An Open-Source, Semisupervised Water End-Use Disaggregation and Classification Tool

Nour A. Attallah, A.M.ASCE[1]; Jeffery S. Horsburgh, Aff.M.ASCE[2]; and Camilo J. Bastidas Pacheco[3]

Abstract: This paper demonstrates a new water end-use disaggregation and classification tool that builds on existing end-use disaggregation studies and addresses the unavailability of code and data used by prior studies. The tool was developed in Python and can be accessed via any current Python programming environment. The base disaggregation and classification model for the tool was developed and tested on high-temporal-resolution data for a single home at which manually labeled end-use event data were also collected. The tool was then applied to four additional homes selected from a larger data set for 31 homes located in the cities of Logan and Providence, Utah, to demonstrate the generalizability of the tool. At homes for which no manually labeled end-use data are available, the tool's base model is extended through a self-learning procedure that trains the model for an individual home using end-use events identified at that home. Results from homes with different meter types and sizes are presented to demonstrate the ability of the tool to disaggregate and classify high-temporal-resolution data into individual end-use events. The results of this paper are reproducible using openly available code and data, representing an accessible platform for advancing end-use disaggregation tools. The tool can be adapted to specific research needs. **DOI: 10.1061/JWRMD5.WRENG-5444.**

Introduction

Water is a crucial resource on which humans depend for survival. Yet, increasing water scarcity, declining water quality, global climate change, and growing water stresses from the residential sector are creating new challenges for water managers to secure water supplies for current and future water demands. In most large urban water systems in the US, the residential sector consumes the majority of total supplied fresh water, using on average about 56% of piped fresh water (Contestabile 2018).

In the US, metering of residential water use is widespread (Boyle et al. 2013). However, most residential meters rely on measurement technology created decades ago, many of the meters themselves are decades old, and most are only read by a water utility monthly to quarterly for billing purposes. It is well-known that monthly data are too infrequent to characterize patterns in water use, leaving gaps in our understanding of behavior at the household and system level (Cardell-Oliver 2013). This limits our ability to identify alternative management strategies and opportunities for water conservation and increased efficiency.

With the advent of smart metering technology, several water-monitoring studies using high-temporal-resolution data (data with temporal resolution <1 min) were conducted from the single household level to large cities (e.g., Anda and Temmen 2014; Boyle et al. 2013; Cominola et al. 2021; Sønderlund et al. 2016; DeOreo et al. 2016; Froehlich et al. 2011; Kowalski and Marshallsay 2005; Mayer et al. 2004; Mayer and DeOreo 1999; Wong et al. 2010; Willis et al. 2013). These studies demonstrated that high-resolution data can enhance the ability to quantify water user behavior, identify and characterize different end uses, and formulate feedback to engage water consumers and foster water conservation. Given the difficulty associated with measuring flow at each point of use (e.g., individual faucets, toilets), high-resolution data are typically collected on a water meter that measures the total water flow to a residence. These data are often referred to as a total water use "trace." Water end-use characterization consists of breaking down the trace data at the household level into different end-use categories (Cominola et al. 2015). Identifying with a high level of certainty where, when, and how much water is used by a household can help water managers in understanding demand patterns, identifying opportunities for conservation, and checking the effectiveness of proposed and implemented conservation actions.

Several water end-use models and software tools that use high-resolution metering data to identify end uses have been developed, including Trace Wizard (DeOreo et al. 1996), Identiflow (Kowalski and Marshallsay 2003), HydroSense (Froehlich et al. 2009), and Autoflow (Beal et al. 2011). These tools are resource and time intensive, with a significant part of the data processing and analysis requiring human intervention. Some involve an intrusive period of data collection to train the algorithm with required water end-use signature data. Despite the relatively large number of papers published about end-use disaggregation tools and studies of water-use behavior that have used the tools, opportunities for reproducing or replicating existing studies or building on their results are limited because neither the data nor the code for existing algorithms are openly available and/or easily accessible (Di Mauro et al. 2020). In our inquiries with authors of these papers, we found that the code was considered proprietary and could not be released, and the data sets were inaccessible due to privacy concerns. Yet many of these

[1]Ph.D. Candidate, Dept. of Civil and Environmental Engineering, Utah State Univ., 8200 Old Main, Logan, UT 84322-8200 (corresponding author). ORCID: https://orcid.org/0000-0002-2227-7321. Email: nour.atallah@usu.edu

[2]Associate Professor, Dept. of Civil and Environmental Engineering and Utah Water Research Laboratory, Utah State Univ., 8200 Old Main Hill, Logan, UT 84322-8200. ORCID: https://orcid.org/0000-0002-0768-3196. Email: jeff.horsburgh@usu.edu

[3]Postdoctoral Researcher, Idaho National Laboratory, 1955 N. Fremont Ave., Idaho Falls, ID 83415; formerly, Postdoctoral Scholar, Dept. of Civil and Environmental Engineering, Utah State Univ., 8200 Old Main, Logan, UT 84322-8200. Email: camilo.bastidas@usu.edu

Note. This manuscript was submitted on June 15, 2021; approved on February 8, 2023; published online on May 9, 2023. Discussion period open until October 9, 2023; separate discussions must be submitted for individual papers. This paper is part of the ***Journal of Water Resources Planning and Management***, © ASCE, ISSN 0733-9496.

papers call for additional research to verify and extend methods as well as for new applications of results, including those that describe Trace Wizard (DeOreo et al. 1996), Identiflow (Kowalski and Marshallsay 2003), HydroSense (Froehlich et al. 2009), and Autoflow (Beal et al. 2011). Thus, there is a clear need for open and reproducible approaches that enable other researchers to test, replicate, reuse, and build on existing work. In response to these issues, this paper presents new, open-source, fully automated, semisupervised, and nonintrusive techniques for collecting and disaggregating high-resolution water-use data into component end uses, along with an openly available, anonymized data set for testing this and potentially other water end-use disaggregation algorithms.

Background

Despite the variety of tools and techniques adopted by existing water end-use studies, they all followed the same four general phases of (1) data gathering, (2) data cleansing, (3) water end-use disaggregation, and (4) classification (Fig. 1) (Cominola et al. 2015; Pastor-Jabaloyes et al. 2018). While data gathering is a necessary first step, a thorough discussion of challenges and variability in data gathering methods is beyond the scope of this paper and we do not discuss it here. Data cleansing is a preprocessing step that prepares raw, high-frequency water-use data for subsequent steps. Disaggregation extracts events from cleansed data, separates overlapping events (i.e., events made up of simultaneous end uses), and then identifies event features (e.g., volume, duration). Classification assigns events to an end-use category based on their features. In the following sections, we describe each subprocess in more detail.

Data Cleansing Techniques

Water-use data consist of time series of flow where the characteristics of the signal (i.e., periods of nonzero flow identified as water-use events and the features of those events, including volume, duration, and flow rate) reflect the type of end uses in a household. Within the time series of flow data, periods of nonzero flow constitute water-use events. Events must first be identified, and their features calculated, before they can be classified. This is not always simple given that some noise or signal distortion is always expected to be embedded in the raw trace data caused by the combination of data recording frequency and the volumetric pulse resolution at which data are collected (in many cases manifesting as a volume of water per electronic pulse generated by a water meter that can be counted). Noise can impede accurate data interpretation, including difficulty in identifying the start and end of events as well as accurate calculation of other event attributes and features.

Filtering can be used to remove undesired noise and make data ready for further analysis. Filtering can remove certain components of the signal (e.g., high-frequency oscillations caused by the meter's pulse resolution) while retaining other components (e.g., the overall shape of an event). Techniques for time series data filtering include empirical mode decompositions (Flandrin et al. 2004) and Monte Carlo techniques (Doucet et al. 2000). However, despite their wide use, their performance on water-use data where the means, variance, and covariance change over time is poor (Nayak et al. 1999). As an alternative, Chen (2014) suggested isolating signal frequencies of interest by removing or keeping them either at the top (low-pass filter), the bottom (high-pass filter), or both sides of the domain (band filter).

Given the variety in temporal and pulse resolutions recorded by smart water meters, filtering is an active area of investigation for cleansing high-resolution water-use data prior to end-use analysis. For example, in this study we describe how a low-pass filtering technique can be applied to the raw trace data to maintain low-frequency signals and adjust or remove high-frequency oscillations caused by the pulse resolution of the meter. This worked well for the data recording frequency and water meter pulse resolutions we encountered in our case study, but we acknowledge other filtering and preprocessing techniques may work better for different data resolutions.

End-Use Disaggregation Techniques

Single events are those where a single fixture is in use, while overlapping events occur when two or more fixtures are in use simultaneously. Trace disaggregation iterates on overlapping events until all subevents are single events. The process involves (1) extracting events from the trace, (2) classifying events into either single or overlapping events, and (3) breaking down overlapping events until all resulting events are single events. An additional step involves calculating the features for all single events (e.g., start time, end time, duration, volume, average flow rate). Features play an important role in classifying events as either single or overlapping. For example, Pastor-Jabaloyes et al. (2018) classified events as single or overlapping based on the number of vertices present in their filtered flow data, where a vertex is a point where the flow rate changes from one nonzero value to another within the same event. They presumed that events having only four vertices should be classified as single events. Events with more than four vertices were classified as overlapping.

Overlapping events have not been consistently handled by existing studies (Table 1). Most have concluded that overlapping events in single-family houses account for a relatively small proportion of total events. Moreover, methods used by those who have tried to break down overlapping events were built on assumptions (e.g., if the flow rate changes within an event, a new event is assumed to be introduced to the trace, and the time at which the flow rate change occurred is assumed to be the start time of the new event). Table 1 summarizes features incorporated in the disaggregation process, methods used in classifying events as single or overlapping, and methods used in breaking down overlapping events into single events for the most used end-use disaggregation software tools.

Although the methods in Table 1 work for identifying single events and in some cases disaggregating two overlapping end uses, to our knowledge no method has been able to accurately disaggregate overlapping events comprised of more than two overlapping end uses. Thus, in our case study, we describe a new method for disaggregating complex, overlapping events formed by more than two simultaneous end uses by incorporating multiple physical features in the disaggregation process.

End-Use Classification Techniques

One accurate approach to better understand household water use traces is through simultaneously monitoring each individual water end use in the household with a water meter. However, developing such instrumentation is time consuming, invasive, and expensive (Längkvist et al. 2014). An alternative is to use machine learning to classify events of different end-use types derived from the raw trace data logged on the main water meter. Supervised techniques require a training data set in which a set of events have been recorded and classified and from which a model is derived for classifying new events based on their features. Unsupervised techniques can cluster similar events regardless of the number of known events and without labeled training data, but it can be difficult to interpret the clusters produced. If labels for a small set of events are available, a class of algorithms called semisupervised learning can be used. Semisupervised techniques combine a small set of labeled

J. Water Resour. Plann. Manage., 2023, 149(7): 04023024

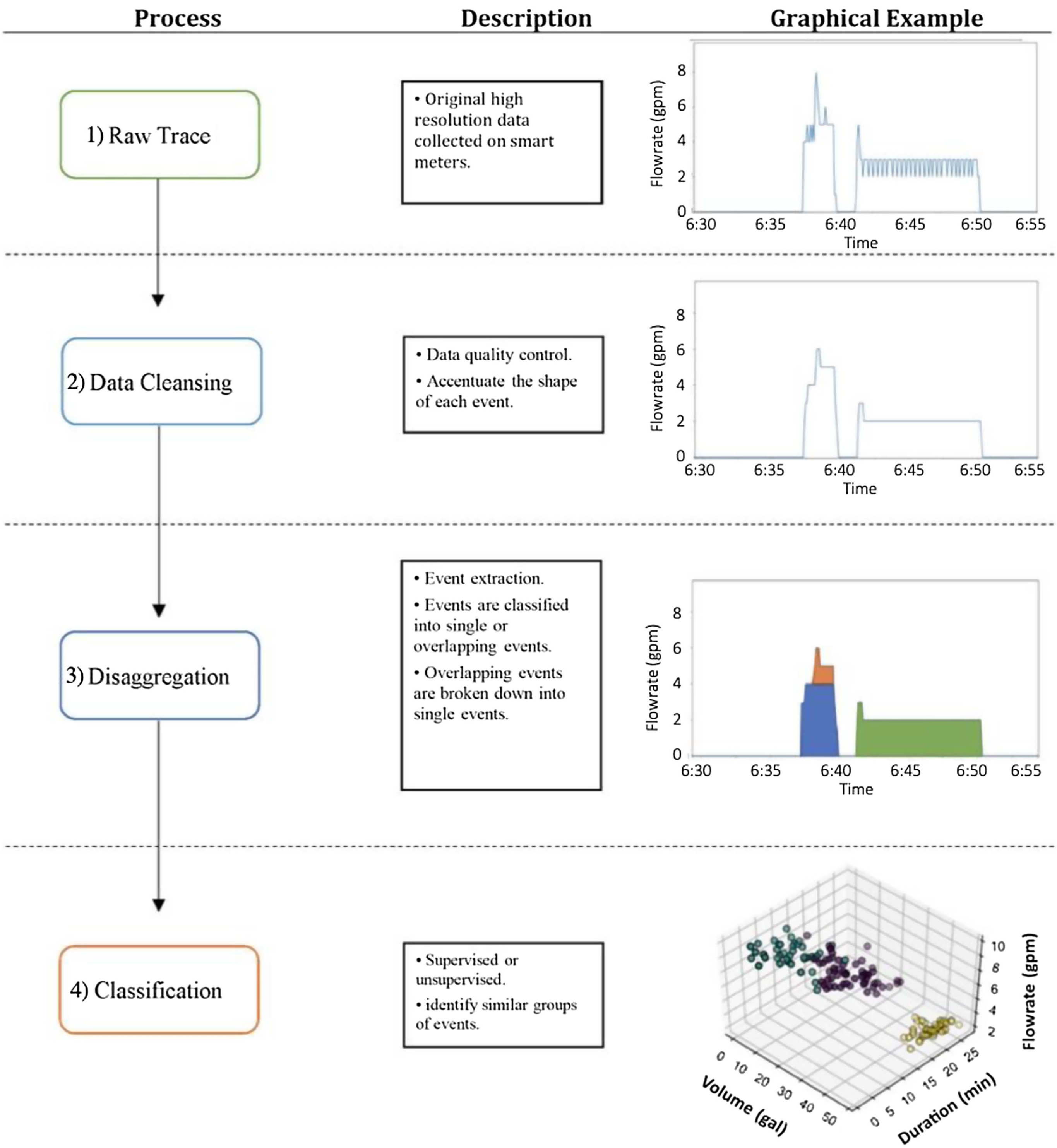

Fig. 1. Water end-use disaggregation methodology.

data with a large set of unlabeled data to aid in the classification process. Here we define a small data set as one that could be created by a single homeowner through manually labeling events (e.g., on the order of hundreds of real events). Supervised machine learning classifiers often require training data sets with on the order of tens of thousands of instances to train and test the algorithm, whereas with semisupervised classifiers, a data set with less than 1,000 instances can be used to train and test the algorithm (Chapelle and Zien 2005).

In the context of end-use disaggregation and classification, most studies have adopted supervised techniques such as decision trees (Kowalski and Marshallsay 2003; DeOreo et al. 1996), hidden Markov model (HMM) (Nguyen et al. 2013), Bayesian probabilistic models (Froehlich et al. 2011), and hybrid signature-based iterative disaggregation (Cominola et al. 2017). Others have utilized unsupervised, partitional clustering such as K-means (Yang et al. 2018) and K-medoids (Pastor-Jabaloyes et al. 2018). However, the practical application of these techniques is limited because they

Table 1. Event disaggregation process for different software tools

Software tool	Features incorporated	Overlapping event identification method	Overlapping event separation method
Trace wizard (DeOreo et al. 1996)	Start time, duration, volume, maximum flow rate, and mode flow rate	Decision tree algorithm	N/A
Identiflow (Kowalski and Marshallsay 2003)	Duration, volume, average flow and maximum flow	Decision tree algorithm	Decision tree algorithm and volume threshold value
Autoflow (Nguyen et al. 2013)	Volume, duration, maximum flow rate, mode flow rate, frequency of mode flow rate, magnitude of initial flow-rate rise, magnitude of flow-rate drop at the end of event, gradient of initial flow-rate rise, and gradient of flow-rate drop at the end of the event	A hybrid of hidden Markov model (HMM), artificial neural networks (ANNs), and the dynamic time warping (DTW) algorithms	A hybrid of hidden Markov model (HMM) and ANN algorithms
HydroSense (Froehlich et al. 2009)	N/A	N/A	N/A
Soft computing technique (Pastor-Jabaloyes et al. 2018)	Total volume, average flow rate, and the number of vertices	Number of vertices	Gradient flow-rate change and average flow-rate threshold value

Note: If a tool does not use any feature in the disaggregation, does not attempt to identify whether the event is single or overlapping, or does not attempt to break down overlapping events, N/A is reported.

either require predefining the number of clusters (i.e., the number of water end uses inside a household) in the case of the unsupervised techniques, or a large group of water end-use events that have been accurately tagged with one or more labels that can be used to train a supervised model. Determining the number of water end uses requires manual surveys of residents, while generating labeled data sets requires manual logging of water use events. Both are labor intensive and difficult to achieve at any scale.

To address the limitations of previous studies, we sought to develop a new, fully automated machine learning technique for identifying and classifying end uses from water trace data that does not require a large, fully labeled data set of water end uses for training or prior knowledge of the number of end uses inside a home for classifying events. We explored a semisupervised approach using a small set of labeled events from a single home to help in the classification of a much larger set of unlabeled events across multiple homes. The expense associated with collecting a large data set of manually labeled events along with the fact that a small labeled data set is not representative of the true variance of the data made development of a fully supervised model impractical. Semisupervised learning makes it possible to combine the advantages of working with a small, labeled data set to guide the learning process and a much larger, unlabeled data set to increase the generalizability of the solution.

Methods

We used a combination of data mining and semisupervised machine learning to extract and classify high-resolution water use data into component end uses, without the need for complementary information on the number of end uses inside a household. Given that our aim was to develop a fully automated water end-use disaggregation and classification tool, we investigated which features can reveal the end-use type of each event. In the following sections, we describe the case study design and methods we used to for testing the primary phases of the water end-use disaggregation and classification process.

Case Study Design and Data Collection

The high-temporal-resolution data we used were collected at 31 residential homes located within the cities of Logan and Providence, Utah. These cities made available to us their monthly water-use data for residential customers. We ranked users based on their annual average water use computed from monthly records and divided them into classes of low (< 33rd percentile), medium (33rd to 66th percentile) and high (> 66th percentile) water users. From these classes, we randomly subsampled and invited potential participants to participate in our detailed data collection. Prospective participants were sent a letter in the mail inviting them to participate in this study. Of 200 letters sent, 11 participants responded positively and enrolled. Given the low response rate to mailed letters, an additional 20 participants were recruited and enrolled through word of mouth and targeted invitations. We achieved a sample size of 31 participants that broadly represent the spectrum of water users within Logan and Providence.

For each participating household, we collected water use data using the home's water meter for 2 weeks in the summer when outdoor water use was ongoing and another 2 weeks in the late fall and winter when there was no outdoor water use (4 weeks total). We utilized low-cost electronic data loggers with high-temporal-resolution data-collection capabilities (Bastidas Pacheco et al. 2020). These data loggers were installed on the existing water meters at each home and recorded observations of water use with a 4-s data recording interval with pulse resolution that depended on the meter type and size.

For each participating household, we conducted a brief survey to identify the water end uses present in the home. We conducted regular visits to download the high-frequency data and saved them as comma-separated values (CSV) data files. Raw data were managed initially on a field laptop before being transferred to a secure file sharing system (for archival of original data files) and an operational database server (to support high-performance queries and analysis) for secure and shared access among the research team.

One difference among participants was meter type and size. For participants in Logan, Neptune T-10 water meters were observed with pipe sizes of 0.625 or 1 in. The meter sizes are described in inches with pulse resolutions in gallons to match manufacturer specifications for how these meters are sold in the US. T-10 meters produce a magnetic pulse detectable by our low-cost data loggers every 0.03 L (0.008 gal.) or 0.076 L (0.02 gal.) of water consumed, respectively. In Providence, Master Meter bottom-loading multijet meters were observed with pipe sizes of 0.625 or 1 in. These meters generate a magnetic pulse every 0.076 L (0.02 gal.) or 0.15 L (0.04 gal.) of water consumed, respectively. Pulse resolutions of these meters were determined through laboratory testing at the Utah Water Research Laboratory (Bastidas Pacheco et al. 2020).

Alongside the high-temporal-resolution water-use data, we asked a resident of one participating household to manually label some water-use events. Each time an end use was initiated inside the home, the type of end use and its start time were recorded. Matching the times of manually labeled events with the high-resolution trace data enabled us to better understand the shape of events for each water end-use type, their characteristics, and, most importantly, the variables that contributed the most in distinguishing between events of different types. A total of 998 different single water end-use events were manually labeled. We collected an initial data set containing 538 events to serve as a training data set, and we conducted an additional data collection period during which we collected 460 events to serve as a testing data set. Besides the manually labeled events, we manually added irrigation events to the data set because they were not labeled by the homeowner and they were easily distinguishable in the trace of water use.

For water end-use analysis, we selected a subset of five households with different meter sizes and types from the larger set of 31 sampled homes. Four of the households were selected because they had the highest number of residents and the highest average daily water use compared with other households with similar meter size and type. The fifth was selected as the household at which manually labeled events were recorded. The results reported in this paper are not meant to present a comprehensive analysis of water-use behavior in all of the homes for which we collected data, but rather are focused on demonstrating the effectiveness of the disaggregation and classification tool we developed. All five households had outdoor water use and used sprinklers for irrigation. Table 2 summarizes the general characteristics of the selected households. Households 1, 2, 3, 4, and 5 in Table 2 correspond to Households 3, 11, 24, 27, and 19, respectively, in the Bastidas Pacheco et al. (2023) HydroShare data set that contains data for all 31 sampled households. Identifiers have been changed here for convenience in referring to them in the text of this paper. To facilitate reproducibility of the results presented in this paper, the high-resolution data for the five households listed in Table 2, along with the manually labeled event data for Household 5, are available in a separate HydroShare resource (Attallah and Bastidas Pacheco 2023).

Data Cleansing

We cleansed the raw trace data for each residence using a low-pass filter [Eq. (1)] (Broesch 2008) to enhance the process of extracting events and their associated physical features (e.g., volume, duration). The filter modifies the raw trace data to accentuate event start and end times and the number of vertices of each event. Values output by the low-pass filter were rounded down to the nearest integer value. Fig. 2 shows an example of the output of this process. We chose a low-pass filter coupled with rounding because it retains the overall shape of events while removing or adjusting oscillations caused by the data recording interval and pulse resolution of the meter, leading to more clearly recognizable events

$$P'_j = \begin{cases} \dfrac{\sum_{i=1}^{n} P'_{j-i} + \sum_{i=0}^{n} P_{j+i}}{2n+1} & \text{for center data points} \\ \dfrac{\sum_{i=0}^{n} P_{j\pm i}}{n+1} - 1 & \text{for edge data points} \end{cases} \tag{1}$$

where P'_j = filtered data point at the jth index; P'_{j-1} = predecessor filtered data point value within the number of time periods n; and P_{j+i} = original data point value at the $j+1$ index within n periods of time. Edge points are the first and the last data points in the series, while center points are all other data points between the two edge data points. We used P_{j+i} for the first data point, and P_{j-i} for the last data point. For filter configuration, we used root-mean-square error (RMSE) to measure the difference between values predicted by the filter and observed values at different time periods n. A value of 1 for n yielded the lowest value of RMSE, and thus was selected for the filter.

Event Disaggregation

The data cleansing process resulted in filtered data points that accentuate the shape of each event within the trace and make it easier to identify the beginning and end of each event. For example, in Fig. 2 the unfiltered trace data peaks at seven pulses near the beginning of the event, which might be mistaken as a different, short-duration event superimposed on the longer event having values of four and five pulses. After filtering, the event is more identifiable as one event. From there, we identified all periods of nonzero flow in the unfiltered trace data as events. After that, we calculated several features for each event using either the filtered or the raw trace data (Table 3). The volume, duration, flow rate, and mode flow rate have been used in other water end-use disaggregation studies and have been proven useful in separating and identifying events. To that list of features we added peak flow rate, peak flow-rate frequency, mode flow-rate frequency, RMS, number of vertices, two irregularity measures, and a complexity measure because we noticed these additional features enhanced the disaggregation process.

The RMS of each event was calculated using the formula suggested by Coppack (1990)

$$\mathrm{RMS} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} FR_i^2} \tag{2}$$

Table 2. Characteristics of the five households selected for water end-use analysis

Household	Meter size (in.)	Meter type	Meter resolution (L/pulse)	Number of residents	Number of bathrooms[a]	Irrigation system	Legal property size (m^2)	Building size (m^2)
1	5/8	Neptune T-10	0.033	4	3.5	Manual	1,133	138
2	1	Neptune T-10	0.126	4	3.5	Automatic	890	136
3	1	Master Meter BL	0.157	6	3.5	Automatic	1,052	144
4	5/8	Master Meter BL	0.096	7	3.5	Automatic	3,116	300
5[b]	1	Master Meter BL	0.157	5	3	Automatic	1,133	128

[a]A half bathroom contains only a toilet and sink, but no bathtub or shower.
[b]Manually labeled events were collected for Household 5.

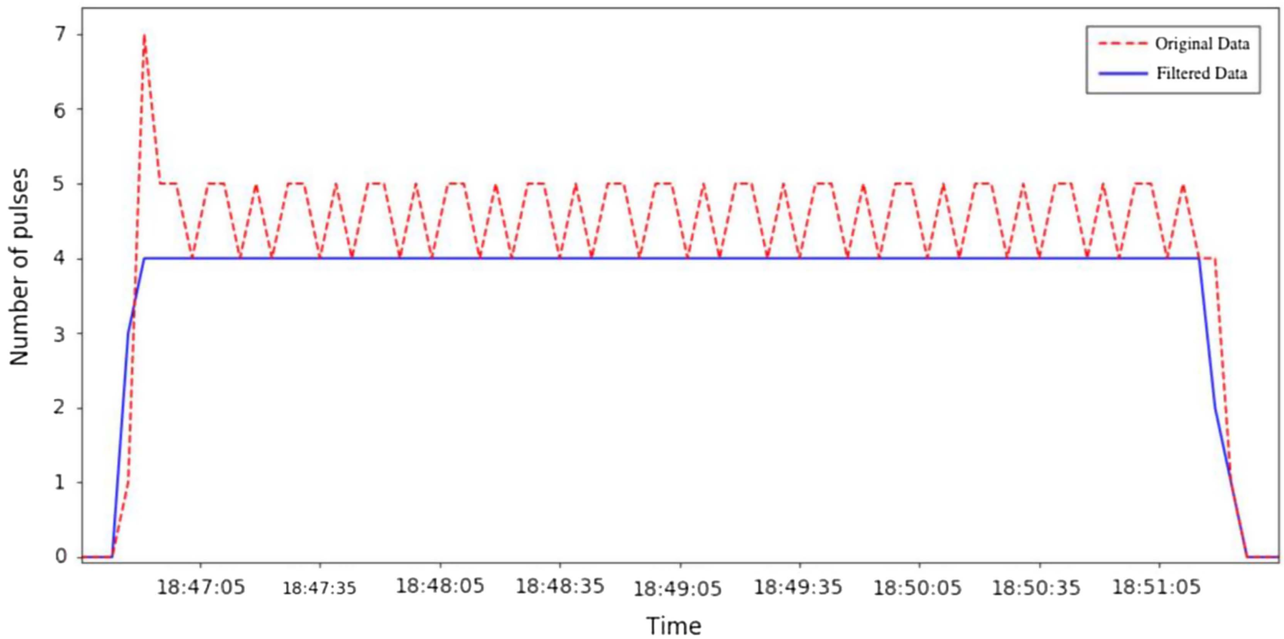

Fig. 2. Example of data filtering output for a single water-use event. Oscillations in the original data are caused by the combination of the data recording interval and pulse resolution of the water meter.

where RMS = square root of the mean square of water use flow rate within an event; FR_i = flow rate at the ith index; and n = number of flow-rate values within an event.

The disaggregation process then uses these features to classify events as single or overlapping and then to disaggregate overlapping events into single events. Before classifying events as single or overlapping, we first eliminated events with durations of one time step and volume of one pulse. Given uncertainty around the nature of these events, we did not classify them. We observed single-pulse, single-time-step faucet events in the set of events manually labeled by our study participant, but we also observed many other events of this type during time periods that were unlikely to be faucet events. Thus, where some other studies have assumed that these single-pulse events are associated with leaks, we were unable to do so given that they occurred in the manually labeled data. Instead, we categorized these events as unclassified and note that unclassified events may result from leaks and/or single time step events that we were unable to classify further.

Classifying events as single or overlapping was carried out using a combination of features from Table 3, including the number of vertices, Irregularity Measure I (IR1), Irregularity Measure II (IR2), and the complexity measure (CX), which are all calculated from the filtered data. These features reflect the shape, irregularity, and complexity of each event, and hence enable determination of whether an event is single or overlapping. Perfect, rectangular-shaped events have a zero value for IR1 (RMS = mode flow rate). As the irregularity of an event increases, RMS and mode flow-rate values become unequal and result in a nonzero value for IR1.

Table 3. Water end-use event features

Feature	Definition	Data used
Volume (L)	Summation of water-use volume between the beginning and end of a set of consecutive nonzero values	Raw
Start time	Time at which the water-use volume changes from zero to any positive value	Raw
End time	Time at which the water-use volume transitions back to zero	Raw
Duration (min)	Elapsed time of the event calculated as the difference between start and end times	Raw
Flow rate (L/min)	Volume of an event divided by its duration	Raw
Peak flow rate (L/min), original	Maximum rate of water flow for any time step within the event	Raw
Peak flow rate (L/min), filtered	Maximum rate of water flow for any time step within the event	Filtered
Peak flow rate frequency, original	Number of occurrences of the maximum flow rate within an event	Raw
Peak flow-rate frequency, filtered	Number of occurrences of the maximum flow rate within an event	Filtered
Mode flow rate (L/min)	Flow-rate value that appears most often within an event	Filtered
Mode flow-rate count	Number of occurrences of the mode flow rate within an event	Filtered
RMS	Square root of the mean square of flow rate (the arithmetic mean of the squares of the flow rates within the event)	Filtered
Number of vertices (V)	Number of points where the flow rate changes from one nonzero value to another nonzero value within the same event	Filtered
Irregularity Measure I (IR1) (L/min)	Difference between the mode flow rate and the RMS value	Filtered
Irregularity Measure II (IR2) (L/min)	Difference between peak flow rate and mode flow rate	Filtered
Complexity measure (CX)	Count of the mode flow rate divided by the number of vertices	Filtered

A constant-flow-rate, single-use event has a zero value for IR2 (peak flow rate = mode flow rate). When water use events overlap, the peak flow-rate value and the mode flow-rate value deviate from each other (IR2 becomes a nonzero value). We considered events to be irregular if both IR1 and IR2 were not equal to zero. Besides their irregular shapes, overlapping events are constituted by more than four vertices and have a CX value greater than 1. After calculating the complexity and the irregularity of each event originally identified in the trace, we identified overlapping and single-use events using the following criteria:

$$\text{Event} = \begin{cases} \text{Overlapping} & \text{if IR1} \neq 0 \text{ and IR2} \neq 0 \text{ and CX} > 1 \text{ and V} > 4 \\ \text{Single} & \text{all other events} \end{cases} \tag{3}$$

The criteria suggest that for an event to be considered overlapping, it must meet the IR1, IR2, CX, and V conditions all together. If one condition is not met, the event is considered a single-use event. Fig. 3 depicts the process of classifying events as single use or overlapping. Graphicly, most single-use events [Fig. 3(a)] are approximately rectangular and exhibit a constant flow rate throughout the entire event. IR1 and IR2 measures are zero for both events, and the CX measure is larger than 1, hence both events are single-use events. In Fig. 3(b) the event exhibits a variable flow rate throughout its entirety. In this case, both IR1 and IR2 measures do not equal to zero, and the event is confined by more than four vertices, but the event violates the CX measure, where the frequency of the mode flow divided by the number of different vertices is less than 1. Thus, the event is classified as a single-use event. Fig. 3(c) shows an overlapping event that satisfies the IR1, IR2, CX, and V conditions all together, having more than four vertices, values of IR1 and IR2 that do not equal zero, and a CX value that is larger than 1.

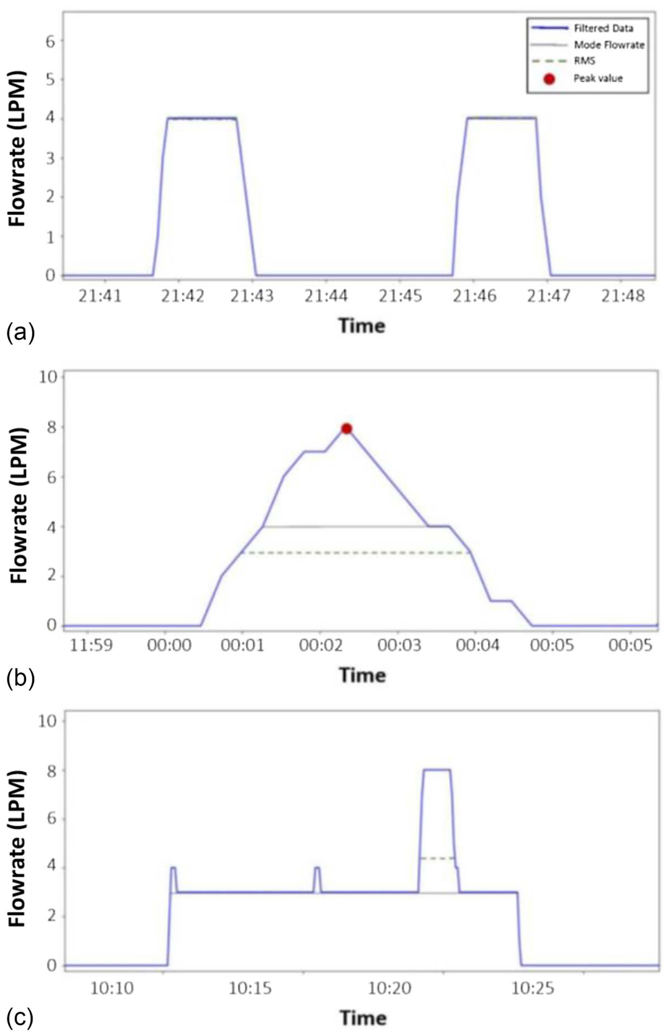

Fig. 3. Examples of single and overlapping events: (a) multiple single-use events; (b) an irregular single-use event; and (c) overlapping events.

Overlapping events identified in the first phase proceed to the second phase where they are split into their single-event components using a sequential splitting procedure (Fig. 4). The first split is applied horizontally at the mode flow rate. The first identified subevent is made up of any value that is less than or equal to the mode flow rate. After the first split, the remaining subevents are considered to be new events whose features are estimated to test whether the event is a single-use event. If yes, no more splitting is needed, and the event proceeds to the clustering process. If not, the splitting procedure at the mode flow rate continues in an iterative manner until all subevents are single-use events.

In Fig. 4, the first split is applied on the mode flow-rate value (3 L/min). Values that are equal to or less than 3 are added together to form Event 1. After the first split, the features of the newly separated subevents are recalculated and used to reevaluate the complexity and irregularity of each subevent. The features of Events 2 and 3 suggest they are both single events, hence no further splitting is performed on these two events. However, Event O1 is an overlapping event and can be further simplified. We applied the second split on the mode flow rate of Event O1 and reevaluated the complexity and irregularity of the resulting subevents. This time, their features suggest that Events 4 and 5 are single-use events. The final split is applied to the last overlapping event in the series, which contains Events 6 and 7. This process is run on the data set for each house, and the result of this iterative end-use disaggregation process is a data set of single-use events associated with their features described in Table 3 for each house. Creating an event data set for each house maintains consistency in end-use types, water-use behavior, and the statistical distribution of water-use events of each data set. We acknowledge that there may be a small number of overlapping events where the first event ends before the second event ends. In these cases, the algorithm will split the second event into two separate events. The separated events will still be classified correctly, but the count of events will include an extra event for each of these instances.

Event Classification

Each single end-use event output by the disaggregation process was assigned an end-use category using a semisupervised clustering and classification technique. We used clustering to identify outliers in the event data set and classification to assign each nonoutlier event

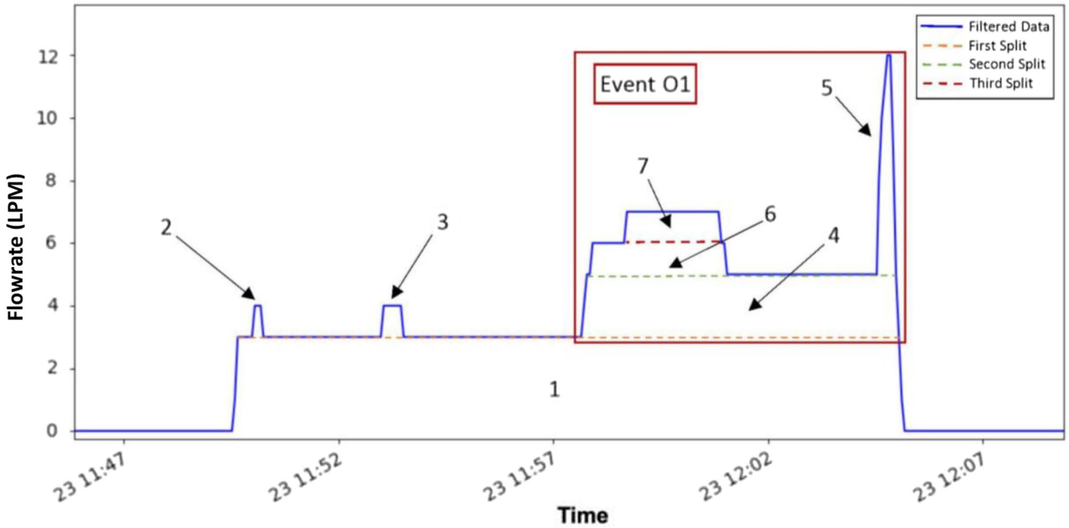

Fig. 4. Splitting procedure for overlapping events.

to an end-use category. This process was executed on the event data sets for each house, one house at a time.

Feature Scaling

The numeric values of the features listed in Table 3 are highly variable in units and range. Considering that most clustering algorithms use distance between data points (e.g., the Euclidean distance in a multidimensional space) as the measure of similarity, features with varying magnitudes will not be weighted equally in the distance calculations. To overcome this, we applied feature scaling prior to clustering. We explored the distributions for all features and observed that each exhibited a multimodal distribution (e.g., Fig. 5).

Because the feature distributions are not Gaussian and we observed outlier data points in the distributions for some of the features, we used the RobustScaler function from the scikit-learn package for Python to scale the range and the magnitude of all features. The RobustScaler function scales the feature values by subtracting the median from each data point and then by dividing by the interquartile range (IQR) of the data. Mathematically, the RobustScaler can be expressed as

$$x_i' = \frac{x_i - Q_2(x)}{Q_3(x) - Q_1(x)} \tag{4}$$

where x_i' = scaled data point; x_i = original data point; $Q_1(x)$ = original first quartile data point; $Q_2(x)$ = second quartile data point (median); and $Q_3(x)$ = original third quartile data point. The output of this step in the process is a data set of events for each house with scaled feature values.

Feature Selection

Including redundant or irrelevant features may inhibit clustering algorithm performance and can misguide results. To avoid this, we utilized a feature selection technique to subset a feature combination that produced the "best" clusters where the intracluster similarity was high and the intercluster similarity was low. There are several algorithms for selecting features, including recursive feature elimination (RFE) (Guyon et al. 2002), Boruta (Kursa et al. 2010), and genetic algorithms (Whitley 1994). We used the Boruta algorithm because it was found to be more robust than other feature selection algorithms in a study that evaluated different feature selection methods (Degenhardt et al. 2017).

The Boruta algorithm uses a random forest–based classifier iteratively and attempts to capture all features that are relevant to the outcome variable (i.e., the predicted end-use labels for each event). The algorithm determines which features to keep and which to eliminate by first creating shadow features that are duplicates of the original features. After the shadow features are created, their values are shuffled to remove any potential correlation with the outcome variable. Next, the algorithm combines the original features with the shadow features into one data set and runs the random forest classifier on the combined data set multiple times (the default is 20). Boruta calculates an importance score for each feature in the combined data set as the decrease in the classification accuracy of the outcome variable when a feature is excluded (mean decrease in accuracy of the outcome variable). The importance score for each original feature in the data set is then compared with a threshold value, defined as the highest importance score among the shadow features. If the feature's importance score is higher than the threshold value, the feature scores a hit. If not, the feature scores a no hit. With each run, each feature is removed one at a time and its importance score is calculated. With the 20 run results, a feature that scores a hit in at least 95% of the total runs is deemed important. If the feature's hit score is lower than 95% of the total runs, the feature is considered irrelevant and is eliminated from the data set.

We used the implementation of the Boruta algorithm included in the scikit-learn contributed packages (Homola 2015) on the manually labeled event data set from Household 5. For the algorithm configuration, we defined the event label attribute as the outcome variable and the scaled features in Table 3 as the predictor variables. We used the default number of runs setting (20). The algorithm suggested that the mode flow rate, duration, volume, peak flow rate, average flow rate, and RMS features have the most impact in predicting the correct labels for events. Fig. 6 summarizes the output of the algorithm and the importance score for each feature. The color of the box in Fig. 6 indicates feature type. The final six boxes on the figure are original features, while the first three boxes are shadow features. Features with an importance score less than the maximum shadow feature score do not appear in the plot (e.g., CX). The output of the feature selection process is

J. Water Resour. Plann. Manage., 2023, 149(7): 04023024

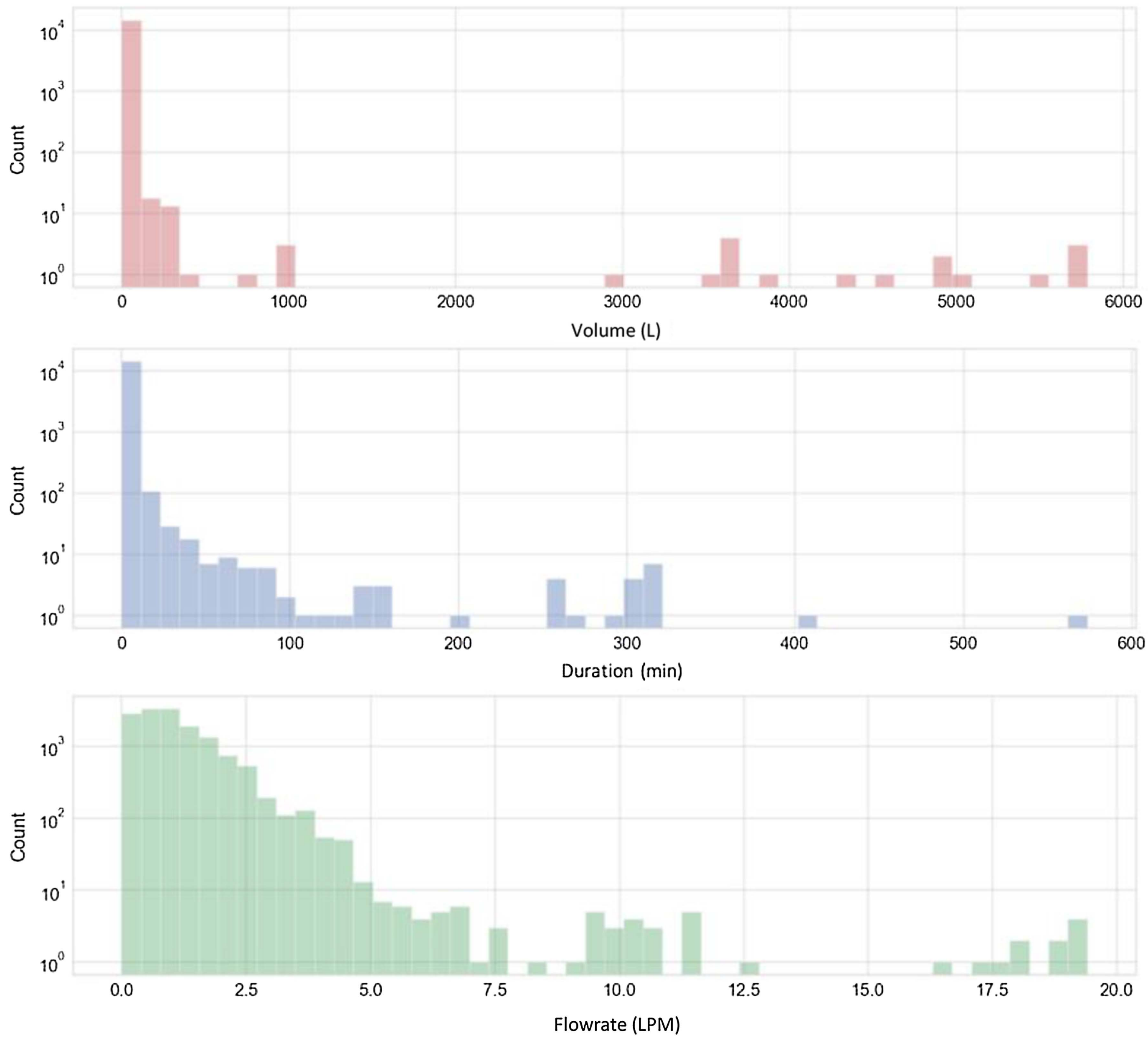

Fig. 5. Distributions of volume, duration, and flow-rate features. Due to the high discrepancies in the frequency of events across the feature values, log-scale y-axes were used to improve readability.

a data set of events for each house with a reduced set of features that can be used in the classification process.

Semisupervised Classification

The final step in the classification process is clustering and label assigning, for which we used a cluster-then-label semisupervised technique. This technique utilizes unsupervised clustering to detect and eliminate outlier events that deviate from other events in the data space and then predicts the category label of each remaining unlabeled event in the developed clusters using semisupervised classification. Similar approaches have been used by Tanha et al. (2015), Gan et al. (2013), and Weston et al. (2005). Due to the peculiar variation in the features of end-use events of the same type caused by residents' water-use behavior (e.g., long versus short showers, faucet partially versus fully open), clusters of events may not have convex, isotropic shapes, which can be problematic for partition-based and hierarchical clustering techniques. Density-based clustering, on the other hand, identifies clusters as groups of data points of high density and is capable of identifying clusters of any shape. On account of this, we used the density-based spatial clustering of applications with noise (DBSCAN) clustering algorithm (Keim and Kriegel 1996).

DBSCAN has two types of outputs: (1) outliers (scattered low-density events that do not fall within an identified cluster), and (2) event clusters (definitive clusters made up of high-density core and border events that include only one type of end use). DBSCAN was run on the event data set consisting of the subset of scaled event features selected by the Boruta algorithm to extract event clusters for each house. We observed that outliers were typically abnormal water-use events (e.g., a very long shower or a dual toilet flush) that exhibit features that are different from anticipated behavior. Outliers identified by DBSCAN were added to the list of events

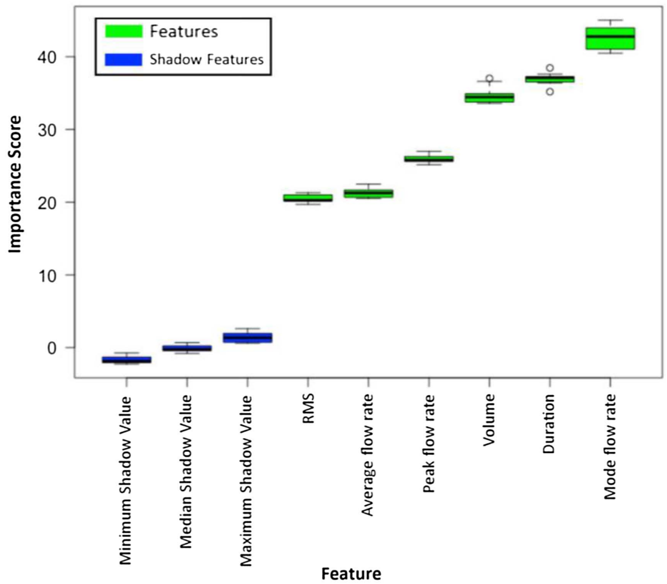

Fig. 6. Boruta algorithm output. Box plots represent the distribution of importance scores over the 20 Boruta runs.

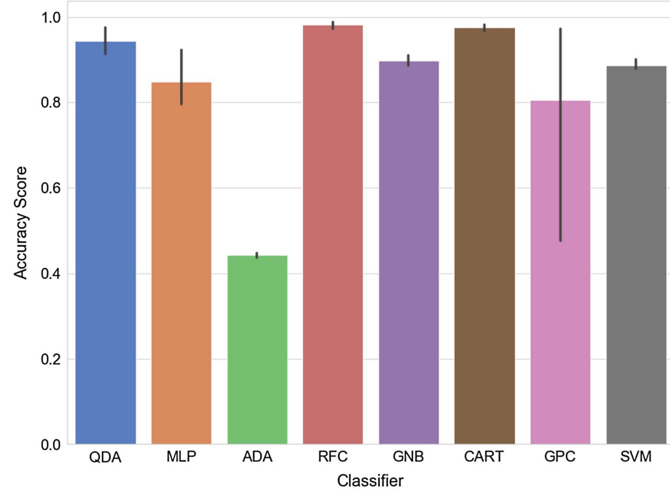

Fig. 7. Accuracy of different supervised classifiers. Error bars represent the standard deviation of the 10 accuracy score values.

that we did not classify, whereas events that were assigned to a specified cluster proceeded to the final step in the procedure, label assigning.

For assigning labels to unlabeled events, we tested several supervised classifiers in the scikit-learn package on the manually labeled training data set, including quadratic discrimination analysis (QDA), multilayer perceptron (MLP), adaptive boost (ADA), random forest (RFC), Gaussian naive Bayes (GNB), classification and decision tree (CART), Gaussian process (GPC), and support vector machine (SVM). A description of each of these classifiers can be found in the scikit-learn documentation (Scikit 2007). Given the limited nature of our manually labeled sample of events, we used k-fold cross-validation analysis to test the accuracy of each classifier (Camacho and Ferrer 2012). The k-fold method splits the data set into a number (k) of subsets (folds). Each data point is assigned to an individual fold and stays in that fold for the duration of the procedure. To begin the procedure, one fold is reserved for testing (called the holdout), and the rest of the folds are combined and used as a training data set for fitting a model. The accuracy of the fitted model is then evaluated on the data in the holdout fold. The accuracy score of the model is calculated as the number of correctly predicted data points divided by the total number of data points within the holdout fold. The accuracy score is then retained, while the model is discarded. This process is repeated k times, each time using a different holdout fold. Thus, each fold serves as the holdout fold once and is used as part of the training data set to fit a model $k-1$ times. For k-fold configuration, we used 10 folds with each fold holding 10% of the total data.

The average accuracy score of each tested classifier was calculated as the summation of accuracy scores for the 10 models developed using that classifier divided by the number of folds (10). By doing this, we were able to create many different models using each of the classifiers and test them on different subsets of the training data rather than doing a single evaluation on one testing data set. Consistently high accuracy scores across all folds indicate that the classifier performs well, and we can be confident that the trained algorithm will always produce a similar performance. We applied k-fold cross validation to all of the classifiers mentioned and evaluated the average and standard deviation of their accuracy scores. We found that the random forest classifier produced the highest accuracy score and selected it for use, although it was only slightly better than the decision tree classifier (Fig. 7). Both classifiers performed consistently across all tested folds as indicated by their small standard deviation values.

We used the random forest classifier on the manually labeled training data set to develop an initial model that was the same for all of the houses. We then implemented an iterative, self-training procedure to develop a final model for each house. The initial random forest classifier was used to predict the labels for all of a home's unlabeled events. We used the similarity score metric function in the random forest classifier to quantify the quality of the predicted label for each event in the home's data set. The similarity score function uses the coefficient of determination (R^2) of the newly labeled events to estimate the probability of each newly labeled event being classified correctly when compared with the manually labeled events. On each iteration, a subset of the newly labeled events with a similarity score of at least 90%, together with their predicted labels, were added to the labeled data set. The random forest classifier was then retrained on the larger set of labeled events to produce a new model for the home. This procedure was repeated, and the model was iteratively trained until all events with similarity scores of at least 90% with their predicted labels were added to the set of labeled events. The enhanced model based on the original and newly labeled events was then used to predict the labels of the remaining unlabeled events. Each household required two to three iterations to reach a point where no new events met the 90% similarity score criterion. The pseudo-code of the algorithm can be formally described as follows:

Let $\mathbf{L} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{l}$ be the input manually labeled training data set where x is the set of labeled events, y is the label set, and l is the number of labeled events. Let $\mathbf{U} = \{\boldsymbol{x}_j\}_{j=l+1}^{l+u}$ be the set of u unlabeled events for a home.

1. Apply DBSCAN on $\mathbf{U}$ to generate an initial set of clusters
2. Using the initial DBSCAN clusters, identify outlier data points and remove them from $\mathbf{U}$
3. Train a predictive random forest classifier $\boldsymbol{f}$ using $\mathbf{L}$ as a training data set
4. Predict the label of all unlabeled events in $\mathbf{U}$
5. Calculate the similarity score for the newly labeled events in $\mathbf{U}$
6. Add $\{(\boldsymbol{x}, \boldsymbol{f}(\boldsymbol{x}))|\boldsymbol{x} \in \mathbf{S}\}$ to $\mathbf{L}$, where $\mathbf{S}$ is the subset of newly labeled events in $\mathbf{U}$ with similarity score > 90%

7. Remove **S** from **U**
8. Retrain the classifier $\boldsymbol{f}$ using the enhanced training data set **L**
9. Repeat Steps 4–8 until there is no $(\boldsymbol{x}, \boldsymbol{f}(\boldsymbol{x}))$ with similarity score $> 90\%$ in **U**, or until $\mathbf{U} = \emptyset$

The output of this process is a random forest classifier model for a home that can be applied to the set of unclassified events for that home to predict their event-type labels. We used the manually labeled testing data set for testing the accuracy of the final model developed for Household 5.

Software Design and Implementation

The tool was designed and developed using the Python programming language as a single script that can be executed using any Python programming environment. The tool was developed using the SciPy, Pandas, NumPy, and scikit-learn packages for Python. The input to the tool is a CSV file that contains high-resolution meter data collected every 4 s for an individual home and a pulse resolution conversion factor that corresponds to the size of the meter (Bastidas Pacheco et al. 2020). The input file has three fields: time, record, and pulses. The time field contains the time stamps at which individual observations of water use were recorded. The record field is a sequential numbering attribute that uniquely identifies rows in the data set. The pulses field contains the number of magnetic pulses recorded by the data logger in a 4-s interval, where each pulse corresponds to a known volume of water use. The first three rows of the file are reserved for a metadata header including the site number, datalogger ID, and meter size. Data start on the fourth row. The length of the file is not restricted.

The tool reads the data file and loads it into a date- and time-indexed Pandas data frame, after which the filtering process on raw water use data is applied. The disaggregation process outputs four different CSV files. The first file contains single water-use events, the second contains unclassified water-use events, the third contains overlapping water-use events, and the fourth contains the disaggregated single water-use subevents derived from overlapping events. All output files from the disaggregation process include values for the features listed in Table 3. The tool then proceeds with the classification process. The classification process outputs one CSV file that contains classified and labeled water-use events along with the features of each event, including duration, volume, flow rate, peak flow rate, and mode flow rate.

Results and Discussion

After data cleansing and removal of single-pulse events, the tool identified 16,420 unprocessed water-use events retrieved from the five households (excluding unclassified events). After breaking down overlapping events into their single-use-event components, the total number of events increased to 18,491. The average processing time for all filtering, disaggregation, and classification operations per single day of data for one household was approximately 50 s. Unclassified events (including single-pulse events) accounted for only 3% of the total water-use volume across the five houses; however, these events appeared frequently in the trace data and accounted for more than 40% of all recorded events in some households. Given the uncertainty around the nature of unclassified events and their relatively small overall volume, we excluded them from further analysis.

End-Use Classification Accuracy

We used the testing data set consisting of manually labeled events from Household 5 to quantify the classification accuracy of the developed model at Household 5. For consistency with other end-use classification studies, we calculated model accuracy as the fraction of events whose end-use category was correctly predicted by the model when compared with the manual labels (Table 4). The overall accuracy of the classification process was 98.2%, with the highest accuracy observed (100%) for clothes washer and shower events. These results represent a significant improvement compared with the 88.4 faucet accuracy reported by Autoflow v3.1 (Yang et al. 2018). We also used the F1 score metric to quantify the model accuracy (Table 4). The weighted average F1 score for all end-use categories was 98%, where weights were assigned based on the number of events in each end-use category. The disparity between the classification accuracy for bathtub events (67%) versus the F1 score metric (80%) indicates that the model may not be finding all bathtub events, but that those events classified by the model as bathtub events are likely to be correct. We also illustrate classification results using a confusion matrix (Fig. 8), which is a summary of predicted labels for water end-use events compared with actual labels. The diagonal elements represent the number of events for which the predicted label is equal to the true label, while off-diagonal elements represent misclassified events. As shown in Fig. 8, all clothes washer and shower events were classified correctly. The classification accuracy for bathtub events was lower than the other event types. Of the bathtub events in the testing data

Table 4. Classification accuracy and F1 score measures for each end-use type

End-use type	Number of events	Classification accuracy (%)	F1 scores (%)
Bathtub	15	66.7	80
Clothes washer	24	100	98
Faucet and dishwasher	190	98.9	99
Shower	60	100	96
Toilet	171	99.4	99
All categories	460	98.2	98

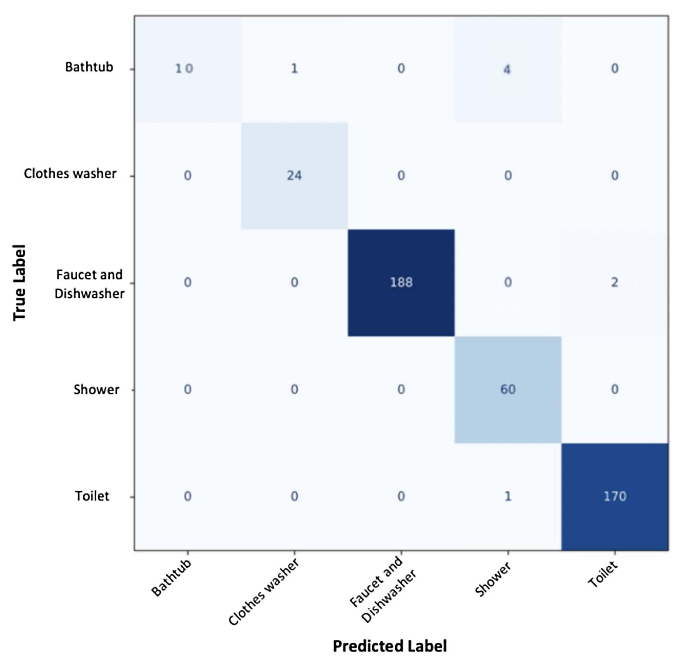

Fig. 8. Confusion matrix results for each end-use type.

set that were incorrectly labeled, 80% (four out of five) were classified as shower events. We attribute this to similarity between the characteristics of bathtub and shower events and the small number of bathtub events in the manually labeled training data set.

For sites where no manually labeled events were available, it was not possible to directly evaluate the accuracy of the classifier. Instead, we applied a manual verification procedure consisting of examining the characteristics of extracted events and their raw trace data for each site. While the characteristics of water fixtures may vary from one household to another, the data show that their overall characteristics are relatively close (e.g., a toilet flush in one house might have an average volume around 6 L/flus versus 7 L/flush in a different house). Events extracted from all homes were compiled into one file. Then, for each home, events within each end-use type were sorted according to their features in a step-by-step procedure (e.g., sorting toilet events for a home by their volumes in descending order and then ascending order). We then investigated differences between events at the end points and events in the middle of the distribution of each end-use type for each home. We assumed that mislabeled events are more likely to occur at the end points of the feature distributions of each end-use type. The number of events we examined for each end-use type varied depending on how similar events at the end points of the distribution for each end-use type compared with the rest of events. We examined between 20 and 50 events per end-use type per house. Besides the examination of extracted classified events, we also examined the raw trace data for the same events to verify similarities with other events having the same label. As a last step in the verification procedure, we relabeled misclassified water events according to where they were most likely to belong, based on the analyst's decision, considering all the elements described previously. Without considering unclassified events, on average, changes were made to less than 2% of the labels assigned by the algorithm at each site.

The information collected at each home during enrollment helped us confirm the verification process and catch mislabeled events (e.g., the algorithm labeling bathtub events that were similar to clothes washer events in homes where bathtubs were not present or used). We retrained these homes using a training data set that did not include bathtub events. The developed method seeks to classify water end-use events without the need for labeled events at each home. Thus, the manual evaluations we did were aimed at ensuring the quality of our analyses.

Although the accuracies we observed were similar to those reported by Yang et al. (2018), it should be noted that the accuracy values are not exactly comparable. We estimated the classification accuracy of the tool using a relatively small set of events that were manually labeled by a study participant. Thus, we were confident in the event labels. The Autoflow classification accuracy was estimated based on a much larger set of events that were labeled by an analyst using the Trace Wizard software (Yang et al. 2018). While we cannot verify the actual accuracy of the labels assigned by the analyst, the Autoflow software was able to match those labels with a high level of accuracy. Using the set of manually labeled events, we were able to identify the features of dishwasher events. However, we found that their features were indistinguishable from faucet events, so we were not able to separate dishwater and faucet events. Thus, dishwasher events were grouped with faucet events (see Appendix S2 for more details).

Overlapping Events

A similar verification procedure was applied to overlapping events, where we manually verified the accuracy of our method in identifying and separating overlapping events by visually inspecting events collected at different sites. A total of 515 of the original 16,420 events we extracted from the water trace data for five households covering a monitoring period of 4 weeks per household were identified as overlapping events. The overlapping events were disaggregated into 2,071 single events, bringing the total number of single events to 18,491. We visually inspected 20% of all overlapping events identified by the algorithm and verified that they all met the algorithm's criteria for overlapping events and that the splits were applied correctly. Thus, we are confident that the algorithm is correctly identifying overlapping events and that the splitting procedure works as designed.

Overall Water Use and Individual End Uses

For completeness and for comparison with other studies, we examined the indoor and outdoor water use for each of the studied households. We also characterized individual end uses for each home to quantify the distribution of volume and frequency of use across the different end uses along with potential seasonal variation. These analyses are provided in Appendix S1 for overall water use and Appendix S2 for detailed end-use analysis.

Conclusions

We presented a new, open-source, semisupervised water end-use disaggregation and classification tool that can break down the total water use observed at the household level into different end uses. The tool uses nonintrusive monitoring data collected at high temporal resolution from a residential home's water meter along with machine learning techniques to disaggregate water use into discrete end-use events. This work was driven by the fact that, for most other studies that have worked on end-use disaggregation algorithms, neither the source code nor the data are available for testing or further advancement. It is our hope that the code and anonymized data we have openly shared can be a platform for advancing the availability and functionality of open tools for water end-use disaggregation studies.

Unlike other end-use disaggregation techniques, we used a semisupervised classification approach to overcome the challenges associated with classifying events. Although our approach required an initial set of manually labeled events, which can be expensive and difficult to collect, we employed this relatively small number of labeled events to show how a semisupervised model can be developed and used for classifying events from any residential home. The data we collected and the developed water end-use disaggregation tool are now available for potential use by others who may want to use similar end-use classification approaches. Additionally, where some other studies validated their results by comparing with events classified in postprocessing by a data technician (which may or may not be correct depending on the data set and experience level of the technician), our approach used actual events manually labeled by a study participant. The openly available data set of manually labeled events paired with the corresponding high-resolution water-use data from the meter that we produced could be expanded by other investigators in other geographical areas to produce a much larger set of labeled data to produce a large corpus of data for training machine learning models as has been done in other fields of study [e.g., the ImageNet data set (Deng et al. 2009)]. Our intermediate results (e.g., the disaggregated events and their features) could also be repurposed for testing other clustering or classification techniques.

In our case study application, the number of events participating in overlapping events accounted for approximately 11% of all recorded events and contributed to about 28% of total water-use volume, demonstrating the importance of handling them correctly in the disaggregation process. Our disaggregation approach extends what has been done in other studies and enabled us to separate the overlapping events into single events prior to classification. Executing the disaggregation and classification algorithms on data from different households with different meter sizes and types and different water-use characteristics verified the reliability of the algorithm across the meter types and sizes we tested, which should mean it can be used across a wide range of residential meter types and sizes, although further testing with new data sets would be needed to confirm this. While we manually verified that the algorithm correctly identified overlapping events according to our criteria and separated them according to the rules we set, a new study with data collection focused specifically on recording and labeling overlapping events could provide benchmark data sets for further testing the accuracy of algorithms for disaggregating overlapping events.

The tool provides significant benefits for water consumers and water utilities. For consumers, it can provide information about how and when they are consuming water. Our detailed results illustrate both behavioral differences in water use across households and technical differences based on the performance of the water fixtures within the homes, which we found were not always meeting plumbing standards. A compilation of household-level water end-use data could assist water utilities in identifying opportunities for incentive programs to encourage water conservation and monitoring effectiveness of those programs. For researchers, the open nature of the data collection hardware and the methods described in this paper for end-use disaggregation present a new opportunity for advancing beyond the limitations imposed by lack of available data and the proprietary nature of existing software. The code we have provided for analyzing the disaggregated water-use data can serve as a base for further work.

The work described in this paper builds on other end-use disaggregation studies and, like those other studies, demonstrates how water end-use studies can provide detailed data to inform water resource management in areas where water is scarce. By characterizing how residential water is utilized in households, these results provide information that may be useful for city engineers and local water managers as they operate existing infrastructure and formulate plans to increase the efficiency of current water supply and distribution infrastructure, as well as in planning for future improvements. Indeed, understanding water use at the end-use level is essential for gaining insight into how, when, and why water is being used. This information is, in turn, critical for water managers in identifying opportunities for conservation, assessing the impact of conservation programs, forecasting demand, and determining how water-use patterns may change over time in response to population growth, demographic shifts, and improvements in technology. Supplying this type of information to water users can also be a tool for impacting water-use behavior and managing demand.

Detailed information about water use is needed by Utah, other states in the US, and other similar areas throughout the world to better project future water needs. High-resolution data from metered households can provide valuable information on daily and seasonal consumption patterns, especially when coupled with both structural (e.g., lot size and landscaping characteristics, appliance and fixture age) and sociodemographic information (e.g., age, family size, income level, ethnicity) about the household. Future water-use projections can then be made based on the demographics of projected growth and not just on the projected number of people. In addition to more accurate demand forecasting, these data provide a potential opportunity for water utilities to reduce operational costs now and in the future through efficiency gains and deferral of upgrades.

Data Availability Statement

Some or all data, models, or code generated or used during the study are available in a repository online in accordance with funder data retention policies. The CIWS Disaggregator software is open source, released under the Creative Commons Attribution CC BY license, and available in the HydroShare repository (Attallah and Bastidas Pacheco 2023). Documentation of hardware and software requirements, Jupyter notebooks with code examples, and instructions for running the code are provided in the HydroShare resource. The high-resolution water-use data set containing the data for all 31 houses we sampled is available in HydroShare (Bastidas Pacheco et al. 2023). The manually labeled event data set and the processed event data resulting from the case study analyses in this paper are also available in HydroShare (Attallah and Bastidas Pacheco 2023). While we anticipate that other geographical areas may have different meter configurations, we anticipate that the disaggregation and classification procedure described in this paper can be applied to high-resolution metering data collected anywhere. Specific instructions for implementing the code on our data set (or other similar data sets) are provided in HydroShare (Attallah and Bastidas Pacheco 2023).

Reproducible Results

Amber Spackman Jones (Utah State University) downloaded the CIWS Disaggregator code and input data set from HydroShare (Attallah and Bastidas Pacheco 2023). She installed the code using instructions available in the HydroShare resource and ran the CIWS Disaggregator using the input data set provided in the HydroShare resource to reproduced results in Figs. S1 and S2 and Tables S1–S3 in Appendix S1, and Figs. S3–S7 and Tables S4–S8 in Appendix S2.

James Stagge (Ohio State University) downloaded, installed, and ran the code using the input data set and reproduced results in Figs. S1 and S2 and Tables S1–S3.

Acknowledgments

This research was funded by the United States National Science Foundation under Grant No. 1552444. Any opinions, findings, and conclusions or recommendations expressed are those of the authors and do not necessarily reflect the views of the National Science Foundation. Additional financial support was provided by the Utah Water Research Laboratory at Utah State University. We acknowledge the cities of Logan and Providence for their cooperation in the realization of the field data collection campaigns. We also gratefully acknowledge the homeowners who participated in our data collection efforts.

Supplemental Materials

Appendixes S1 and S2, including Figs. S1–S7 and Tables S1–S8, are available online in the ASCE Library (www.ascelibrary.org).

References

Anda, M., and J. Temmen. 2014. "Smart metering for residential energy efficiency: The use of community based social marketing for behavioural change and smart grid introduction." *Renewable Energy* 67 (Jul): 119–127. https://doi.org/10.1016/j.renene.2013.11.020.

Attallah, N., and C. J. Bastidas Pacheco. 2023. “Supporting data and tools for ‘an open-source, semi-supervised water end use disaggregation and classification tool’.” *HydroShare*. https://doi.org/10.4211/hs.3143b3b1bdff48e0aaebcb4aedf02feb.

Bastidas Pacheco, C. J., N. Atallah, and J. S. Horsburgh. 2023. “High resolution residential water use data in Cache County, Utah, USA.” *HydroShare*. https://doi.org/10.4211/hs.0b72cddfc51c45b188e0e6cd8927227e.

Bastidas Pacheco, C. J., J. S. Horsburgh, and R. J. Tracy. 2020. “A low-cost, open source monitoring system for collecting high temporal resolution water use data on magnetically driven residential water meters.” *Sensors* 20 (13): 3655. https://doi.org/10.3390/s20133655.

Beal, C., R. A. Stewart, A. Spinks, and K. Fielding. 2011. “Using smart meters to identify social and technological impacts on residential water consumption.” *Water Supply* 11 (5): 527–533. https://doi.org/10.2166/ws.2011.088.

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

Broesch, J. D. 2008. *Digital signal processing: Instant access*. 1st ed. Burlington, MA: Newnes.

Camacho, J., and A. Ferrer. 2012. “Cross-validation in PCA models with the element-wise k-fold (EKF) algorithm: Theoretical aspects.” *J. Chemom.* 26 (7): 361–373. https://doi.org/10.1002/cem.2440.

Cardell-Oliver, R. 2013. “Water use signature patterns for analyzing household consumption using medium resolution meter data.” *Water Resour. Res.* 49 (12): 8589–8599. https://doi.org/10.1002/2013WR014458.

Chapelle, O., and A. Zien. 2005. “Semi-supervised classification by low density separation.” In *Proc., 10th Int. Workshop on Artificial Intelligence and Statistics*, 57–64. Cambridge, MA: Proceedings of Machine Learning Research.

Chen, W. 2014. “The regularized low pass filter.” *J. Signal Inf. Process.* 5 (1): 14–16. https://doi.org/10.4236/jsip.2014.51003.

Cominola, A., M. Giuliani, A. Castelletti, P. Fraternali, S. L. H. Gonzalez, J. C. G. Herrero, J. Novak, and A. E. Rizzoli. 2021. “Long-term water conservation is fostered by smart meter-based feedback and digital user engagement.” *npj Clean Water* 4 (1): 29. https://doi.org/10.1038/s41545-021-00119-0.

Cominola, A., M. Giuliani, D. Piga, A. Castelletti, and A. Rizzoli. 2017. “A hybrid signature-based iterative disaggregation algorithm for non-intrusive load monitoring.” *Appl. Energy* 185 (Jan): 331–344. https://doi.org/10.1016/j.apenergy.2016.10.040.

Cominola, A., M. Giuliani, D. Piga, A. Castelletti, and A. E. Rizzoli. 2015. “Benefits and challenges of using smart meters for advancing residential water demand modeling and management: A review.” *Environ. Modell. Software* 72 (Oct): 198–214. https://doi.org/10.1016/j.envsoft.2015.07.012.

Contestabile, M. 2018. “Fresh water input.” *Nat. Sustainability* 1 (3): 119. https://doi.org/10.1038/s41893-018-0042-z.

Coppack, S. W. 1990. “Correspondence section: Limitations of the Pearson product-moment correlation.” *Clin. Sci.* 79 (3): 287. https://doi.org/10.1042/cs0790287.

Degenhardt, F., S. Seifert, and S. Szymczak. 2017. “Evaluation of variable selection methods for random forests and omics data sets.” *Brief Bioinf.* 20 (2): 492–503. https://doi.org/10.1093/bib/bbx124.

Deng, J., W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei. 2009. “ImageNet: A large-scale hierarchical image database.” In *Proc., 2009 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. New York: IEEE.

DeOreo, W. B., J. P. Heaney, and P. W. Mayer. 1996. “Flow trace analysis to access water use.” *J. Am. Water Works Assn.* 88 (1): 79–90. https://doi.org/10.1002/j.1551-8833.1996.tb06487.x.

DeOreo, W. B., P. W. Mayer, B. Dziegielewski, and J. C. Kiefer. 2016. *Residential end uses of water, Version 2*. Denver: Water Research Foundation.

Di Mauro, A., A. Cominola, A. Castelletti, and A. Di Nardo. 2020. “Urban water consumption at multiple spatial and temporal scales. A review of existing datasets.” *Water* 13 (1): 36. https://doi.org/10.3390/w13010036.

Doucet, A., S. Godsill, and C. Andrieu. 2000. “On sequential Monte Carlo sampling methods for Bayesian filtering.” *Stat. Comput.* 10 (3): 197–208. https://doi.org/10.1023/A:1008935410038.

Flandrin, P., G. Rilling, and P. Goncalves. 2004. “Empirical mode decomposition as a filter bank.” *IEEE Signal Process Lett.* 11 (2): 112–114. https://doi.org/10.1109/LSP.2003.821662.

Froehlich, J., E. Larson, S. Gupta, G. Cohn, M. Reynolds, and S. Patel. 2011. “Disaggregated end-use energy sensing for the Smart Grid.” *IEEE Pervasive Comput.* 10 (1): 28–39. https://doi.org/10.1109/MPRV.2010.74.

Froehlich, J. E., E. Larson, T. Campbell, C. Haggerty, J. Fogarty, and N. S. Patel. 2009. “HydroSense: Infrastructure-mediated single-point sensing of whole-home water activity.” In *Proc., 11th Int. Conf. on Ubiquitous Computing*, 1053–1056. New York: Association for Computing Machinery. https://doi.org/10.1145/1620545.1620581.

Gan, H., N. Sang, R. Huang, X. Tong, and Z. Dan. 2013. “Using clustering analysis to improve semi-supervised classification.” *Neurocomputing* 101 (Feb): 290–298. https://doi.org/10.1016/j.neucom.2012.08.020.

Guyon, I., J. Weston, S. Barnhill, and V. Vapnik. 2002. “Gene selection for cancer classification using support vector machines.” *Mach. Learn.* 46 (1/3): 389–422. https://doi.org/10.1023/A:1012487302797.

Homola, D. 2015. “BorutaPy: An all relevant feature selection method based on Random Forest estimators.” Accessed December 13, 2020. https://danielhomola.com/feature%20selection/phd/borutapy-an-all-relevant-feature-selection-method/.

Keim, D. A., and H.-P. Kriegel. 1996. “Visualization techniques for mining large databases: A comparison.” *IEEE Trans. Knowl. Data Eng.* 8 (6): 923–938. https://doi.org/10.1109/69.553159.

Kowalski, M., and D. Marshallsay. 2003. “A system for improved assessment of domestic water use components.” In *Proc., Efficient 2003–2nd Int. Conf. on Efficient Use and Management of Urban Water Supply.* London: International Water Association.

Kowalski, M., and D. Marshallsay. 2005. “Using measured microcomponent data to model the impact of water conservation strategies on the diurnal consumption profile.” *Water Supply* 5 (3–4): 145–150. https://doi.org/10.2166/ws.2005.0094.

Kursa, M. B., A. Jankowski, and W. R. Rudnicki. 2010. “Boruta—A system for feature selection.” *Fundam. Inform.* 101 (4): 271–285. https://doi.org/10.3233/FI-2010-288.

Längkvist, M., L. Karlsson, and A. Loutfi. 2014. “A review of unsupervised feature learning and deep learning for time-series modeling.” *Pattern Recognit. Lett.* 42 (Jun): 11–24. https://doi.org/10.1016/j.patrec.2014.01.008.

Mayer, P. W., and W. B. DeOreo. 1999. *Residential end uses of water.* Denver: American Water Works Association Research Foundation.

Mayer, P. W., W. B. DeOreo, E. Towler, L. Martien, and D. Lewis. 2004. *Tampa Water Department residential water conservation study: The impacts of high efficiency plumbing fixture retrofits in single-family homes*. Rep. Prepared for Tampa Water Department and the United States Environmental Protection Agency. Boulder, CO: Aquacraft.

Nayak, M. R., W. J. Emery, and R. E. Thomson. 1999. “Data analysis methods in physical oceanography.” *Estuaries* 22 (3): 728–730. https://doi.org/10.2307/1353059.

Nguyen, K. A., R. A. Stewart, and H. Zhang. 2013. “An intelligent pattern recognition model to automate the categorisation of residential water end-use events.” *Environ. Modell. Software* 47 (Sep): 108–127. https://doi.org/10.1016/j.envsoft.2013.05.002.

Pastor-Jabaloyes, L., F. J. Arregui, and R. Cobacho. 2018. “Water end use disaggregation based on soft computing techniques.” *Water* 10 (1): 46. https://doi.org/10.3390/w10010046.

Scikit. 2007. “User guide.” Accessed March 1, 2021. https://scikit-learn.org/stable/user_guide.html.

Sønderlund, A. L., J. R. Smith, C. J. Hutton, Z. Kapelan, and D. Savic. 2016. “Effectiveness of smart meter-based consumption feedback in curbing household water use: Knowns and unknowns.” *J. Water Resour. Plann. Manage.* 142 (12): 04016060. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000703.

Tanha, J., M. van Someren, and H. Afsarmanesh. 2015. "Semi-supervised self-training for decision tree classifiers." *Int. J. Mach. Learn. Cybern.* 8 (1): 355–370. https://doi.org/10.1007/s13042-015-0328-7.

Weston, J., C. Leslie, E. Ie, D. Zhou, A. Elisseeff, and W. S. Noble. 2005. "Semi-supervised protein classification using cluster kernels." *Bioinformatics* 21 (15): 3241–3247. https://doi.org/10.1093/bioinformatics/bti497.

Whitley, D. 1994. "A genetic algorithm tutorial." *Stat. Comput.* 4 (2): 65–85. https://doi.org/10.1007/BF00175354.

Willis, R. M., R. A. Stewart, D. P. Giurco, M. R. Talebpour, and A. Mousavinejad. 2013. "End use water consumption in households: Impact of socio-demographic factors and efficient devices." *J. Cleaner Prod.* 60 (Dec): 107–115. https://doi.org/10.1016/j.jclepro.2011.08.006.

Wong, J. S., Q. Zhang, and Y. D. Chen. 2010. "Statistical modeling of daily urban water consumption in Hong Kong: Trend, changing patterns, and forecast." *Water Resour. Res.* 46 (3): W03506. https://doi.org/10.1029/2009WR008147.

Yang, A., H. Zhang, R. A. Stewart, and K. Nguyen. 2018. "Enhancing residential water end use pattern recognition accuracy using self-organizing maps and k-means clustering techniques: Autoflow v3.1." *Water* 10 (9): 1221. https://doi.org/10.3390/w10091221.