (see combinations C8, C9, and C7 with p_2 = 97th percentile of the rate of change), it is possible to obtain higher *F*-measures regardless of the values for the p_1 parameter. Thus, a higher value for the p_2 parameter is preferable (in this case, the 97th percentile of the rate of change was the higher tested value).

Table 3 presents the overall results for the three best combinations (C8, C9, and C7) obtained in the 10,000 MC simulations. Parameter p_2 is equal in the three best combinations, being the obtained differences due to the different values for the p_1 parameter. Combination C8 has the highest average *F*-measure of 0.62 by considering p_1 equal to 3 times the 50th percentile of the time-step intervals. Combination C9 leads to the highest number

Table 3
Results of the Calibration of p_1 and p_2 Parameters for the Three Best Combinations

Best combination	p_1	p_1 (s)	p_2	p_2 (m³/[h.s])	Average *F*-measure	TP	TN	FP	FN
C8	3 × P50th of time step intervals	57	P97th of rate of change	0.487	0.62	39,081	32,035,048	49,256	10,914
C9	5 × P50th of time step intervals	95	P97th of rate of change	0.487	0.47	45,099	31,936,897	152,882	4,899
C7	1 × P50th of time step intervals	19	P97th of rate of change	0.487	0.42	18,189	32,048,829	14,622	31,804

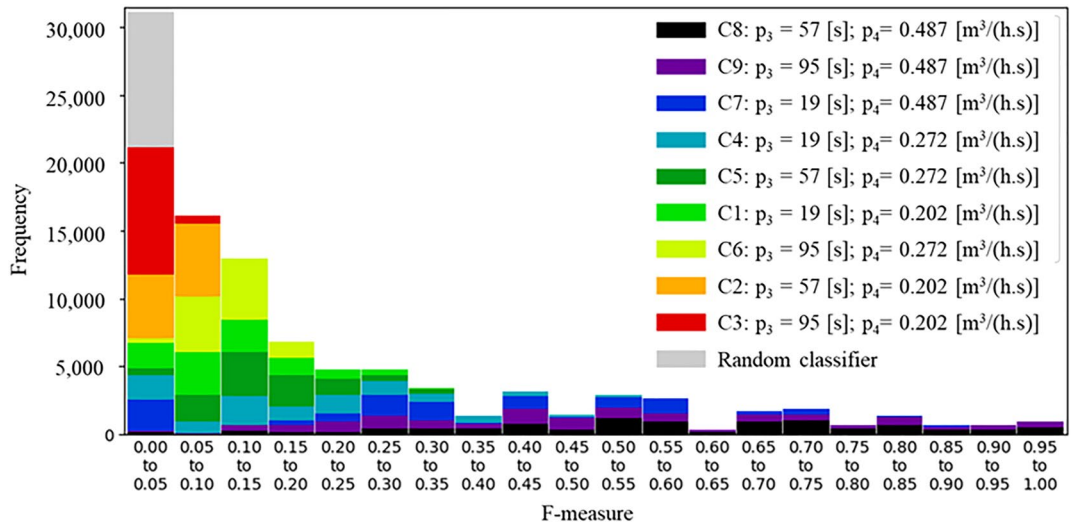

Figure 6. Distribution of *F*-measure for 10,000 MC simulations for each of the nine combination of parameters p_3 and p_4.

of measurements correctly identified as abnormally high values, with a TP with 45,099, and by considering a wider p_1 parameter of 5 times the 50th percentile of the time-step intervals. This is achieved with the cost of increasing the number of measurements misclassified as abnormally high, increasing the FP to a total of 152,882 measurements and ultimately decreasing the average *F*-measure to 0.47. Combination C7 considers a narrower p_1 parameter of 50th percentile of the time-step intervals. This combination leads to both the lowest number of FP and TP (of the three best combinations), resulting in the average *F*-measure 0.42.

The identification test for abnormally low values uses two specific parameters, namely, the maximum duration, p_3, and the maximum acceptable rate of change, p_4. The same three values previously used for p_1 are considered herein for p_3, namely, 1, 3, and 5 times the 50th percentile of the time step intervals of the raw flowrate time series. Similarly, the three values that are analyzed for p_2, namely, the 80th, 90th, and 97th percentile of the rate of change for successive raw flowrate measurements, are used for p_4. This leads to nine combinations of parameters for this second test.

Figure 6 shows the distribution of the *F*-measure across the 10,000 simulations for all the nine combinations plus a random classifier. Higher *F*-measure values can be obtained by considering a higher value for the p_4 parameter (C8, C9, and C7 with p_4 = 97th percentile of the rate of change), regardless of the values for the p_3 parameter. Thus, a higher value for p_4 is preferable (in this case, the 97th percentile of the rate of change was the higher tested value).

Table 4 presents the overall results for the three best combinations (C8, C9, and C7) obtained in the 10,000 MC simulations. Combination C8 presents the highest average *F*-measure of 0.58 by considering p_3 equal to 3 times the 50th percentile of the time step intervals, followed by C9 and C7 with an *F*-measure of 0.49 and 0.32, respectively. Note that the highest p_3 value of 5 times the 50th percentile of the time step intervals of combination C9 leads to both the highest number of *TP* and of *FP*, whilst the lowest p_3 value of 1 times the 50th percentile leads to both the lowest number of *TP* and of *FP*.

Table 4
Results of the Calibration of p_3 and p_4 Parameters for the Three Best Combinations

Best combination	p_3	p_3 (s)	p_4	p_4 (m³/[h.s])	Average *F*-measure	TP	TN	FP	FN
C8	3 × P50th of time step intervals	57	P97th of rate of change	0.487	0.58	34,823	32,021,007	51,178	15,167
C9	5 × P50th of time step intervals	95	P97th of rate of change	0.487	0.49	40,422	31,959,392	120,234	9,571
C7	1 × P50th of time step intervals	19	P97th of rate of change	0.487	0.32	14,033	32,058,891	17,291	35,964

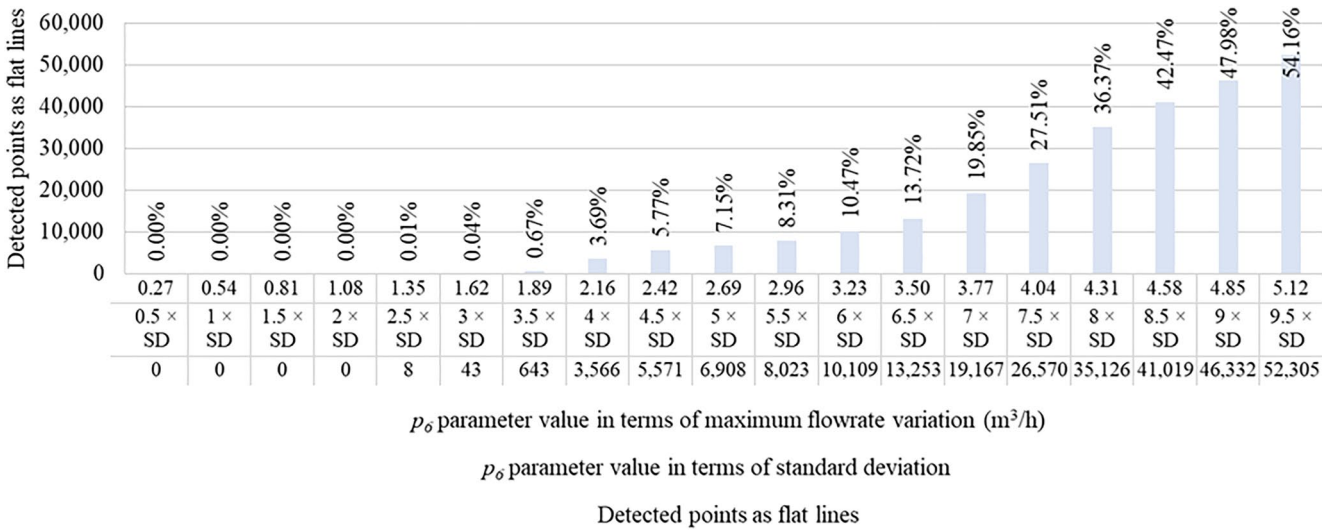

Figure 7. Number and total percentage of detected points as flat lines using different values for the maximum flowrate variation parameter p_6.

The calibration of the maximum flowrate variation in a flat line, p_6, is carried out as described in Section 4.3 by conducting an exhaustive sensitivity analysis to the percentage of points detected as flat lines for different values of p_6. A total of 19 possible values are considered between 0.5% and 9.5% of the SD of the raw flowrate values. For each possible value for the p_6 parameter (out of the 19 possible values), the same month of September 2018 is considered and the flat lines are detected using the proposed technique described in Section 3 (p_5 parameter was considered as previously presented).

Figure 7 depicts the number of detected flat line points for each of the 19 values of p_6. This figure shows that 2.5, 3, and 3.5 times the SD of raw flowrate values yield the best balance between identifying flat lines of small variations (i.e., 1.35, 1.62, and 1.89 m³/hr, respectively) whilst not identifying a significant percentage of the total number of measurements (0.01%, 0.04%, and 0.67%, respectively). A higher percentage of total number of measurements is identified as flat lines for higher values of p_6, with the values of 5 and 9.5 times the SD (2.69 and 5.12 m³/hr, respectively) identifying a total of 6,908 and 52,305 measurements (7.15% and 54.16%, respectively) as part of flat lines.

5.2. Assessment of the Validity of the Proposed Methodology

Results of the application of the proposed methodology with the calibrated set of parameters are presented herein for the three different Portuguese case studies. The three case studies have very distinct characteristics, in terms of consumption pattern, sensor equipment characteristics and acquisition system settings (see Section 2). Three days (Friday, Saturday, and Sunday) of raw flowrate data of each case study, collected at the inlet section of each sector, are used to test and validate the proposed methodology. The anomalous values due to acquisition and transmission problems described in Section 3 are artificially added in each utility 3-day flowrate data, namely, a negative value, five abnormally high and low values (by increasing or decreasing by a factor between 2 and 3), a flat line (of equal values) with a duration of 3 hr and a long period without measurements with a duration of 3 hr Figure 8 shows the flowrate time series (in blue line) with the introduced abnormal values for each case study.

The set of parameters is calculated for each case study considering that the time series should have its time step normalized to 900 s (15 min) after the validation process. This is a common value for flowrate time step to be used in engineering tools. The parameter values are calculated for the raw flowrate time series being processed (i.e., the 3 days) according to the best values found in Section 5.1 (see Table 2). For instance, the parameter p_1 of each case study is calculated as 3 times the 50th percentile of the time step intervals for the raw flowrate time series (i.e., the 3-days period). Table 5 presents parameter values for each case study.

Note that some parameter values for CS1 change when compared to the used values in Section 5.1. For instance, the value for p_1 is 48 s, when 57 s was previously used in Section 5.1 (although 3 times the 50th percentile of time

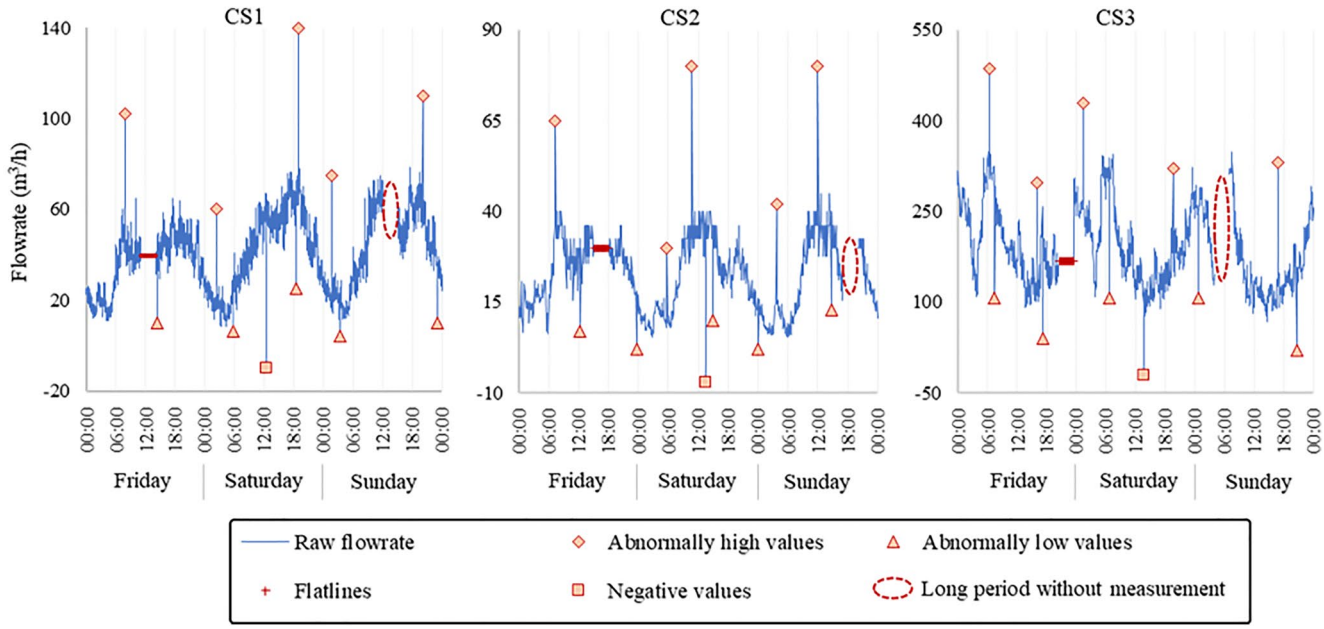

Figure 8. Three days of raw flowrate time series with introduced abnormal values for the three case studies.

step intervals was considered in both situations). This is due to the fact that the raw flowrate timeseries used in Section 5.1 was related to September 2018 (with a particular consumption pattern and level), whilst the 3 days period occurred in June 2018 (with a distinct consumption).

For each case study, an already processed flow rate time series of the same sensor with a duration of a month (and a time step of 15 min) is used during the reconstruction of long-duration gaps by using the method presented in Section 3.4. These already processed flow rate time series are the same as those previously used to calculate the consumption patterns (see Section 2) presented in Figure 1.

Figure 9 presents the processed flowrate time series (in red line) for each case study after using the proposed methodology with the parameter values in Table 5. The raw flowrate time series with introduced abnormal values are depicted in thin light blue lines. The original raw flowrate time series (i.e., without the introduced abnormal values) is depicted in a thin dark blue line. Both blue lines overlap except during periods of flat lines and periods without measurement.

The proposed methodology is capable of detecting the different introduced anomalies (abnormally high and low values, flat lines, negative values, and long periods without measurements) in the three case studies. This demonstrates that the calibrated set of parameters is robust enough to accommodate distinct characteristics of the time series. Thus, and by using the calibrated set of parameters, no input from the user is required for anomaly identification and removal.

Table 5
Calibrated Set of Parameters for Each Case Study

	CS1	CS2	CS3
p_1 (s)	48	903	198
p_2 (m^3/[h.s])	0.512	0.041	0.401
p_3 (s)	48	903	198
p_4 (m^3/[h.s])	0.512	0.041	0.401
p_5 (s)	300	752	300
p_6 (m^3/h)	0.444	0.304	1.913
$p_7 = p_8$ (s)	900	900	900

The time step normalization is also carried out, ensuring that the same flowrate behavior could be described with a much smaller number of measurements (10,633, 822, and 3,520 measurements were reduced to just $96 \times 3 = 288$ measurements). Note that the time step after normalization is customizable using the proposed methodology and its proper definition depends on the final usage for the processed flowrate time series in engineering tools. For instance, water and energy balances and hydraulic model calibration can be performed for an hourly based time-step. On the other hand, leak detection and location techniques might require a smaller time step of 15 min (or even shorter). The study on how the definition of the time step after normalization affects the futures uses of the time series by engineering tools is out of the scope of this study and will be carried out in future works.

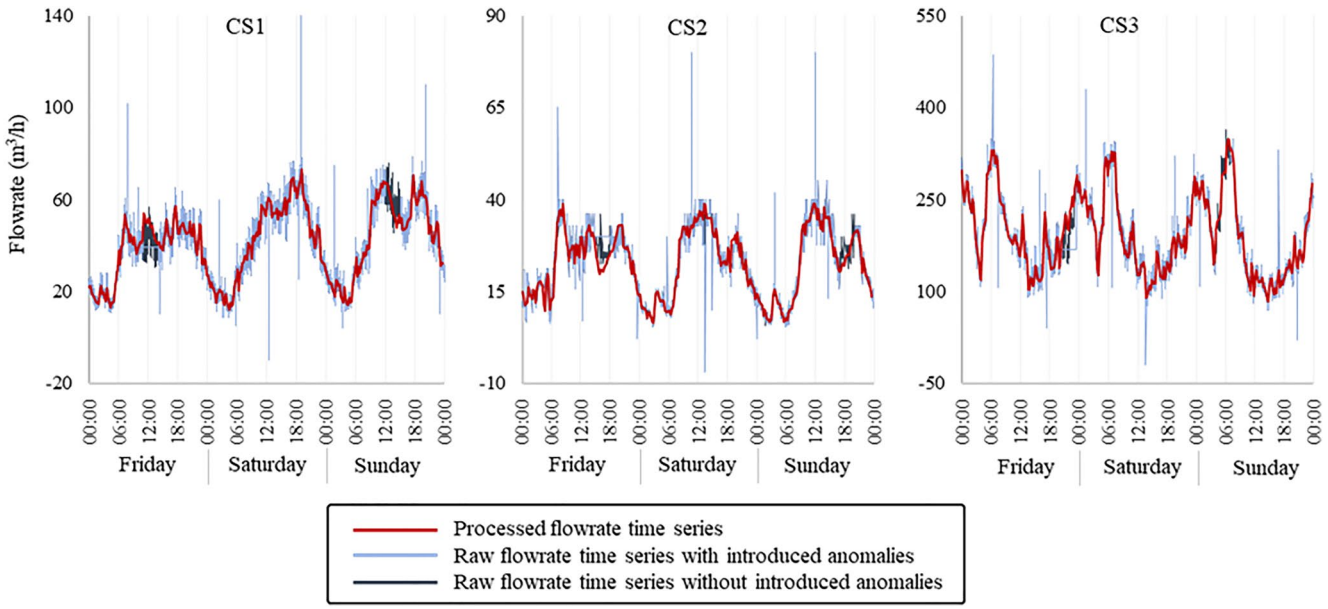

Figure 9. Raw flowrate time series with and without introduced anomalies and processed flowrate time series for each case study.

Lastly, the time series are reconstructed (with emphasis on the long duration gaps) according to the expected consumption patterns (presented in Figure 1). Two major periods are being reconstructed in each time series, namely, a flat line on a Friday and a long period without measurements on a Sunday (see Figure 8). The long periods without measurements are correctly reconstructed for the three case studies (the red line can describe the behavior of the dark blue line in such periods). Minor differences can be seen in the reconstruction of the flat lines occurring on Friday. This may be due to the fact that a pattern model is used and in which Fridays are considered along the remaining Weekday to create the Weekday pattern (which can be seen in Figure 1). The definition of a specific pattern for each Weekday could improve the reconstruction model. This study, as well as how the duration of the already processed flow rate time series of the same sensor affects the reconstruction process are out of the scope of this study and will be carried in future works.

These results demonstrate the validity of the proposed methodology on processing flowrate time series of different water utilities, with distinct characteristics in consumption patterns and magnitude, as well as acquisition and transmission settings. Note that no input from the user was required to produce the red lines in the three case studies, besides the definition of the 15 min time step period after normalization.

Despite the proposed methodology having demonstrated to be successful in the processing of time series related to the normal operational (i.e., no pipe burst), this approach should be also capable of preserving the time series behavior during abnormal events, such as pipe burst events. Thus, the validity of the proposed methodology on processing flowrate time series under abnormal operation is assessed herein. One day of raw flowrate measurements containing a real pipe burst event is considered in each case study.

Figure 10 presents the raw (in thin blue line) and processed flowrate time series (in red line) considering a normalized time step of 15 min. The pipe bursts events are highlighted with a dashed black line: in CS1, a burst occurred between 15:00 and 17:00 (approximate times), whilst in CS2 it occurred between 12:00 and 16:00; in CS3, the pipe burst event occurred between 16:00 and 20:00.

The proposed methodology is used for each case study time series. The processed flowrate time series (in red line) preserves the abnormal behavior associated with the pipe burst event (besides the identification and removal of abnormal values as those previously described in Section 3.2). Furthermore, the time step normalization is carried (to 15 min), as required by most pipe burst detection techniques. These results demonstrate that the proposed methodology can successfully process flowrate time series of different water utilities during the occurrence of abnormal events, without changing the series behavior.

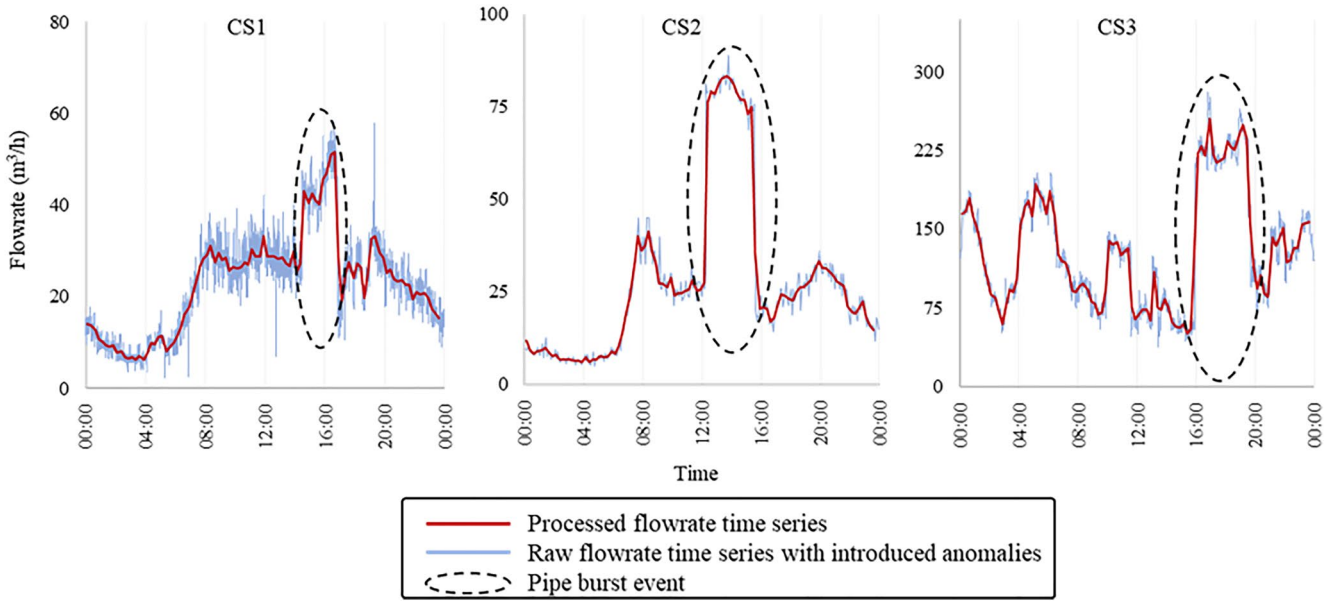

Figure 10. Raw and processed flowrate time series with a pipe burst event for each case study.

5.3. Comparison of Reconstruction Methods

The Portuguese national holiday of the 25th of April (occurring during a Weekday) is reconstructed for the two most different case studies (CS2 and CS3). Two techniques are compared: the original reconstruction method introduced by Quevedo et al. (2010) and the improved reconstruction method proposed in Section 3.4. The main difference between techniques lies in the daily model, specifically on how the total daily volume of a holiday occurring during a weekday is estimated. In the original method, this volume is estimated using an ARIMA model; for instance, the total daily volume of the holiday occurring during a Wednesday would be estimated as a regular Wednesday (i.e., i.e., not a holiday) based on past 7 days. The improved reconstruction technique, on the other hand, uses a simple exponential smoothing model and a subset of the total daily volumes of past Sundays and holidays. The flowrate in this holiday is reconstructed for a time step of 15 min (leading to 96 steps). Figure 11 presents the real (raw) flowrate time series for each case study (in blue line), the estimated values by using the original reconstruction method (in green line) and the improved reconstruction method (in red line).

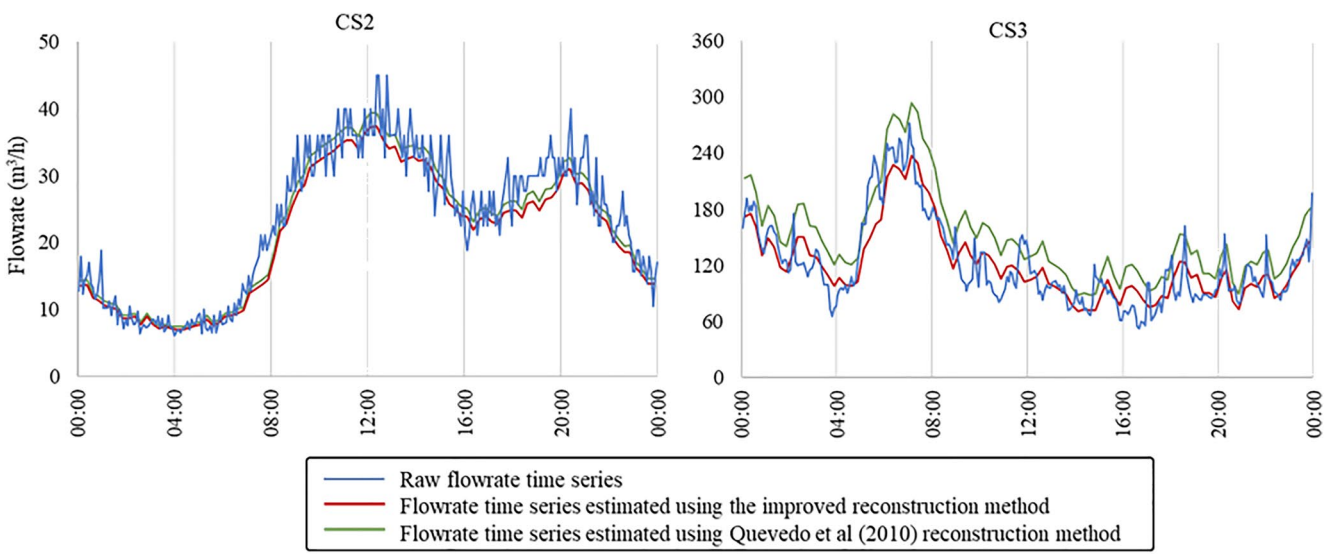

Figure 11. Flowrate variation of national holiday for two case studies (CS2 and CS3).

For CS3, the original method (green line) overestimate the flowrate when compared to the real-time series (in blue line). This is due to the fact that the total daily volume is estimated for a weekday and, for this particular case study, is significantly greater then the total daily volume of a holiday. The improved method, on the other hand, estimates total daily volume as a sequence of Sundays/holidays, thus leading to more accurate results. CS2 presents a small variability between Weekdays and holidays (in terms of total daily volume); both methods capture the pattern of a holiday and correctly estimate the total daily volume of a holiday occurring during a weekday for CS2. On the overall, the original reconstruction method is able to correctly estimate a holiday occurring during a weekday as long as the total daily volume remains relatively constant across the 7 days of the week. If, on the other hand, the total daily volume greatly varies between the weekdays and Sundays (as in CS3), the improved method is recommended.

6. Open-Source Computer Application

The developed methodology has been implemented in an open-source computer application for Windows using Python programming language. The main functionality is the possibility to import a raw flowrate time series to be processed using this approach. Once the time series have been imported (in *CSV* or *TXT* format), the set of parameters is automatically calculated (see Table 2) and suggested, by default, to the user, which has the possibility to accept or to change the value of each parameter (including the desired time step after normalization). Already processed flowrate time series of the same sensor can also be imported to the computer application (in *CSV* or *TXT* format) to be used in the reconstruction step (as described in Section 3.4).

The computer application and its source code are available in a GitHub repository (https://github.com/Ferreira-B/Flowrate-time-series-processing).

7. Conclusions

The current paper proposes a novel and comprehensive methodology for the processing of unevenly and evenly spaced flowrate time series for use in multiple engineering computer applications, namely, for creating early warning systems against failures, for the calibration of hydraulic models or for pipe bursts detection and location. The most common anomalies in flowrate time series are identified by thoroughly assessing data of several Portuguese water utilities. Tests are developed for the automatic identification of the most common anomalies in flowrate time series due to acquisition and transmission problems. The time step normalization is carried by numerical procedures prior to the time series reconstruction using a pattern model coupled with regression techniques (ARIMA and exponential smoothing). An open-source tool with the implemented methodology has been developed and is freely available for the technical and scientific community use (https://github.com/Ferreira-B/Flowrate-time-series-processing).

The methodology requires the definition of specific parameters. The values of these parameters can be obtained by a calibration process based on MC simulations applied to a real case study, whose characteristics are representative of most Portuguese water utilities sensors and data acquisition systems. The overall methodology is demonstrated through application to three different Portuguese case studies, with distinct characteristics both in consumption magnitude and pattern, sensor equipment and acquisition settings. The proposed methodology demonstrated to be capable of processing flowrate time series whilst preserving the behavior of abnormal events, such as a pipe burst event, in the three different case studies.

In future research, the proposed methodology can be automatically combined with pipe burst detection techniques. That is, the flowrate time series processing methodology can be continuously carried with a certain time step (e.g., every 15 min), processing the raw flowrate measurements that occurred during this period and, then, returning validated data to a pipe burst detection and location technique. Distinct flowrate processing methods could be compared to assess how they affect the performance of burst detection and location techniques. Additionally, the methodology can be extended to process other types of time series collected in the water sector, such as pressure data or chlorine concentration data, as well as to data from other types of utilities (electricity or gas).

Data Availability Statement

The datasets and source code are publicly accessible at https://github.com/Ferreira-B/Flowrate-time-series-processing.

Acknowledgments
This research was funded by Fundação para a Ciência e a Tecnologia, through studentship (reference number SFRH/BD/149392/2019) and WISDom project (reference number DSAI-PA/DS/0089/2018).

References

Ascensão, C., Ferreira, B., Barreira, R., & Carriço, N. (2021). Comparison of reconstruction methods for water supply systems flow rate time series. In *Proceedings of the 1st International Conference on Water Energy Food and Sustainability (ICoWEFS 2021)* (pp. 851–858). Springer International Publishing. https://doi.org/10.1007/978-3-030-75315-3_90

Barrela, R., Amado, C., Loureiro, D., & Mamade, A. (2017). Data reconstruction of flow time series in water distribution systems—A new method that accommodates multiple seasonality. *Journal of Hydroinformatics*, *19*(2), 238–250. https://doi.org/10.2166/hydro.2016.192

Blocher, C., Pecci, F., & Stoianov, I. (2020). Localizing leakage hotspots in water distribution networks via the regularization of an inverse problem. *Journal of Hydraulic Engineering*, *146*(4), 04020025. https://doi.org/10.1061/(ASCE)HY.1943-7900.0001721

Boyle, T., Giurco, D., Mukheibir, P., Liu, A., Moy, C., White, S., & Stewart, R. (2013). Intelligent metering for urban water: A review. *Water*, *5*(3), 1052–1081. https://doi.org/10.3390/w5031052

Capelo, M., Brentan, B., Monteiro, L., & Covas, D. (2021). Near–real time burst location and sizing in water distribution systems using artificial neural networks. *Water*, *13*(13), 1–23. https://doi.org/10.3390/w13131841

Capponi, C., Ferrante, M., Zecchin, A. C., & Gong, J. (2017). Leak detection in a branched system by inverse transient analysis with the admittance matrix method. *Water Resources Management*, *31*(13), 4075–4089. https://doi.org/10.1007/s11269-017-1730-6

Chen, J., & Boccelli, D. L. (2018). Forecasting hourly water demands with seasonal autoregressive models for real-time application. *Water Resources Research*, *54*(2), 879–894. https://doi.org/10.1002/2017WR022007

Chinchor, N. (1992). MUC-4 evaluation metrics. In *4th Message Understanding Conference, MUC 1992—Proceedings* (pp. 22–29). https://doi.org/10.3115/1072064.1072067

Clifford, E., Mulligan, S., Comer, J., & Hannon, L. (2018). Flow-signature analysis of water consumption in nonresidential building water networks using high-resolution and medium-resolution smart meter data: Two case studies. *Water Resources Research*, *54*(1), 88–106. https://doi.org/10.1002/2017WR020639

Cominola, A., Nguyen, K., Giuliani, M., Stewart, R. A., Maier, H. R., & Castelletti, A. (2019). Data mining to uncover heterogeneous water use behaviors from smart meter data. *Water Resources Research*, *55*(11), 9315–9333. https://doi.org/10.1029/2019WR024897

Covas, D., & Ramos, H. (2010). Case studies of leak detection and location in water pipe systems by inverse transient analysis. *Journal of Water Resources Planning and Management*, *136*(2), 248–257. https://doi.org/10.1061/(ASCE)0733-9496

Covas, D., Ramos, H., Graham, N., & Maksimovic, C. (2004). Application of hydraulic transients for leak detection in water supply systems. *Water Supply*, *4*(5–6), 365–374. https://doi.org/10.2166/ws.2004.0127

Creaco, E., Galuppini, G., Campisano, A., & Franchini, M. (2021). Bottom-up generation of peak demand scenarios in water distribution networks. *Sustainability*, *13*(1), 1–18. https://doi.org/10.3390/su13010031

Cugueró-Escofet, M., García, D., Quevedo, J., Puig, V., Espin, S., & Roquet, J. (2016). A methodology and a software tool for sensor data validation/reconstruction: Application to the Catalonia regional water network. *Control Engineering Practice*, *49*, 159–172. https://doi.org/10.1016/j.conengprac.2015.11.005

Do, N. C., Simpson, A. R., Deuerlein, J. W., & Piller, O. (2016). Calibration of water demand multipliers in water distribution systems using genetic algorithms. *Journal of Water Resources Planning and Management*, *142*(11), 1–13. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000691

Duan, H.-F. (2017). Transient frequency response based leak detection in water supply pipeline systems with branched and looped junctions. *Journal of Hydroinformatics*, *19*(1), 17–30. https://doi.org/10.2166/hydro.2016.008

Fereidooni, Z., Tahayori, H., & Bahadori-Jahromi, A. (2021). A hybrid model-based method for leak detection in large scale water distribution networks. *Journal of Ambient Intelligence and Humanized Computing*, *12*(2), 1613–1629. https://doi.org/10.1007/s12652-020-02233-2

Fiorillo, D., Galuppini, G., Creaco, E., De Paola, F., & Giugni, M. (2020). Identification of influential user locations for smart meter installation to reconstruct the urban demand pattern. *Journal of Water Resources Planning and Management*, *146*(8), 04020070. https://doi.org/10.1061/(asce)wr.1943-5452.0001269

Hu, X., Han, Y., Yu, B., Geng, Z., & Fan, J. (2021). Novel leakage detection and water loss management of urban water supply network using multiscale neural networks. *Journal of Cleaner Production*, *278*, 123611. https://doi.org/10.1016/j.jclepro.2020.123611

Huang, Y., Zheng, F., Kapelan, Z., Savic, D., Duan, H. F., & Zhang, Q. (2020). Efficient leak localization in water distribution systems using multi-stage optimal valve operations and smart demand metering. *Water Resources Research*, *56*(10), 1–21. https://doi.org/10.1029/2020WR028285

Kara, S., Karadirek, I. E., Muhammetoglu, A., & Muhammetoglu, H. (2016). Real time monitoring and control in water distribution systems for improving operational efficiency. *Desalination and Water Treatment*, *57*(25), 11506–11519. https://doi.org/10.1080/19443994.2015.1069224

Kirstein, J. K., Høgh, K., Rygaard, M., & Borup, M. (2019). A semi-automated approach to validation and error diagnostics of water network data. *Urban Water Journal*, *16*(1), 1–10. https://doi.org/10.1080/1573062X.2019.1611884

Kossieris, P., Tsoukalas, I., Makropoulos, C., & Savic, D. (2019). Simulating marginal and dependence behaviour of water demand processes at any fine time scale. *Water*, *11*(5), 885. https://doi.org/10.3390/w11050885

Lepot, M., Aubin, J. B., & Clemens, F. H. L. R. (2017). Interpolation in time series: An introductive overview of existing methods, their performance criteria and uncertainty assessment. *Water*, *9*(10), 796. https://doi.org/10.3390/w9100796

Loureiro, D., Amado, C., Martins, A., Vitorino, D., Mamade, A., & Coelho, S. T. (2016). Water distribution systems flow monitoring and anomalous event detection: A practical approach. *Urban Water Journal*, *13*(3), 242–252. https://doi.org/10.1080/1573062X.2014.988733

Machell, J., Mounce, S. R., Farley, B., & Boxall, J. B. (2014). Online data processing for proactive UK water distribution network operation. *Drinking Water Engineering and Science*, *7*(1), 23–33. https://doi.org/10.5194/dwes-7-23-2014

Meseguer, J., & Quevedo, J. (2017). Real-time monitoring and control in water systems. In V. Puig, C. Ocampo-Martínez, R. Pérez, G. Cembrano, J. Quevedo, & T. Escobet (Eds.), *Advances in industrial control* (pp. 1–19). Springer International Publishing. https://doi.org/10.1007/978-3-319-50751-4_1

Moasheri, R., & Jalili-Ghazizadeh, M. (2020). Locating of probabilistic leakage areas in water distribution networks by a calibration method using the imperialist competitive algorithm. *Water Resources Management*, *34*(1), 35–49. https://doi.org/10.1007/s11269-019-02388-4

Mounce, S. R., Boxall, J. B., & Machell, J. (2010). Development and verification of an online artificial intelligence system for detection of bursts and other abnormal flows. *Journal of Water Resources Planning and Management*, *136*(3), 309–318. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000030

Quevedo, J., Garcia, D., Puig, V., Saludes, J., Cugueró, M. A., Espin, S., et al. (2017). Sensor data validation and reconstruction. In V. Puig, C. Ocampo-Martínez, R. Pérez, G. Cembrano, J. Quevedo, & T. Escobet (Eds.) *Real-time monitoring and operational control of drinking-water systems* (pp. 175–193). Springer International Publishing. https://doi.org/10.1007/978-3-319-50751-4_10

Quevedo, J., Pascual, J., Espin, S., & Roquet, J. (2016). Data validation and reconstruction for performance enhancement and maintenance of water networks. *IFAC-PapersOnLine*, *49*(28), 203–207. https://doi.org/10.1016/j.ifacol.2016.11.035

Quevedo, J., Puig, V., Cembrano, G., Blanch, J., Aguilar, J., Saporta, D., et al. (2010). Validation and reconstruction of flow meter data in the Barcelona water distribution network. *Control Engineering Practice*, *18*(6), 640–651. https://doi.org/10.1016/j.conengprac.2010.03.003

Romano, M., Kapelan, Z., & Savić, D. A. (2014). Automated detection of pipe bursts and other events in water distribution systems. *Journal of Water Resources Planning and Management*, *140*(4), 457–467. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000339

Saludes, J., Quevedo, J., & Puig, V. (2017). Demand forecasting for real-time operational control. In V. Puig, C. Ocampo-Martínez, R. Pérez, G. Cembrano, J. Quevedo, & T. Escobet (Eds.) *Real-time monitoring and operational control of drinking-water systems* (pp. 99–111). Springer International Publishing. https://doi.org/10.1007/978-3-319-50751-4_6

Sophocleous, S., Savić, D., & Kapelan, Z. (2019). Leak localization in a real water distribution network based on search-space reduction. *Journal of Water Resources Planning and Management*, *145*(7), 04019024. https://doi.org/10.1061/(ASCE)WR.1943-5452.0001079

Xenochristou, M., Hutton, C., Hofman, J., & Kapelan, Z. (2020). Water demand forecasting accuracy and influencing factors at different spatial scales using a gradient boosting machine. *Water Resources Research*, *56*(8), 1–15. https://doi.org/10.1029/2019WR026304

Xu, W., Zhou, X., Xin, K., Boxall, J., Yan, H., & Tao, T. (2020). Disturbance extraction for burst detection in water distribution networks using pressure measurements. *Water Resources Research*, *56*(5), 1–17. https://doi.org/10.1029/2019WR025526

Zhang, Q., Zheng, F., Duan, H.-F., Jia, Y., Zhang, T., & Guo, X. (2018). Efficient numerical approach for simultaneous calibration of pipe roughness coefficients and nodal demands for water distribution systems. *Journal of Water Resources Planning and Management*, *144*(10), 04018063. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000986

Zhou, X., Xu, W., Xin, K., Yan, H., & Tao, T. (2018). Self-adaptive calibration of real-time demand and roughness of water distribution systems. *Water Resources Research*, *54*(8), 5536–5550. https://doi.org/10.1029/2017WR022147