Urban Water Journal

ISSN: 1573-062X (Print) 1744-9006 (Online) Journal homepage: https://www.tandfonline.com/loi/nurw20

A semi-automated approach to validation and error diagnostics of water network data

Jonas Kjeld Kirstein, Klavs Høgh, Martin Rygaard & Morten Borup

To cite this article: Jonas Kjeld Kirstein, Klavs Høgh, Martin Rygaard & Morten Borup (2019) A semi-automated approach to validation and error diagnostics of water network data, Urban Water Journal, 16:1, 1-10, DOI: 10.1080/1573062X.2019.1611884

To link to this article: https://doi.org/10.1080/1573062X.2019.1611884

URBAN WATER JOURNAL
2019, VOL. 16, NO. 1, 1–10
https://doi.org/10.1080/1573062X.2019.1611884

Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE

A semi-automated approach to validation and error diagnostics of water network data

Jonas Kjeld Kirstein[a], Klavs Høgh[b], Martin Rygaard[a] and Morten Borup[a]

[a]Urban Water Systems, Department of Environmental Engineering, Technical University of Denmark, Kgs. Lyngby, Denmark; [b]NIRAS A/S, Allerød, Denmark

ABSTRACT

We propose a method for quality assurance of raw data from water distribution networks in near real-time. Well-known and novel data analysis methods, including a timestamp drift test, are combined to produce a malfunction indicator database for diagnosing anomalies within data acquisition practices. The method was applied to 112 flow and 111 pressure data sets, covering on average 32 months, located throughout the distribution networks of three Danish utilities. Around 10% of measurements in the utilities' meter data sets were absent and 3–35% were categorized as dubious or erroneous. The most common types of anomalies for flow and pressure data were flatline and time stamp inconsistencies. Time drifts were identified in all three utilities and a similarity analysis revealed a simultaneous occurrence of many anomalies. These high rates could have been avoided if the proposed method had been implemented to automatically highlight meter errors and system-wide problems in data collection.

ARTICLE HISTORY
Received 30 April 2018
Accepted 22 April 2019

KEYWORDS
Data validation; error diagnostics; water supply

1. Introduction

Erroneous meter data from water distribution networks can lead to incorrect conclusions during data fusion and data analysis in the water supply sector. Reliable data is a fundamental prerequisite for leakage detection, water quality monitoring, hydraulic modelling and network optimization. To secure a high data integrity it is therefore essential to validate the collected data (Quevedo et al. 2017) and to maintain high data accuracy and data reliability, as promoted by the International Water Association (IWA) (Alegre et al. 2006). This is of growing importance, as utilities are drawing inferences at an increasing rate from data-driven applications and the collected data itself as sensors and data transmission become cheaper. Combining all data sources can be used proactively to improve network operations in a utility (Machell et al. 2014). Andrews et al. (2017) found that it is common for many utilities to struggle with the validation of their own data. However, they also found that, once validated, the data builds the foundation for advanced applications, such as a successful water loss control program. Data validation and anomaly detection methods are not new to urban water engineering. Examples include rainfall data (Jørgensen et al. 1998; Estévez, Gavilán, and Giráldez 2011), urban hydrology (Mourad and Bertrand-Krajewski 2002; Branisavljević, Kapelan, and Prodanović 2011), wastewater treatment (Rosen, Röttorp, and Jeppsson 2003; Puig et al. 2008) and the collection of data on water quality (García et al. 2017) and flow (Quevedo et al. 2010; Cugueró-Escofet et al. 2016) in water distribution networks. Typically, validation and anomaly detection are performed in a combination of 'low-level' methods, based on simple heuristics and limited statistical knowledge, and 'high-level' model-based approaches. High-level approaches use the spatial and temporal redundancy in the available data sets to flag suspicious and erroneous data (Quevedo et al. 2017). Even though data validation does not represent a new research field, it is still not a priority issue at many water supply utilities. Whereas larger Danish water utilities are benchmarked on an annual basis to compare their performance in terms of operational costs, unregistered water losses, etc. (DANVA 2016), no comparison or benchmark exists in their data acquisition and management practices. In the American Water Works Association's Free Water Audit Software (WLCC 2014) utility practices are graded; the software, however, does not focus on anomalies in raw meter data. In this paper, we make a call for a renewed interest in data quality and improved management of errors in logged data.

Defining dubious data points as outliers (in this paper referred to as anomalies) is a subjective concept, as classification varies among practitioners, researchers and the applied methodologies (Helsel and Hirsch 2002; Rosen, Röttorp, and Jeppsson 2003; Branisavljević, Kapelan, and Prodanović 2011). The cause for erroneous or abnormal data is a combination of many factors, such as meter malfunctions, problems occurring in the data transmission and storage processes, changes in network system operations or burst pipes (Loureiro et al. 2016; Quevedo et al. 2017). We suggest dividing anomalous data into three categories:

- Type 1: Anomalies caused by faults internally in the meter, during transmission, storage, etc. that cannot be due to any state of the water distribution network (illegitimate data).
- Type 2: Measurements that affect the data quality negatively but have some valuable information left; examples include a loss of sensitivity and miscalibration of sensors.

CONTACT Jonas Kjeld Kirstein jkir@env.dtu.dk
Supplemental data for this article can be accessed here.

- Type 3: Abnormal measurements that are caused by actual events appearing in the network, such as leakage, irregular consumption or valve opening.

Type 1 contains data with no relevance for the operation of the system, because the integrity of the data is damaged or missing. This type of anomaly has also been described as 'dirty data' (Mounce, Boxall, and Machell 2010; Machell et al. 2014). Examples include missing or illegitimate observations, flatline segments, erroneous timestamps or duplicates. Applying time series data that includes type 1 data in hydraulic, leakage or water quality models can lead to incorrect conclusions. Since type 1 data contains no valuable information, it is preferable to remove type 1 anomalies from the time series before further analyses are made. Compared to type 1, anomalies of types 2 and 3 have a higher level of information and the major difference lies in the integrity. Type 2 includes erroneous data that can be corrected at times. These include, for example, errors due to wrong time settings, miscalibration of instruments or loss of sensitivity as seen in water quality sensors (García et al. 2017). Anomalous data of type 3 represent valuable information from correct physical measurements but comprises all abnormal observations that are caused by actual events in the network, such as burst pipes, pump trips, irregular consumption or valve operations. Research on abnormal consumption, leakage and/or burst detection methodologies is dependent on anomalies of type 3, while types 1 and 2 reduce the reliability of such applications (Wu and Liu 2017). It is our aim to develop a structured and improved identification of type 1 and 2 anomalies without misclassifying type 3 anomalies.

The existing literature focuses on batch validation of already collected data to prepare the data for further applications. The current work focuses on near real-time validation of the data as it is collected, while producing diagnostic plots that help the operators to detect errors in the data collection. Proper visualization of anomalies is just as important as their detection, since this enables the operator to correct for errors on a daily basis. If this is not to some extent automated in near real-time, experience shows that errors can prevail for months or years. Also, benchmarking, in terms of the number of anomalies registered over a selected period, can help utilities evaluate their data collection and acquisition performance.

To produce maximum benefits from the collected information, a well-maintained operational database containing validated and easily accessible data should be constructed (Cugueró-Escofet et al. 2016; Loureiro et al. 2016). In such a database, invalid and missing observations can be represented by coexisting qualified estimates based on, for example, time series analysis, physically based models or machine learning approaches such as artificial neural networks (Mounce, Boxall, and Machell 2010; Quevedo et al. 2010; Branisavljević, Kapelan, and Prodanović 2011; Cugueró-Escofet et al. 2016; García et al. 2017). Such reconstruction can benefit from the prior identification of anomalies, but we consider reconstruction of data outside the scope of the article.

Here we present a data processing framework that identifies anomalies in the collected data and store these in a malfunction indicator database (MAID) that is used for further analysis and visualization of anomalies. The method can validate the collected data without expert assistance and without sensor-specific parameterization, which will be useful when applied in future systems that include thousands of online sensors. If needed, the test parameters can be tuned to change the sensitivity of anomaly detection. To identify patterns in error occurrence, a diagnostic tool based on the Jaccard coefficient is proposed. The tool gives operators an idea of where to improve future data collection procedures. The method is demonstrated on 223 meter data sets from three Danish utilities where no data validation procedures are currently in place.

2. Methodology

2.1. Data processing framework

A conceptual scheme of a utility's data collection, processing and analysis system is shown in Figure 1. The system starts with the collection and storage of data from various sources, such as water quality sensors, pressure and flow meters, in a raw meter database. The next step is to categorize every raw data point as either an anomaly or a valid data point, based on a series of anomaly testing. The results from the tests are binary flags (true or false) stored in the MAID as an amendment to the operational database. The MAID highlights anomalies logged by sensors and facilitates the investigation of possible patterns in errors occurring in the data series. Since

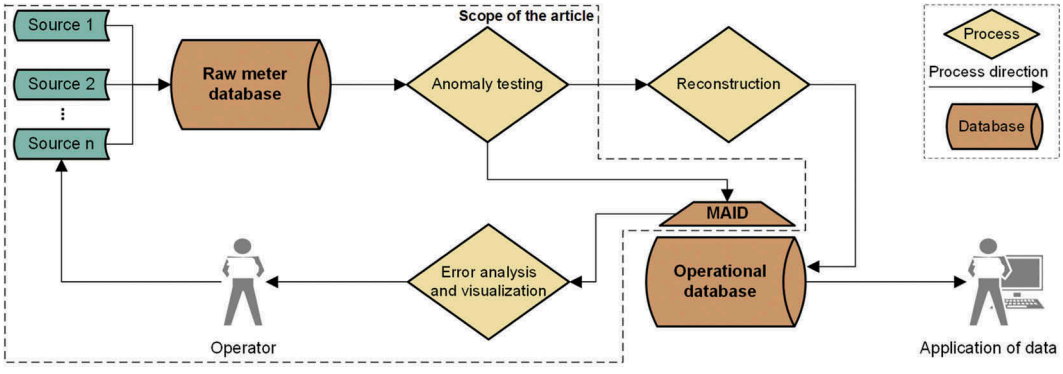


Figure 1. Data processing framework illustrating the transformation of data collected in a raw meter database to an operational database. The malfunction indicator database (MAID) amendment stores flags from the anomaly testing process, then uses this for the analysis and visualization of errors, ultimately suggesting improvements in current data collection procedures.

the test flags are binary values, the storage demand for the MAID entries is negligible compared to, for example, timestamp entries in the raw and operational database.

In the following, the entries in the databases are referred to as matrices and vectors:

- A raw data set from one meter consists of n observations $\boldsymbol{x}= [x_1,\ldots,x_n]$ with respective timestamps $\boldsymbol{t}= [t_1,\ldots,t_n]$.
- The MAID data for each data series is an n x m matrix $\mathbf{M}$ of binary values, where m is the number of tests included in the anomaly testing phase. The value of $\mathbf{M}_{k,i}$ is true only if the k-th test for the i-th time step detects an anomaly.

Flagged and missing data should be reconstructed and stored together with the validated data in an operational database. Such data are often stored in a uniform manner to account for differences in timestamp intervals between various data streams. This application of the data is outside the scope of this article (Figure 1). However, our error analysis and visualization step can use the flags stored in the MAID to provide both short- and long-term diagnostics as well as day-to-day visualizations of errors for use in the daily operation. Having detected anomalies, operators can use this information to investigate whether the data are, in fact, erroneous and to improve future data collection processes (Figure 1).

2.2. Anomaly testing of raw data

Seven tests (I–VII) form the anomaly testing process of the raw data that aim to identify type 1 and type 2 anomalies. A short summary explaining the occurrence of selected anomalies and justifying the necessity of the tests is given in the supporting information (SI) A. Tests I and II are run in a sequential manner to provide regular data streams to the subsequent tests III–VI that can be run in parallel. Finally, all data points flagged by tests I–VI are excluded in test VII. Whereas parts of the described tests are included in standard extract, transform, load (ETL) processes, their implementation may vary among practitioners. Moreover, it is important to note that this list of anomaly tests is not exhaustive, as it is based on the most common anomalies that appeared in the analysed data sets. Figure 2 exemplifies periods with validated and flagged data for each of the seven tests in a simplified manner. Based on time, a data point is seen as valid (marked green) until a new data point, pattern or inconsistency violating the tests' constraints has been registered (marked red).

I. Duplicate timestamp test

The duplicate timestamp test (Figure 2(a)) flags any t_i that is not unique in $\mathbf{M}_1$. Similarly, the raw data is checked for a regular data stream at the 'communications level' in Cugueró-Escofet et al. (2016). If duplicate timestamps can be replaced by known values, such as by an operator, it is possible to adjust the data set without compromising future applications. Otherwise, all duplicates represent a type 1 anomaly.

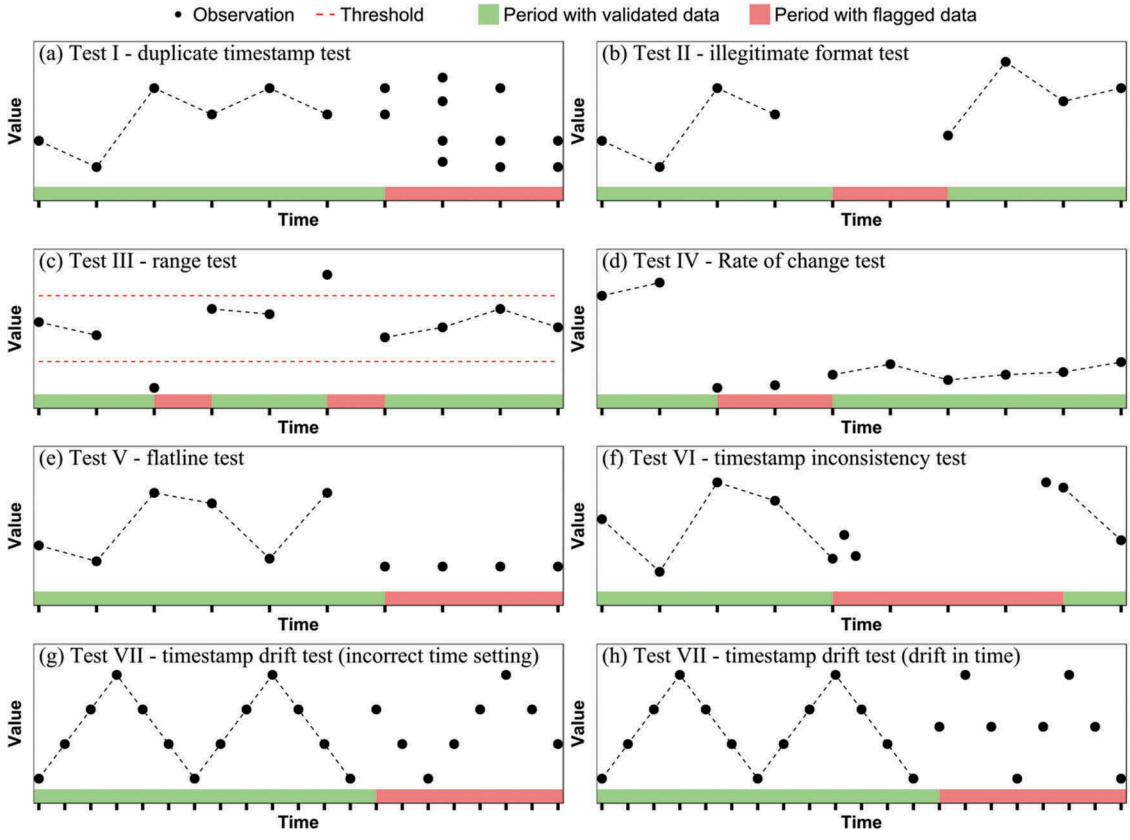


Figure 2. Overview of anomaly tests. Green and red distinguish valid and invalid data. Succeeding validated data are connected with a dashed line.

II. Illegitimate format test

In the illegitimate format test (Figure 2(b)), all timestamps t_i that are linked with non-numerical observations x_i are flagged in $\mathbf{M}_2$.

III. Range test

The range test (Figure 2(c)) identifies all x_i below or above a minimum or maximum threshold, i.e. physically unfeasible or dubious compared to the distribution of the data series. Inspired by the locally realistic range test (Mourad and Bertrand-Krajewski 2002), we flag values in $\mathbf{M}_3$ based on the distance between the median and the 2.5th and 97.5th percentiles of the data series, multiplied by a factor of α and β, respectively.

$$\mathbf{M}_{3,i} = \text{true}\begin{cases} < P_{50\%}(x) + [P_{2.5\%}(x) - P_{50\%}(x)] * \alpha, \alpha > 1 \\ > P_{50\%}(x) + [P_{97.5\%}(x) - P_{50\%}(x)] * \beta, \beta > 1 \end{cases} \quad (1)$$

Being similar to the interquartile range, this test is not as sensitive to outliers in data series as standard score-based methods (i.e. using distance to the mean relative to the standard deviation). This is particularly important, as anomalies in the data series should not be allowed to obstruct the identification of future anomalies. There is always a risk of type 3 being flagged, since the test is based on the historical distribution of the data (e.g. if the data set has not yet covered legitimate seasonal changes). This risk will decrease if α and β are increased but at the cost of sensitivity. Note that α and β values of 1 will result in 5% of the data being flagged as anomalies, while larger values for these parameters rapidly decrease this fraction, such that α and β values of 1.5 for normal distributed data will result in just 0.3% being flagged.

IV. Rate of change test

A rate of change test (Figure 2(d)) identifies unusual changes in observations over time based on the concepts of the step test (Estévez, Gavilán, and Giráldez 2011), signal gradient test (Mourad and Bertrand-Krajewski 2002) and trend level test (Cugueró-Escofet et al. 2016). A threshold θ for a likely rate of change is defined by computing the 97.5th percentiles of all absolute rates of change in a meter data set multiplied by a factor of λ:

$$\theta \geq P_{97.5\%}\left(\frac{|\Delta x|}{\Delta t}\right) * \lambda \quad (2)$$

Here, the rate of change $\Delta x/\Delta t$ is the ratio between the difference of two subsequent observations and their respective differences in time. In the case that the rate of change between two observations is above θ, the measured value is flagged in $\mathbf{M}_4$. Having flagged a value for two consecutive observations, the rate of change for the next observation is computed based on the last valid measurement and is tested for compliance with the threshold of θ. Flagging is stopped if a measurement is below θ. As in test III, the factor θ is based on the historical distribution of the data and might lead to false positives. Additional information about test IV is listed in SI B.

V. Flatline test

The flatline test identifies successive identical observations (Figure 2(e)). Here, flatline segments are flagged in $\mathbf{M}_5$ if successive identical observations more or equal to s cover a period of more than p minutes:

$$\mathbf{M}_{5,[i,\ldots,i+s-1]} = \text{true}\{x_i = x_{i+1} \ldots = x_{i+s-1} \text{ and } t_{i+s-1} - t_i > p\} \quad (3)$$

VI. Timestamp inconsistency test

The timestamp inconsistency test (Figure 2(f)) identifies irregularities in timestamp intervals. This is done to check that data is recorded and stored at a constant interval between measurements. Varying timestamp intervals indicate errors between measuring devices and data collection. Data are flagged in $\mathbf{M}_6$ if 1) the gap between measured data is more than q minutes; or 2) there are identical subsequent timestamp intervals less or equal to r. The latter check is needed to avoid data being marked as flawed because of a persistent change in logging frequency:

$$\mathbf{M}_{6,i} = \text{true}\{t_{i+1} - t_i > q \text{ or } t_{i+1} - t_i \neq (t_i - t_{i-1}, \ldots, t_{i-r+2} - t_{i-r+1})\} \quad (4)$$

VII. Timestamp drift test

The timestamp drift test is used to identify meters with drifting or changing time settings. Figure 2(g–h) shows an example of a meter suddenly having a wrong time setting and a meter drifting in time. The test works by assuming similar patterns in the data over a selected period of days and weeks, which allows for the identification of time shifts by comparing a given week with previous weeks. This implies that the test only works if the overall patterns in the data do not change with time. For the test, data excluding flagged values from tests I–VI is aggregated to hourly data to provide uniform time series without changing time intervals, and to reduce the computational time. Changing time settings are identified by the following method for a data set of length j days.

First, a reference day rd and a test day td, with a period of w weeks between, are selected, i.e. the test and reference day represent the same weekday. Next, the reference day (including the lags $l = \pm 12$ h) is cross-correlated against the test day and the peak correlation value $C_{daily}(d)$ and the corresponding hourly time lag $P_{daily}(d)$ are determined:

$$C_{daily}(d) = \arg\max\left\{corr[rd(l), td] = \frac{s_{rd(l)td}}{s_{rd(l)}s_{td}}, l = 0, \pm 1, \ldots, \pm 12\right\}, \quad d = 1, 2, \ldots, n \quad (5)$$

$$P_{daily}(d) = \text{lag } l \text{ of } C_{daily}(d) \quad (6)$$

where $s_{rd(l)td}$ is the covariance of $rd(l)$ and td, and $s_{rd(l)}$ and s_{td} the standard deviation of $rd(l)$ and td, respectively. $C_{daily}(d)$ and $P_{daily}(d)$ are invalidated if more than one third of the data in the reference or test period is unavailable. Next, $C_{daily}(d)$ is used to compute a weekly correlation measure, $C_{weekly}(d)$:

$$C_{weekly}(d) = \frac{\sum_{l=0}^{6} C_{daily}(d-l)}{\text{Valid test days in } C_{daily}[d:(d-6)]}, d = 7, 8, \ldots, j \quad (7)$$

$C_{weekly}(d)$ is invalidated if more than three days of data are missing. Finally, a weekly measure of the hourly drift, $P_{weekly}(d)$, is determined:

$$P_{weekly}(d) = \frac{\sum_{l=0}^{6} C_{daily}(d-l) * P_{daily}(d-l)}{\sum_{l=0}^{6} C_{daily}(d-l)}, d = 7, 8, \ldots, j \quad (8)$$

$P_{weekly}(d)$ is not computed in the case of previously invalidated values of $C_{weekly}(d)$.

To avoid a single week with a changed consumption pattern (such as a vacation week) resulting in time drift flags, steps 1–3 are run with three different reference periods and w is set to a 1-, 4- and 8-week interval. A drift is identified in the case that $P_{weekly}(d)$ exceeds a threshold of ±2 hours, allowing the time to fluctuate slightly before a flag is set. Additionally, a flag is only raised in $\mathbf{M}_7$ if the following two conditions are met:

(1) $C_{weekly}(d) \geq 0.8$, for at least two reference weeks w.
(2) $|P_{weekly}(d)| > 2$, for more than two subsequent test days d.

These conservative conditions ensure that the test is run only on data sets with a regular daily and weekly pattern, thus reducing the risk of a false alarm.

Parameter selection. It is unavoidable that type 3 anomalies are sometimes misclassified as type 1 or 2. This is one reason why dubious data points should not be deleted or replaced but rather flagged. Most tests presented have parameters that can change the sensitivity of the test and thereby also result in a trade-off between the misclassification of type 1 and 2 anomalies versus type 3 anomalies. To what extent such misclassification is a problem is dependent on the usage of the data. Thus, favourable parameter settings may differ between specific data sets. In Sec. 4.1, we propose a default parameter setting that works well with all data from the three utilities in the current study.

When a high proportion and variety of type I and II anomalies are present in raw data sets, the distribution of the data, independent of meter and data type, differs significantly (SI C). In such cases, outlier cut-off values based on historic averages and standard deviations may have an unwanted influence on the detection of anomalies. For tests III and IV, it was decided to base anomaly detection on the distance between median and low/high percentiles in the raw data sets, which will provide a more robust approach in case of skewed distributions.

Finally, utilities may assign varying quality levels to flagged anomalies that depend on the severity of a test's outcome, e.g. for prioritization during operational troubleshooting. In this study, no quality criterion was included.

2.3. Similarity analysis of anomalies

Similarities between flagged attributes in the MAID are analysed based on the Jaccard coefficient J, suitable for sparse and asymmetric binary data (Tan, Steinbach, and Kumar 2006). The similarity between a single binary feature $\mathbf{M}_k$ in the MAID of two selected meters φ and ω are assessed by:

$$J(\varphi, \omega) = \frac{f_{11}}{f_{01} + f_{10} + f_{11}}, where\ t(\varphi) \cap t(\omega) \quad (9)$$

Here, f_{11} is the number of attributes where φ and ω are equal to one (i.e. flagged attributes), f_{01} is the number of attributes where only φ is equal to one, f_{10} is the number of attributes where only ω is equal to one, and $\boldsymbol{t}(\varphi)$ and $\boldsymbol{t}(\omega)$ represent the meters' timestamps. The Jaccard coefficient ranges between no similarity (0) and a perfect similarity (1) and is only computed for matching timestamps. Here, we consider data entries with identical minute stamps to be matching.

A Jaccard coefficient close to unity means that flags occur simultaneously for two meters, which can be used to interpret the source and nature of the anomalies. If, for example, multiple sensors in an area are flagged as 'out of range' during the same periods of time, it is most likely due to actual physical conditions in the system (type 3 anomalies) rather than sensor errors. However, if sensors are located far from each other but share a high Jaccard coefficient, it might be due to database or transmission errors.

3. Case studies

Data from three Danish water utilities of varying size were analysed (Table 1). The three utilities represent typical water network data acquisition practices in Danish utilities, where mainly flow and pressure data are collected without a systematic analysis and tracking of errors and anomalies. Data is currently used for estimating minimum night flows and producing daily averages. The data sets from utilities A and C are based on pressure and flow measurements at storage locations (waterworks and tanks), DMA inlets and outlets, and other critical locations of the WDN such as pumping stations. In utility B, flow and pressure data is only collected in pipes that are highly important to the utility. Most of the installed devices in utilities A and C are Siemens MAG6000 and MAG8000 flow meters, whereas utility B measures flow with Primayer's PrimeProbe3. On average, a meter data set covered a period of 32 months. Additional information about the analysed utilities and an exhaustive description of the analysed raw meter data sets, including examples of analysed time series, can be found in SI C.

4. Results and discussion

All results listed in the following are a part of the error analysis and visualization process (Figure 1).

4.1. Anomaly testing of tests I–VI

We investigated the impact of various test parameters by manually looking through time series from the utilities, hereby identifying to what extent we agreed with the test results. Eventually, we selected a single set of parameters that works well for all three utilities (Table 2). It is this parameter set that is used for all results presented in the following, and we envisage that this

Table 1. Summary of analysed data sets from three utilities in Denmark. Consumption data from DANVA (2016). Meter types: P = pressure; Q = flow. See also SI C for additional information.

Utility	A	B	C
Name	Halsnæs Vand A/S	HOFOR (Greater Copenhagen Utility)	Nordvand A/S
Water supplied [Mm^3/yr]	0.6	50.5	7.1
Consumers [10^3]	10.4	585.8	142.7
Utility mains [km]	169	1085	529
Consumer/utility mains [km]	61.5	539.9	269.8
Waterworks	3	7	3
District Metered Areas (DMAs)	29	-	22
Number of meters/data sets [type]	68 (30P & 38Q)	44 (22P & 22Q)	111 (59P & 52Q)
Data period (avg. duration)	11/13–01/17 (Ø ≈ 25 months)	01/14–07/17 (Ø ≈ 33 months)	01/14–05/17 (Ø ≈ 35 months)
Total data points (P:Q)	33,487,833 (46:54)	5,691,133 (50:50)	157,876,298 (55:45)
Q – Average (min – max)[(1)] of medians [m3/h]	5.7 (0–60.8)	2.9 (−3.4–12.1)	61 (−12.6–463.1)
P – Average (min – max)[(1)] of medians [bar]	2.1 (0.3–4.3)	2.9 (1.4–5.1)	6.3 (4.8–7.4)
Q – Timestamp intervals [%]	91% in 1 min, 9% in other	54% in 5 min, 46% in 15 min	98% in 1 min, 2% in other
P – Timestamp intervals [%]	92% in 1 min, 8% in other	54% in 5 min, 46% in 15 min	97% in 1 min, 3% in other

Notes: 1) average, min and max of medians based on the median values of a utility's individual raw meter data sets.

Table 2. Default set of parameters applied to the anomaly tests III–VI.

Test	Parameters
III	$\alpha, \beta = 2.5$
IV	$\lambda = 2.5$
V	$s = 3$, $p = 60$ min
VI	$q = 60$ min, $r = 3$

parameter set could be used as a default parameter set for other utilities as well. This default parameter set can be seen as a qualified estimate that provides a good starting point; however, each utility company can alter the parameters according to its knowledge about its system and individual sensors. The distribution of raw and validated data for utilities A–C on different time scales is shown in SI C. In SI D we discuss how a variety of different parameter selections affects the overall test results, and an example of data validation and analysis for operational use is shown in SI E. In a near real-time implementation of tests III and IV, the test parameters that are based on the historical distribution should be updated at a user-defined interval. In this study, fixed parameters based on the entire data set available were selected.

In the following, we focus on the results stored in the MAID. Two mean error rates for tests I–VI were computed based on each meter's individual data set (Figure 3). The 'data point error rate' is the average rate of errors for each sensor in the utility based on the number of flagged measurements, while the 'time error rate' is the corresponding value based on the time affected by flagged measurements. Independent of utility and meter type, the time error rate varied between 12% and 49%. Thus, it is common for all three utilities to struggle to maintain a consistent and anomaly free data stream. Only utility C has a median value of the time error rate below 5% of data entering the system in constant intervals (test VI), illustrating a better data acquisition, maintenance and handling procedure than the other utilities. For utility A, the time error rate was above 15% for both flow and pressure measurements. A similar high rate for test VI indicates that this is mainly caused by missing or inconsistent data, as there is a large discrepancy between the mean flag rate of time and data points. In utility C, however, the data and time error rate are more alike. This is due to the setup of the utility's database, as only one data point is stored if measurements do not change over 15 minutes. The highest time error rates were found in utility B, with up to 49% and 20% of flow and pressure measurements, respectively. Missing data at such high rates not only indicates clear systematic problems within data acquisition practices but also challenges the

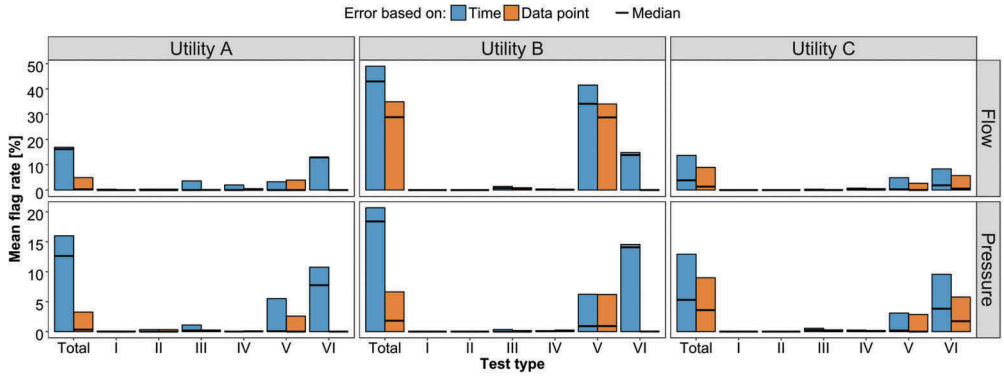


Figure 3. Mean meter flag rate of six different anomaly tests for the three utilities. The line across each bar represents the median flag rate. The *data point* bars show the number of flagged data points relative to the total number of data points, while the *time* bars show the period of time represented by the flagged data points relative to the total operating time of individual meters. Test types: I – duplicate timestamp, II – illegitimate timestamp, III – range, IV – rate of change, V – flatline and VI – timestamp inconsistency.

use and reliability of the data for later applications. For utilities A and B, the timestamp inconsistency bar reveals almost no error rate based on data points but a high rate based on time, averaging 10% for all three utilities. Thus, long periods of no data collection are not unusual. On average, 35% of the flow measurements were flagged as flatlines in utility B. The lowest average error rate based on data points was found in pressure meters of utility A (3%).

Parameter selection and implications

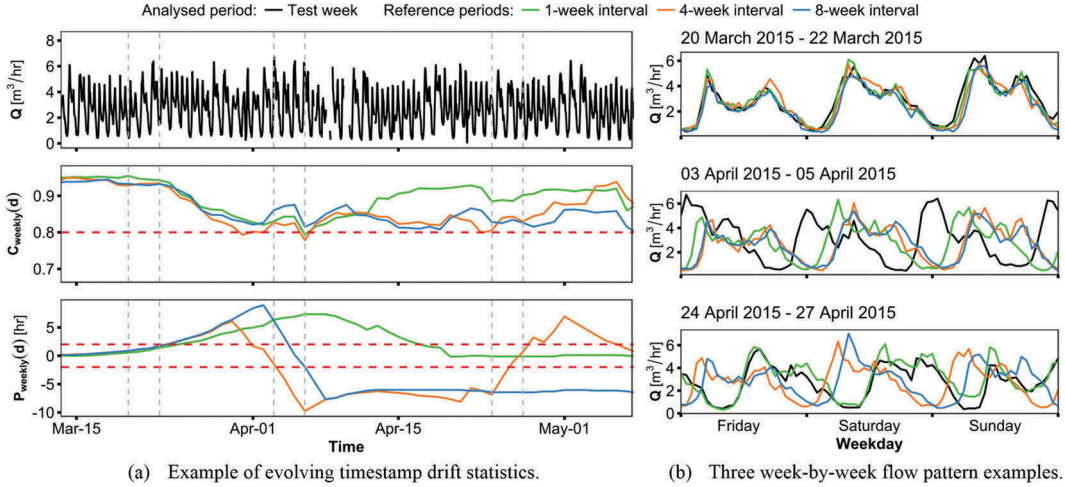
The selected parameters have some implications for the generated results. For example, there is a high risk that the parameter selection of the flatline test (Table 2) has led to neglect of anomalies at lower timestamp intervals. A future implementation could include a varying resolution of the number of significant digits in the flatline test or minimum rate of change test to identify meters measuring data with only minor changes, such as highlighting failing pressure sensors. In the case of data sets with true flatline segments (e.g. an emergency pump flow meter) utilities should have the possibility to suppress the generated anomalies or to skip certain tests from the overall analysis. It is possible that physical boundaries in the range and rate of change tests are known, and these should be applied instead of the formulated methods. Moreover, water supply data can exhibit strong seasonality on daily, weekly and yearly scales, and ideally, the period used for determining the various test parameters should be large enough to cover all of these scales. When this is not the case, the operator should be aware that the reliability of the tests is reduced. In general, it would be beneficial to compare anomalies identified by utility personnel or experts with the flags being raised by the method, as conducted by Branisavljević, Kapelan, and Prodanović (2011).

4.2. Anomaly testing of timestamp drift test (test VII)

It is commonly assumed that the internal clock of the data sources (i.e. meter) has an adequate accuracy and only needs few adjustments over its lifetime. To check if this is true, we applied the timestamp drift test. We applied various combinations of the parameters defining the conditions raising a flag and discussed their outcome in SI D. In the following, the test was applied according to the parameters and conditions stated in test VII (Section 2.2).

Figure 4 illustrates an example of a flow meter where a time drift is visualized. The upper row in Figure 4(a) displays the measured flow over time, and it is difficult to identify any problems from the time series itself except for missing data around the 8th of April. However, a clear drift is visible, when illustrating the time series by a week-by-week profile sectioned into the applied test and reference weeks (Figure 4(b)). The second row of Figure 4(a) shows the computed correlation value [$C_{weekly}(d)$] for the illustrated period, based on the timestamp drift test with a 1-, 4- or 8-week interval. A correlation value above the threshold of 0.8 is observed for most of the period, i.e. a clear daily/weekly recurring pattern is present in the meter time series. The third row shows at which hour the highest correlation value [$P_{weekly}(d)$] was determined. It can be seen that there is an upwards trend, illustrating the drift in time. The drift started at the end of March and was identified by all three reference intervals. It appears that the logger time is corrected at the beginning of April. However, whereas the 1-week interval slowly decreases towards a peak correlation hour of ±2 (approx. 15th of April), the 4- and 8-week intervals indicate that the new time setting is approximately 6 hours different to the setting before the drift started. Thus, the drift test is capable not only of identifying drifts in time but also of showing when a time setting is different (potentially erroneous) from earlier settings.

Test VII is summarized for all utilities in Figure 5. The first test condition can be interpreted as the percentage of data where the test could be run. For example, column 1 of utility A shows that 58% of the flow data passed test condition 1. In other words, 58% of the data included a regular weekly flow pattern as required by the test. Column 2 represents the number of meters meeting test condition 2, i.e. meters where a time drift was found at least once. This was the case for 10 meters in utility A.



(a) Example of evolving timestamp drift statistics. (b) Three week-by-week flow pattern examples.

Figure 4. Example of a drifting flow meter in utility A. (a) The upper row displays the measured flow over time. The second and third row illustrate the computed correlation values [$C_{weekly}(d)$] and [$P_{weekly}(d)$] based on the applied test and three reference weeks needed in test VII. The drift is visible in (b) where the reference weeks are plotted against the test week at three selected periods.

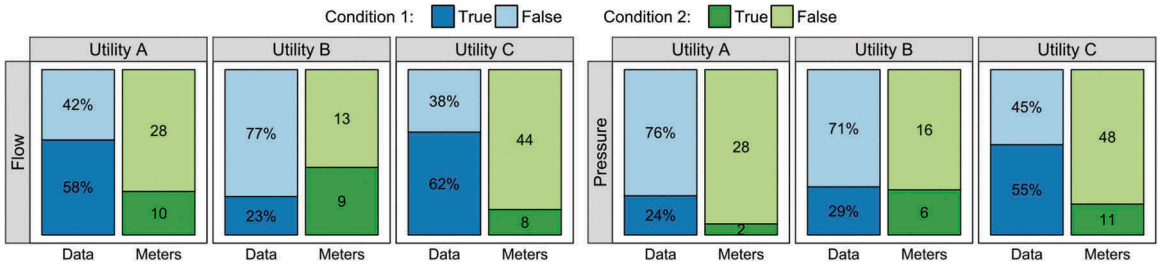


Figure 5. Summary of timestamp drift test. The data column accounts for the percentage of data where test condition 1 was met ($C_{weekly}(d) \geq 0.8$, for at least two reference weeks w). The meter column accounts for the number of meters meeting condition 2 at least once over the course of the measured period ($|P_{weekly}(d)| > 2$, for more than two subsequent test days d).

The test was conducted only on periods excluding flawed data (Figure 3). Consequently, the lowest percentage of data feasible for drift checking was found in utility B, with only 23% of flow and 29% of pressure data (Figure 5). However, 15 meters out of 44 drifting at some point is still a very high proportion of meters with a drift error. The highest share of data passing condition 1 was found in utility C, with up to 62% of flow and 55% of pressure data. Utility C confirmed that it observed drifting time in many of its battery-driven flow meters. To some extent, this validates the applied method. However, utility C had the lowest number of flow meters with a drift. A likely explanation includes the more proactive approach taken by the utility to correct meter errors. If the utility identifies and corrects a drift at a sufficiently early stage, the test will no longer raise an alarm owing to the conditions stated. In general, the test depends on a regular recurring pattern in the data and thus will not be applicable to all data series. Also, it is possible that the test might identify when a time setting is very different (potentially erroneous) compared to earlier settings but does not represent a drift. Nevertheless, our example illustrates the usefulness of testing for drifts.

4.3. Similarity analysis of anomalies

The Jaccard coefficient determines whether there are similarities or convergences in anomalies throughout the network, which can be an effective indicator of where to look for errors. Figure 6 illustrates the Jaccard coefficient for selected tests from utilities A and C. A Jaccard coefficient between 0 (marked green) and 1 (marked red) illustrates the degree of similarity in the occurrence of anomalies. In the figure, each flow and pressure meter was assigned an ID according to the total number of meter data sets and the Jaccard coefficient was evaluated for each set. Duplicate (M_1) and illegitimate datum flags (M_2) were evaluated as one category. These anomalies, as well as timestamp inconsistencies, occur in most cases simultaneously in meters ID 1–8 and ID 38–44 in utility A, evident from a Jaccard coefficient close to unity for these meters (Figure 6). These meters were not connected to the same supply areas nor did they use the same data transmission, but their data was collected in the same database, different from the remaining meters. This observation can be used by the utility to revise and improve its database setup. Furthermore, in the timestamp inconsistency test (test VI), a large number of flow meters and pressure meters have a high similarity in sets of two, visualized in the form of an almost straight line between flow and pressure sensor IDs in the Jaccard coefficient plot. As most pressure meters were installed to send data through a flow meter, any lack of data transmission would affect both flow and pressure observations at the same time. In the case of utility C, flagged values within test VI share a high similarity within a group of pressure meters and these meters are all located in the same supply area. Interestingly, measurements below the minute step appear simultaneously at the highlighted meters, indicating unknown meter or database settings or malfunctions. Jaccard coefficients for all remaining tests I–VI in utility A–C can be found in SI F.

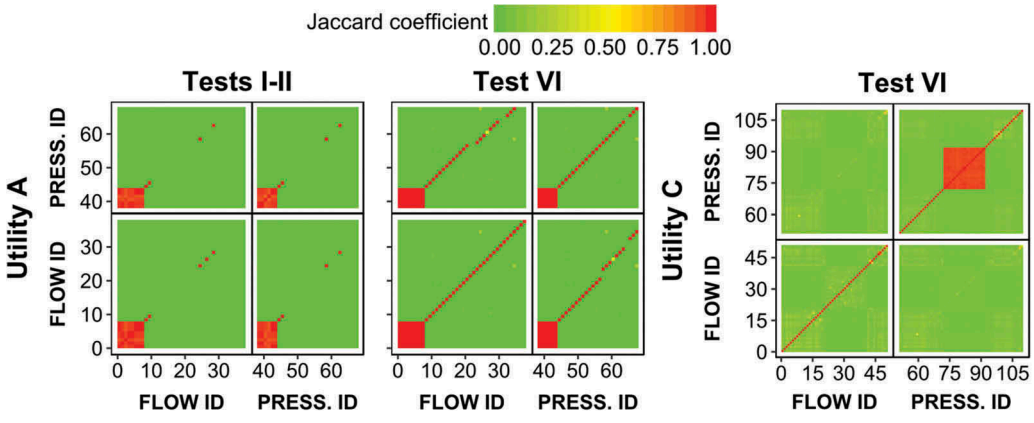


Figure 6. Jaccard coefficient computed for different tests from the anomaly testing covering flow and pressure meters. Results from the illegitimate and duplicate timestamp test were merged and illustrated in one column. Test types: I – duplicate timestamp, II – illegitimate timestamp and VI – timestamp inconsistency.

The occurrence of systematic errors (Figure 6) and large numbers of anomalies in the various tests (Figure 3) emphasises that there is a need for systematic data validation. It is our hope that the proposed methodology can assist a utility in systematically analysing errors and thereby markedly increase data quality in the future.

4.4. Future work

The next logical step includes a reconstruction process, generating qualified estimates of data being flagged as anomalies. The reconstruction method can utilize the MAID database in the training/calibration process to avoid being influenced negatively by anomalies of types 1 and 2. Additional data sets that can help to signal false misclassification of anomalies and improve the reconstruction process should be incorporated where possible. This includes data sets signalling the status of pumps, valves, external temperature or battery charge in selected devices. During reconstruction, the spatial and temporal redundancy between meters can be exploited (see, e.g. Cugueró-Escofet et al. (2016)); this, among other things, can make it possible to actively identify and flag type 3 anomalies. For example, due to a sudden drop in pressure, it is possible that a 'rate of change' flag was stored in the MAID. If the reconstructed value based on nearby sensors was close to the observed flagged value, the anomaly is unlikely to be due to a sensor error and can thus be categorized as type 3 anomaly.

5. Conclusion

This study revealed the need for the validation of meter data collected in water distribution networks. We have proposed a semi-automatic method to verify the collected data and highlight anomalies to deal with the increasing volume of collected data. The implementation of this method in a near real-time version, including a proper visualization of the error flags, will make it easy for the operators, on a day-to-day basis, to see whether sensors and data collection systems are working properly and to identify and correct errors when they arise. Having applied the method on 223 meter data sets from three Danish utilities, we found that, on average, at least 10% of the time that a meter collected data it was flagged as anomalous. For one utility, the collected flow data contained anomalous characteristics for an average of more than 35% of the time. Certain anomalies occur simultaneously throughout the network. Highlighting the occurrence of such similarities can help to improve future data collection and thus reduce the number of stored anomalies. The high rates of errors/anomalies could have been avoided if the proposed method had been implemented to automatically highlight meter errors in near real-time.

Acknowledgements

We thank all participating partners of the LEAKman project, Halsnæs Forsyning and Martin Brandt-Ewon from Nordvand A/S for providing data and answering questions related to data management practices.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the LEAKman project and partners under the Danish Eco-Innovation Program (MST-141-01277/NST-404-00378).

ORCID

Jonas Kjeld Kirstein http://orcid.org/0000-0002-0752-3061
Martin Rygaard http://orcid.org/0000-0001-8578-8842
Morten Borup http://orcid.org/0000-0002-6531-9464

References

Alegre, H., J. M. Baptista, C. E. Jr, F. Cubillo, P. Duarte, W. Hirner, W. Merkel, and R. Parena. 2006. *Performance Indicators for Water Supply Services*. London: IWA Publishing.

Andrews, L., K. Gasner, R. Sturm, W. Jernigan, S. Cavanaugh, and G. Kunkel. 2017. *Utility Water Audit Validation: Principles and Programs*. Denver: Water Research Foundation. http://www.waterrf.org/PublicReportLibrary/4639b.pdf.

Branisavljević, N., Z. Kapelan, and D. Prodanović. 2011. "Improved Real-Time Data Anomaly Detection Using Context Classification." *Journal of Hydroinformatics* 13 (3): 307–323. doi:10.2166/hydro.2011.042.

Cugueró-Escofet, M. À., D. García, J. Quevedo, V. Puig, S. Espin, and J. Roquet. 2016. "A Methodology and A Software Tool for Sensor Data Validation/Reconstruction: Application to the Catalonia Regional Water Network." *Control Engineering Practice* 49 (Apr.): 159–172. doi:10.1016/j.conengprac.2015.11.005.

DANVA (Danish Water and Wastewater Association). 2016. *Water in Figures 2016*. Skanderborg: DANVA. https://www.danva.dk/publikationer/benchmarking-og-statistik/water-in-figures-pdf/water-in-figures-2016/.

Estévez, J., P. Gavilán, and J. V. Giráldez. 2011. "Guidelines on Validation Procedures for Meteorological Data from Automatic Weather Stations." *Journal of Hydrology* 402 (1–2): 144–154. doi:10.1016/j.jhydrol.2011.02.031.

García, D., R. Creus, M. Minoves, X. Pardo, J. Quevedo, and V. Puig. 2017. "Data Analytics Methodology for Monitoring Quality Sensors and Events in the Barcelona Drinking Water Network." *Journal of Hydroinformatics* 19 (1): 123–137. doi:10.2166/hydro.2016.048.

Helsel, D. R., and R. M. Hirsch. 2002. *Statistical Methods in Water Resources Techniques of Water Resources Investigations*. *Book 4, chapter A3*. Reston, VA: U.S. Geological Survey. https://pubs.usgs.gov/twri/twri4a3/.

Jørgensen, H. K., S. Rosenørn, H. Madsen, and P. S. Mikkelsen. 1998. "Quality Control of Rain Data Used for Urban Runoff Systems." *Water Science and Technology* 37 (11): 113–120. doi:10.1016/S0273-1223(98)00323-0.

Loureiro, D., C. Amado, A. Martins, D. Vitorino, A. Mamade, and S. T. Coelho. 2016. "Water Distribution Systems Flow Monitoring and Anomalous Event Detection: A Practical Approach." *Urban Water Journal* 13 (3): 242–252. doi:10.1080/1573062X.2014.988733.

Machell, J., S. R. Mounce, B. Farley, and J. B. Boxall. 2014. "Online Data Processing for Proactive UK Water Distribution Network Operation." *Drinking Water Engineering and Science* 7 (1): 23–33. doi:10.5194/dwes-7-23-2014.

Mounce, S. R., J. B. Boxall, and J. Machell. 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. doi:10.1061/(ASCE)WR.1943-5452.0000030.

Mourad, M., and J.-L. Bertrand-Krajewski. 2002. "A Method for Automatic Validation of Long Time Series of Data in Urban Hydrology." *Water Science and Technology* 45 (4–5): 263–270. doi:10.2166/wst.2002.0601.

Puig, S., M. C. M. van Loosdrecht, J. Colprim, and S. C. F. Meijer. 2008. "Data Evaluation of Full-Scale Wastewater Treatment Plants by Mass Balance." *Water Research* 42 (18): Elsevier Ltd: 4645–4655. doi:10.1016/j.watres.2008.08.009.

Quevedo, J., D. Garcia, V. Puig, J. Saludes, M. A. Cugueró, S. Espin, J. Roquet, and F. Valero. 2017. "Sensor Data Validation and Reconstruction." In *Real-Time Monitoring and Operational Control of Drinking- Water Systems*, edited by V. Puig, C. Ocampo-Martínez, R. Pérez, G. Cembrano, J. Quevedo, and T. Escobet, 175–193. Cham: Springer International Publishing.

Quevedo, J., V. Puig, G. Cembrano, J. Blanch, J. Aguilar, D. Saporta, G. Benito, M. Hedo, and A. Molina. 2010. "Validation and Reconstruction of Flow Meter Data in the Barcelona Water Distribution Network." *Control Engineering Practice* 18 (6): 640–651. doi:10.1016/j.conengprac.2010.03.003.

Rosen, C., J. Röttorp, and U. Jeppsson. 2003. "Multivariate on-Line Monitoring: Challenges and Solutions for Modern Wastewater Treatment Operation." *Water Science and Technology* 47 (2): 171–179. doi:10.2166/wst.2003.0113.

Tan, P.-N., M. Steinbach, and V. Kumar. 2006. *Introduction to Data Mining*. Boston: Pearson Addison Wesley.

WLCC (Water Loss Control Committee). 2014. "AWWA Water Audit Software (Version 5)." Microsoft Excel file. Denver: American Water Works Association.

Wu, Y., and S. Liu. 2017. "A Review of Data-Driven Approaches for Burst Detection in Water Distribution Systems." *Urban Water Journal* 14 (9): 972–983. doi:10.1080/1573062X.2017.1279191.