Dynamic Time Warping Clustering to Discover Socioeconomic Characteristics in Smart Water Meter Data

Steffelbauer, D. B.; Blokker, M.; Buchberger, S. G.; Knobbe, Arno; Abraham, E.

DOI
10.1061/(ASCE)WR.1943-5452.0001360

Publication date
2021

Document Version
Final published version

Published in
Journal of Water Resources Planning and Management

Citation (APA)
Steffelbauer, D. B., Blokker, M., Buchberger, S. G., Knobbe, A., & Abraham, E. (2021). Dynamic Time Warping Clustering to Discover Socioeconomic Characteristics in Smart Water Meter Data. *Journal of Water Resources Planning and Management*, *147*(6), Article 04021026. https://doi.org/10.1061/(ASCE)WR.1943-5452.0001360

Dynamic Time Warping Clustering to Discover Socioeconomic Characteristics in Smart Water Meter Data

D. B. Steffelbauer, Ph.D.[1]; E. J. M. Blokker, Ph.D.[2]; S. G. Buchberger, Ph.D., M.ASCE[3]; A. Knobbe, Ph.D.[4]; and E. Abraham, Ph.D.[5]

Abstract: Socioeconomic characteristics are influencing the temporal and spatial variability of water demand, which are the biggest source of uncertainties within water distribution system modeling. Improving current knowledge of these influences can be utilized to decrease demand uncertainties. This paper aims to link smart water meter data to socioeconomic user characteristics by applying a novel clustering algorithm that uses a dynamic time warping metric on daily demand patterns. The approach is tested on simulated and measured single-family home data sets. It is shown that the novel algorithm performs better compared with commonly used clustering methods, both in finding the right number of clusters as well as assigning patterns correctly. Additionally, the methodology can be used to identify outliers within clusters of demand patterns. Furthermore, this study investigates which socioeconomic characteristics (e.g., employment status and number of residents) are prevalent within single clusters and, consequently, can be linked to the shape of the cluster's barycenters. In future, the proposed methods in combination with stochastic demand models can be used to fill data gaps in hydraulic models. **DOI: 10.1061/(ASCE)WR.1943-5452.0001360.** *© 2021 American Society of Civil Engineers.*

Introduction

Water utilities make use of hydraulic simulation software to design and operate their systems in a more effective way (Walski et al. 2003). However, models of water distribution systems (WDSs) consist of thousands or tens of thousands of parameters (length, diameter, and roughness of every pipe or the water demand at every node). These parameters are mostly unknown, have to be estimated through model calibration (Zhou et al. 2018), and are fraught with uncertainties. Especially, water demand plays a crucial role in the dynamics of WDS because it fluctuates over a variety of temporal and spatial scales depending on the type of consumers (Hutton et al. 2014; Díaz and González 2020). Additionally, due to the low density of metered consumers and the difficulty in obtaining large amounts of demand data in real time, the variability of water demand is the biggest source of uncertainty in WDS modeling.

Over the last decade, smart water meters (SWMs) that measure and transmit water consumption data at the single household level are available in high temporal resolution from the subsecond up to 1 h (Boyle et al. 2013), potentially overcoming limitations of current metering practices. These devices have the potential to revolutionize WDS modeling (Gurung et al. 2014; Nguyen et al. 2018; Stewart et al. 2018). However, the large-scale roll-out of SWMs globally is yet to happen because technology adoption barriers—caused by financial, cybersecurity, and privacy issues—hinder the widespread deployment of this new technology (Cominola et al. 2015). Furthermore, for water utilities adopting this new technology, the cost-benefit trade-off has not yet been quantitatively justified (Cominola et al. 2018; Monks et al. 2019).

Besides data management challenges associated with big data streams (Shafiee et al. 2020), water companies are further challenged to generate relevant knowledge from the raw consumption data that has to be useful for their hydraulic computer models. However, a combination of system wide consumer information and data from a few SWMs represents a promising approach to reduce modeling uncertainties by filling in data gaps at the unmeasured locations according to their consumers' characteristics without extensively measuring real-time water demand at every node in a WDS. The question is whether it is possible to link raw SWM data to rather general consumer information.

Deriving valuable information from SWMs is by far not trivial because water use is stochastic by nature (Blokker et al. 2010); no one operates water end-use devices (shower, water tap, and toilet, among others) exactly at the same time each day and extracts precisely the same amount of water during each usage. Nevertheless, by building periodic means over a certain number of days, patterns in water-usage behavior emerge. These patterns are called daily demand patterns. The patterns contain information about consumer's daily routines, reveal irregular consumption behaviors, and are shaped by their socioeconomical characteristics, e.g., age, gender, economic situation, employment status, or family composition. Hereinafter, this information will be referred to as the underlying

[1] Associate Professor, Dept. of Civil and Environmental Engineering, Norwegian Univ. of Science and Technology (NTNU), S.P. Andersens veg 5, 7031 Trondheim, Norway; Marie Skłodowska Curie Fellow of the Leading Fellows PostDoc Programme, Dept. of Water Management, Faculty of Civil Engineering and Geosciences, Delft Univ. of Technology, Stevinweg 1, 2628 CN Delft, Netherlands (corresponding author). ORCID: https://orcid.org/0000-0003-2137-985X. Email: david.steffelbauer@ntnu.no

[2] Principal Scientist, Drinking Water Infrastructure Team, KWR Water Cycle Research Institute, Groningenhaven 7, 3433 PE Nieuwegein, Netherlands; Associate Professor, Dept. of Water Management, Faculty of Civil Engineering and Geosciences, Delft Univ. of Technology, Stevinweg 1, 2628 CN Delft, Netherlands.

[3] Professor, College of Engineering and Applied Science, Univ. of Cincinnati, Cincinnati, OH 45221. ORCID: https://orcid.org/0000-0002-8795-1583

[4] Associate Professor, Leiden Institute of Advanced Computer Science, Leiden Univ., Niels Bohrweg 1, 2333 CA Leiden, Netherlands. ORCID: https://orcid.org/0000-0002-0335-5099

[5] Assistant Professor, Dept. of Water Management, Faculty of Civil Engineering and Geosciences, TU Delft, Stevinweg 1, 2628 CN Delft, Netherlands. ORCID: https://orcid.org/0000-0003-0989-5456

Note. This manuscript was submitted on March 6, 2020; approved on November 7, 2020; published online on March 31, 2021. Discussion period open until August 31, 2021; 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.

high-level information. This paper will show that data-mining techniques can be used to automatically distinguish daily demand patterns into groups according to their underlying high-level consumer information.

This study aims to answer following two questions by applying a novel clustering algorithm on daily demand patterns generated from SWM data:

- How many distinct daily demand patterns are in a specific SWM data set?
- Can one draw conclusions from these pattern shapes on the consumers' underlying high-level information?

The proposed methods are tested on two artificial SWM data sets generated with the stochastic demand modeling software SIMDEUM (Blokker et al. 2010) and a real-world SWM data set from Milford, Ohio (Buchberger et al. 2003).

Related Work

Cluster analysis belongs to the family of unsupervised learning algorithms and is a technique to find groups in data sets (Lin et al. 2012). For SWM data analysis, clustering can be used to segment users into groups with similar water-use behavior (Cominola et al. 2019), e.g., commercial versus residential or single households versus multifamily homes. Although customer segmentation was mostly focusing in the past on smart meters that measure energy consumption (Espinoza et al. 2005; Nizar and Dong 2009; Nambi et al. 2016; Kwac et al. 2014), only a few studies applied customer segmentation to SWM data.

Most research used k-means clustering (Lloyd 1982) in combination with different metrics. McKenna et al. (2014) clustered SWM data into commercial and residential patterns by applying k-means on features extracted through fitting Gaussian mixture models. Mounce et al. (2016) used k-means++ (Arthur and Vassilvitskii 2007) with correlation distance to cluster data in residential and commercial groups. Garcia et al. (2015) classified demand patterns using k-means clustering. Cheifetz et al. (2017) made use of Fourier-based time-series models for clustering demand patterns. The patterns were qualitatively interpreted as residential, commercial, office, industrial, or noise. Cardell-Oliver et al. (2016) identified groups of similar households by features of their high-magnitude water-use behaviors based on previous work (Cardell-Oliver 2013a, b; Wang et al. 2015). Cominola et al. (2018) applied customer segmentation analysis simultaneously on water and electricity data by clustering extracted eigenbehaviors and linked the clusters to a list of user psychographic features. Recently, Cominola et al. (2019) coupled nonintrusive end-use disaggregation with customer segmentation to identify and cluster primary water-use behaviors.

Clustering techniques are also used as a prior step to demand forecasting. For example, Aksela and Aksela (2011) constructed clusters with k-means according to their average weekly consumption before forecasting future demand. Candelieri (2017) clustered SWM data using a k-means with cosine metric, first to split data into weekdays and weekends and then to split the data into residential, nonresidential, and mixed-type clusters.

In contrast to the aforementioned studies, this paper employs soft dynamic time warping (SDTW) as time-series clustering metric. This metric is capable of optimally aligning two sequences in time by nonlinearly warping the time axes of the sequences until their dissimilarity is minimized (Dürrenmatt et al. 2013). The time when people use water is highly variable among different users with otherwise similar socioeconomic characteristics (Blokker et al. 2008). The SDTW metric can expose similarities in daily schedules of inhabitants' water use that are shifted in time, which are not detectable by using linear time metrics (Euclidean or correlation). Originally developed for speech recognition (Sakoe and Chiba 1978), dynamic time warping (DTW) has been applied in the field of water management in the past, but never in the context of time-series clustering. Past applications of DTW in water management included burst detection (Huang et al. 2018), analyzing residence times in wastewater treatment plant reactors (Dürrenmatt 2011), sewer flow monitoring (Dürrenmatt et al. 2013), or identifying water demand end uses (Yang et al. 2018; Nguyen et al. 2014).

Contributions

Although clustering of SWM data has been done before, this paper's approach is innovative in many aspects. First, a novel method is proposed to cluster SWM data that is capable of finding similarities in daily demand patterns even if they are shifted in the time domain. Hence, it should outperform clustering methods with fixed time metrics [e.g., Euclidean (Mounce et al. 2016; Garcia et al. 2015)]. Second, the proposed methods are tested on SWM data sets that were simulated with the stochastic demand modeling software SIMDEUM. Because the ground truth of these data sets are known, it enables measuring and comparing the performance of different clustering approaches. Third, whereas former work focused on identifying different types of consumer classes by mainly distinguishing between residential and commercial use, this work goes one step further by investigating which underlying high-level information of residential customers is prevalent in different clusters, e.g., by looking at work schedules or the number of household residents.

Furthermore, it is sought to highlight the simplicity of the proposed approach compared with other methods. Whereas most of the discussed studies used clustering on burdensome obtained surrogate parameters [e.g., eigenbehaviors (Cominola et al. 2019), high-magnitude water-use behaviors (Cardell-Oliver et al. 2016), or parameters from fits from Gaussian distributions (McKenna et al. 2014) or Fourier regression mixture models (Cheifetz et al. 2017)], the SDTW clustering method is applied directly on the demand patterns and, hence, does not risk losing valuable information contained in the raw data. Additionally, SDTW enables user segmentation using water consumption data without the need for additional information as, for example, electricity (Cominola et al. 2018) or end-use disaggregated water consumption data (Cominola et al. 2019). Moreover, the SDTW clustering approach is an unsupervised algorithm with no need for prior information nor previous calibration of, for example, consumption threshold parameters.

Materials and Methods

Water Demand Pattern Generation

SWM data (and demand patterns) are time series (Shumway and Stoffer 2010). A time series $\mathbf{x}$ of length M is a sequence of data points in strict chronological order

$$\mathbf{x} = \{x_t\} = (x_1, x_2, \dots, x_M) \in \mathbb{R}^M \tag{1}$$

A water demand pattern $\overline{\mathbf{x}}_P$ is generated by building periodic means from a SWM time series $\mathbf{x}$

$$\overline{\mathbf{x}}_j^P = \frac{1}{N_P}\sum_{i=1}^{N_P} x_{P(i-1)+j} \quad \forall\ j = 1, 2, \dots, P \tag{2}$$

where P = period length; and $N_P = \lfloor M/P \rfloor$ is the number of full periods in $\mathbf{x}$. More specifically, this paper will deal with daily demand patterns ($P = 24$ h).

Soft Dynamic Time Warping

Unlike Euclidean distance, SDTW is able to compare time series of variable sizes and is robust to shifts or dilations across the time dimension (Cuturi and Blondel 2017). The SDTW method computes the best possible alignment between time series. This is relevant for SWM data because the daily water-use behaviors of different households might be similar but shifted in the time domain due to different daily schedules, e.g., caused by different wake-up, working, or commuting times.

Two Time Series

Let $\mathbf{x} \in \mathbb{R}^M$ and $\mathbf{y} \in \mathbb{R}^N$ be two time series, where $\mathbf{x}$ and $\mathbf{y}$ do not have to be equally long or have the same sampling rate. Because this paper focuses on clustering daily demand patterns, the time series are always of the same length and sampling rate introduced by the periodic mean in Eq. (2). First, the elements of a pairwise distance matrix $\mathbf{D} \in \mathbb{R}^{M\times N}$ are computed between the points of two time series $\mathbf{x}$ and $\mathbf{y}$

$$D_{mn} = \delta(x_m, y_n) \tag{3}$$

where δ = arbitrary distance metric. A path connecting the upper-left corner and bottom-right corner of $\mathbf{D}$ that only allows moves to the right, diagonal, or down is called a warping path $\mathbf{p}$. This path is used to align the two time series $\mathbf{x}$ and $\mathbf{y}$. The warping path $\mathbf{p}$ is linked to the binary alignment matrix $\mathbf{A} \in \mathbb{R}^{M\times N}$ as follows:

$$\mathbf{p} = (\rightarrow, \searrow, \downarrow, \rightarrow, \dots) \triangleq \mathbf{A} = \begin{bmatrix} 1 & 1 & 0 & 0 & \cdots & \\ 0 & 0 & 1 & 0 & & \\ 0 & 0 & 1 & 1 & & \\ \vdots & & & & \ddots & \\ & & & & & 1 \end{bmatrix} \tag{4}$$

In the following, $\aleph \in \{0,1\}^{M\times N}$ is the set of all possible (binary) alignment matrices $\mathbf{A}$. The warping distance d along a warping path $\mathbf{p}$ is defined through

$$d(\mathbf{A},\mathbf{D}) = \sum_{ij} \mathbf{A}_{ij} D_{ij} = \mathrm{Tr}(\mathbf{A}^T D) \tag{5}$$

where $\mathrm{Tr}(\cdot)$ = trace of a matrix. The optimal warping path $\mathbf{p}^*$ with minimal distance d^* is computed with SDTW in the following way (Cuturi and Blondel 2017):

$$d^*(\mathbf{x},\mathbf{y}) = \min_{A\in\aleph}(d(\mathbf{A},\mathbf{D})) = -\log\left(\sum_{A\in\aleph} e^{-d(\mathbf{A},\mathbf{D})}\right) \tag{6}$$

The SDTW metric integrates over all possible alignments, is differentiable, and leads to a robust smooth solution in an optimization framework (Cuturi and Blondel 2017). Although the set of all possible alignment matrices $\aleph$ grows exponentially with M and N, Eq. (6) can be recursively solved with computational complexity of order $\mathcal{O}(MN)$ starting from $r_{0,0} = 0$

$$r_{i,j} = \delta(x_i, y_j) + \min(r_{i-1,j-1}, r_{i-1,j}, r_{i,j-1}) \tag{7}$$

Multiple Time Series

The optimal distance d^* can be used to average over multiple time series. The ability to build such averages is a necessary condition for time-series clustering. Let $\{\mathbf{y}_l\} = (\mathbf{y}_1, \dots, \mathbf{y}_L)$ be a family of L time series. To average $\{\mathbf{y}_l\}$ with SDTW, the following minimization problem has to be solved:

$$\min_{\mathbf{x}^*\in\mathbb{R}^M} \sum_{i=1}^{L} d^*(\mathbf{x}^*, \mathbf{y}_i) \tag{8}$$

This problem is solved using a quasi-Newton method, the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm (Nocedal and Wright 2006). The solution $\mathbf{x}^*$ is called the barycenter of $\{\mathbf{y}_l\}$.

Clustering

The k-means clustering method (Lloyd 1982) is used to identify different demand patterns in the SWM data. The principle behind the k-means algorithm is to separate the data into a preset number of k clusters that minimize intracluster variability and maximize intercluster differences (McKenna et al. 2014). Let $\{\mathbf{y}_l\}$ be again a family of L time series. Then k-means clustering in the Euclidean metric equals minimizing the following nested sums:

$$\min_c \sum_{j=1}^{k} \sum_{\mathbf{y}\in C_j} \|\mathbf{y} - \boldsymbol{\mu}_j\|_2 \tag{9}$$

where $\boldsymbol{\mu}_j$ = within-cluster mean of C_j. The Euclidean norm $\|\cdot\|_2$ is only valid when the time series are of equal length ($M = N$). Analogously, clustering in the SDTW metric is defined

$$\min_c \sum_{j=1}^{k} \sum_{\mathbf{y}\in C_j} d^*(\mathbf{x}_j^*, \mathbf{y}) \tag{10}$$

where $\mathbf{x}_j^*$ = barycenter of the jth cluster.

Furthermore, a simplified version of the k-means clustering algorithm is introduced here. Instead of investigating the whole daily demand time series $\mathbf{y}$, this algorithm uses only the mean ($E[\mathbf{y}]$) and standard deviation ($\sqrt{E[\mathbf{y}^2] - (E[\mathbf{y}])^2}$) of the daily pattern during work hours (10:00 a.m.–4:00 p.m.). Hence, the feature space is reduced to a two-dimensional space through following transformation:

$$\mathbf{y} \Rightarrow \begin{pmatrix} E[\mathbf{y}] \\ \sqrt{E[\mathbf{y}^2] - (E[\mathbf{y}])^2} \end{pmatrix} \tag{11}$$

Within this work, the problem defined in Eq. (9) will be called Euclidean clustering, along with simple clustering [Eq. (11)], and SDTW clustering [Eq. (10)]. The latter is capable of finding more general similarities in patterns by allowing more freedom in the time domain and will be benchmarked against simple and Euclidean clustering. The initial cluster centers are seeded according to the k-means++ algorithm (Arthur and Vassilvitskii 2007) to increase the method's robustness. Prior to clustering, the time series $\mathbf{y}_i$ can be normalized

$$\mathbf{y}_i' = \frac{\mathbf{y}_i - \min(\mathbf{y}_i)}{\max(\mathbf{y}_i) - \min(\mathbf{y}_i)} \tag{12}$$

Performance Measures

Success and error rates are used to validate the clustering results based on the ground truth (Witten et al. 2011), whereas silhouette coefficients are used to validate the clusters based on the dissimilarities and similarities of their members (Rousseeuw 1987). Two cases have to be distinguished for validating clustering results: (1) if the ground truth is known; and (2) if there exists no information about the true nature of the outcomes. For the first case, the correct allocation of distinct patterns is known, and one can compute a success and an error rate (Witten et al. 2011). In the second case, the allocation and the number of distinct patterns is unknown. For data sets with unknown ground truth, the clustering results can still be validated based on silhouette coefficients.

Success and Error Rate

True positive (TP) is the case if a pattern is assigned to the correct cluster. False positive (FP) means that a pattern belonging to another cluster is wrongly assigned to the current cluster. A true negative (TN) is the case when a pattern from another cluster is correctly assigned to the other cluster. Finally, a false negative (FN) is the case when a pattern belonging to the cluster is wrongly assigned to another cluster. One can define an overall success rate (SR) with all aforementioned cases through the following equation (Witten et al. 2011):

$$\mathrm{SR} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{13}$$

The error rate (ER) is the complement of SR

$$\mathrm{ER} = 1 - \mathrm{SR} \tag{14}$$

Silhouette Coefficients

Silhouette coefficients S_i are properties of a single time series $\mathbf{y}_l$ (Rousseeuw 1987). They can be used to determine the quality of clusters when the ground truth is unknown. The S_i values are computed as a combination of two scores: (1) the mean intracluster distance, and (2) the distance between a sample and the nearest cluster of which $\mathbf{y}_l$ is not part. The mean intracluster distance is defined as the average distance of time series $\mathbf{y}_l$ to all other time series $\mathbf{y}_j$ that are members of the same cluster C_i. Let $\mathbf{y}_l$ be the lth member of the time series belonging to cluster C_i; then, its intracluster distance (mean distance between all members $\mathbf{y}_j$ of cluster C_i) $a(\mathbf{y}_l)$ is defined

$$a(\mathbf{y}_l) = \frac{1}{|C_i| - 1} \sum_{\substack{\mathbf{y}_j \in C_i \\ j \neq l}} d(\mathbf{y}_j, \mathbf{y}_l) \tag{15}$$

where $|C_i|$ = number of samples in cluster C_i; and d = arbitrary distance metric. The second score $b(\mathbf{y}_l)$ is the distance d between the time series $\mathbf{y}_l$ of C_i and its nearest cluster C_j^* as follows:

$$b(\mathbf{y}_l) \coloneqq \min_{i \neq j} \frac{1}{|C_j|} \sum_{\mathbf{y}_k \in C_j} d(\mathbf{y}_k, \mathbf{y}_l) = \frac{1}{|C_j^*|} \sum_{\mathbf{y}_k \in C_j^*} d(\mathbf{y}_k, \mathbf{y}_l) \tag{16}$$

where $\mathbf{y}_k \in C_j$ = members of the cluster C_j. The two scores are combined in the following way to obtain the Silhouette coefficient of a time series $\mathbf{y}_l$ (Rousseeuw 1987):

$$S(\mathbf{y}_l) = \frac{b(\mathbf{y}_l) - a(\mathbf{y}_l)}{\max(a(\mathbf{y}_l), b(\mathbf{y}_l))} \tag{17}$$

By definition, the silhouette coefficient is $-1 \leq S_i \leq 1$. Higher S_i values relate to a model with better defined clusters, i.e., each time series is closer to its own cluster members than to the nearest cluster.

The specific number of clusters k is required as an input parameter for the k-means algorithm. The average silhouette coefficient for all L time series can be used to compare clustering results for different k to decide on the number of clusters that are in the data

$$\overline{S} = \frac{1}{L} \sum_{l=1}^{L} S(\mathbf{y}_l) \tag{18}$$

Investigating the behavior of $\overline{S}$ as a function of k is subsequently called cluster analysis.

Data Sets

The SDTW methodology is tested on three data sets of water use at single-family homes: (1) an artificial SWM data set generated with SIMDEUM (Blokker et al. 2010) consisting of single-person households, (2) another SIMDEUM data set with multiple-person homes, and (3) a measured SWM data set from Milford, Ohio (Buchberger et al. 2003). Daily demand patterns with a time resolution of 30 min are generated and smoothed with a 2-h moving average. A more detailed description of the data sets can be found in the Supplemental Materials.

SIMDEUM Single-Person Households

SIMDEUM is a water demand end-use model that is capable of simulating water usage at household level with a time resolution of down to 1 s (Blokker et al. 2010). The model generates randomly water end-use events based on statistical information of users and end-use devices. The information includes census data for the number of residents in a household, their age distribution, the average number of appliances, and their daily routines, as well as frequency, duration, and intensity for different end-uses such as numbers of kitchen tap uses, toilet flushes, showers taken per day, and washing machine or dishwasher uses, among others.

To generate the first simulated data set, 100 single-person households with different daily routines were simulated. For each household, the water consumption of its residents for a period of 100 days was simulated. The data set consists of adult inhabitants with (50) and without (50) jobs away from home. Consequently, the employment status of the occupants is the underlying high-level information responsible for the different pattern shapes. Throughout this paper, the first half of the household demand patterns are referred to as work patterns and the second half as home patterns.

SIMDEUM Multiple-Person Households

The second simulated data set was produced by generating SIMDEUM simulations for 200 multiperson homes with 1–5 residents according to the household statistics in the Netherlands [Table 2 in Blokker et al. (2010)]; again simulated for 100 days for each household. The high-level information consists of the type of the household (one-person, two-person, or family), the number of residents, their age distribution, their daily schedules, and their profession (employed, unemployed, retired, child, or teenager).

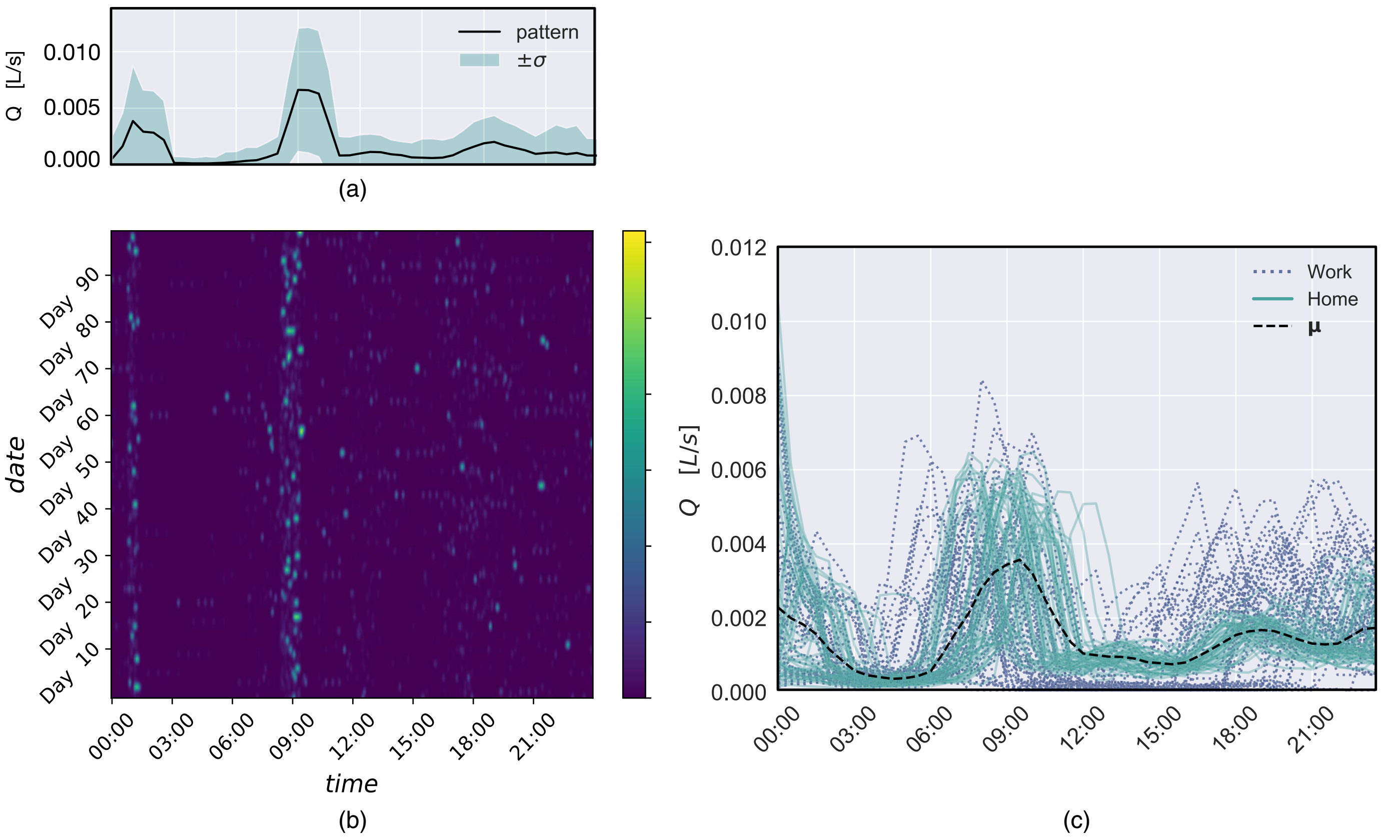
Measured SWM Data from Milford, Ohio

The measured SWM data set contains data from SWM installed at 21 single-family houses in Milford, Ohio, recorded between April 1 and October 31, 1997 (Buchberger et al. 2003). Because user behavior and, hence, daily demand patterns, can differ significantly between weekends and weekdays (Alvisi et al. 2007), patterns were generated separately for weekdays and weekends. The underlying high-level information contains the number of residents and the pattern type (home or work).

Results

Simulated Data Set with Single-Person Households

The data set consists of work and home patterns. A cluster consisting predominantly of work patterns is called a work cluster, and a cluster whose majority patterns are home patterns is a home cluster. The intention of this numerical experiment is to see (1) if the SDTW clustering approach is able to extract the employment status of the residents; (2) if the correct number of distinct patterns in the data set can be identified with cluster analysis; and (3) how SDTW clustering performs compared with the benchmark algorithms

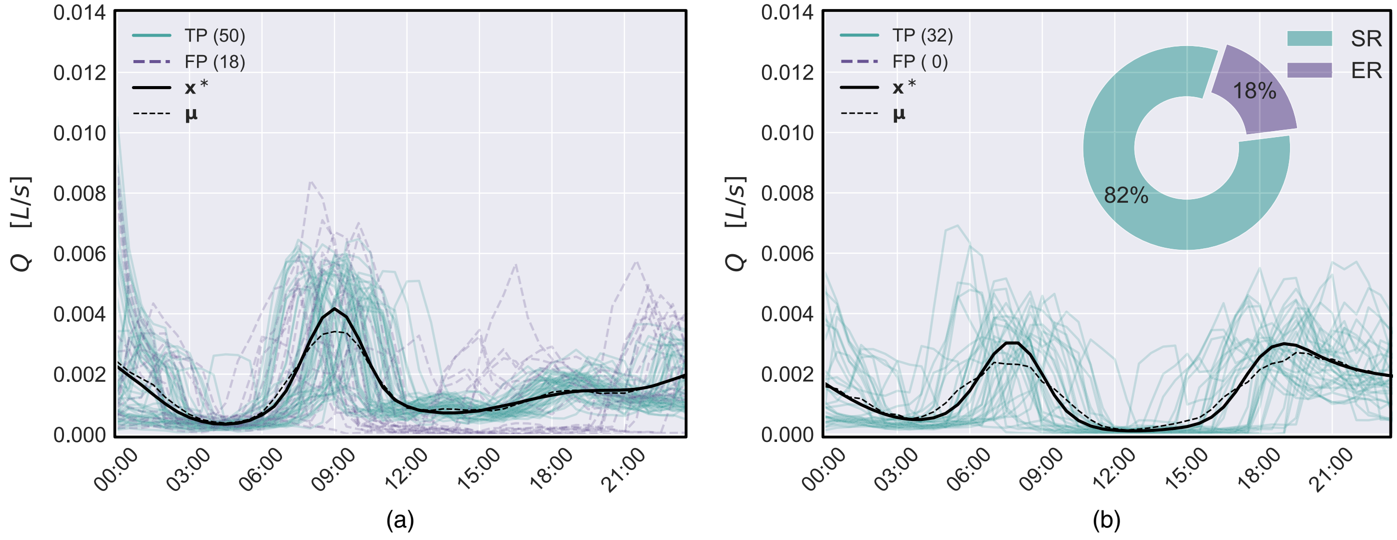


(a)

(b)

(c)

Fig. 1. SIMDEUM single-person household data set: (a) daily demand pattern constructed from 100 days of SWM data and its standard deviation σ of an example household; (b) example household's water consumption over 100 days; and (c) demand patterns (mean over 100 days) of the whole data set showing 50 work patterns, 50 home patterns, and the mean overall demand patterns $\boldsymbol{\mu}$.



(a)

(b)

Fig. 2. Clustering results for SIMDEUM single-person household data set showing the (a) home and (b) work cluster, including a doughnut chart showing the SR and ER.

(simple and Euclidean clustering). Fig. 1 presents the SIMDEUM data set. Fig. 1(a) shows the daily demand pattern of a single-person household where the occupant is staying at home throughout the day. Additionally, the standard deviation σ is shown. Fig. 1(b) provides the 100 days water-usage data of this household. Fig. 1(c) presents the demand patterns of the whole SIMDEUM data set (100 households). The work patterns are shown as dotted lines, the home patterns are shown as solid lines. Furthermore, the mean over all patterns is shown as a dashed line. Variations of the patterns from the mean are clearly visible both in time and in magnitude.

First, SDTW clustering is applied on the data set. The patterns are normalized prior to clustering to suppress the influence of different average consumption, so that the algorithm focuses only on pattern shapes and not on magnitudes. Figs. 2(a and b) present

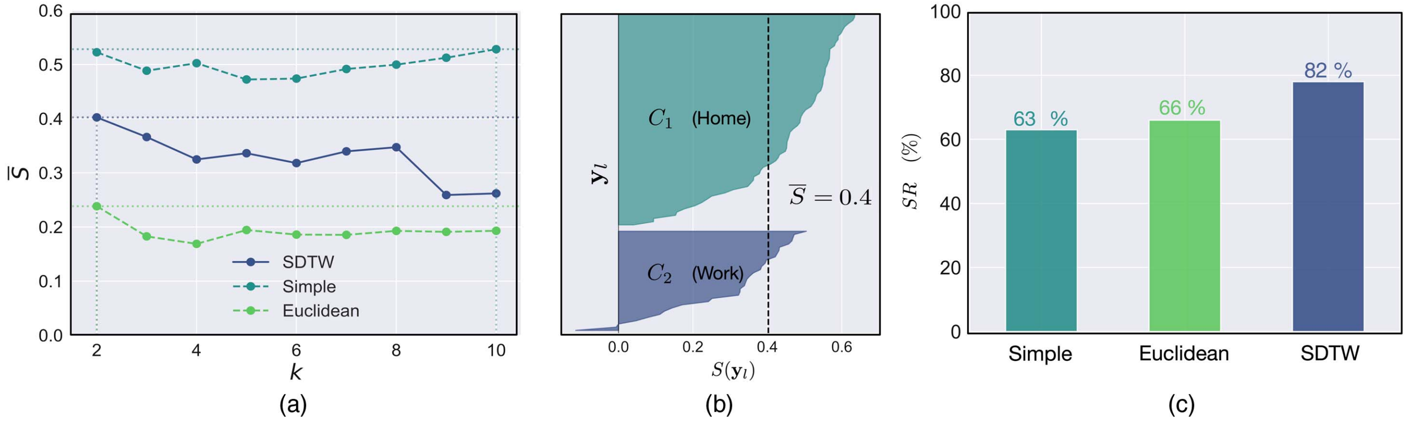


Fig. 3. Cluster analysis plots for SIMDEUM single-person household data set: (a) average silhouette coefficient $\overline{S}$ as a function of the number of clusters k for SDTW clustering and the benchmark algorithms, where dashed lines highlight the maximum $\overline{S}$ values; (b) silhouette plot for the SDTW metric for $k = 2$ showing the silhouette coefficients of each member of the two clusters C_1 (home) and C_2 (work); and (c) comparison of the performance between SDTW clustering and the benchmark algorithms with respect to the SRs.

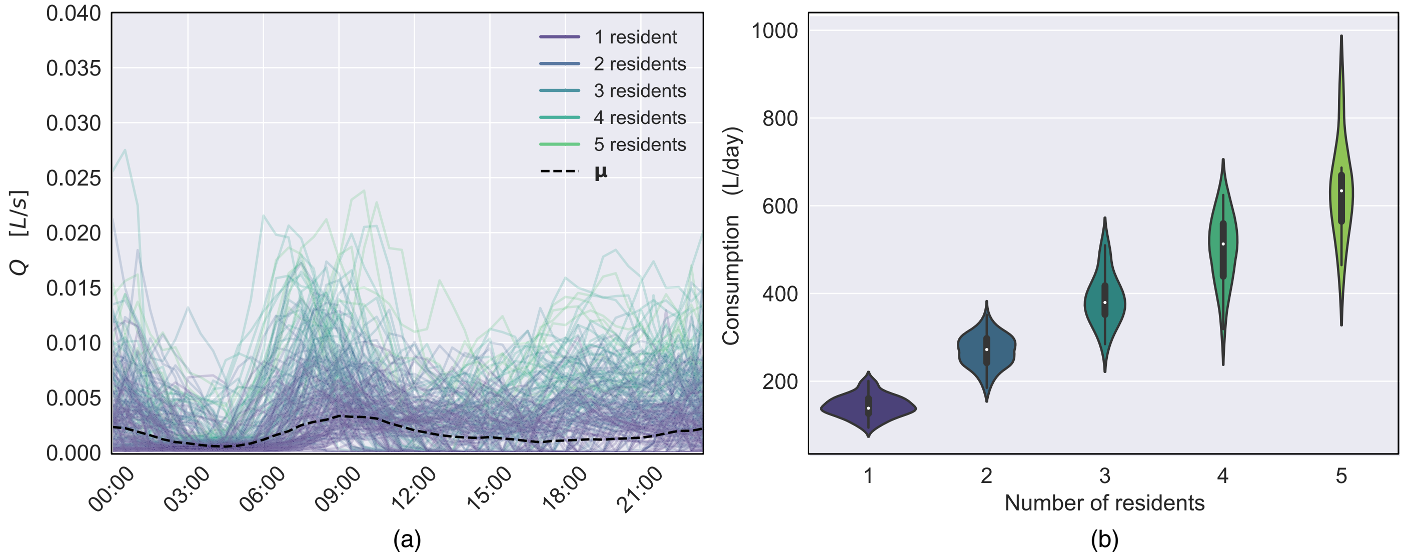


Fig. 4. Representation of 200 simulated multiperson households: (a) daily demand pattern of the households indicating the number of residents, where mean $\boldsymbol{\mu}$ is depicted as a dashed line; and (b) violin plots for the average daily consumption versus the number of residents.

the clustering results for $k = 2$ and the performance measures, with the home cluster in Fig. 2(a) and the work cluster in Fig. 2(b). Demand patterns classified correctly (TP) are shown as solid lines, and FP are depicted as dashed lines. The work cluster is a pure cluster consisting only of work patterns. The SR is 82% and ER is 18% [doughnut chart in Fig. 2(b)]. Furthermore, Figs. 2(a and b) show the barycenters $\mathbf{x}^*$. Additionally, the within-cluster mean $\boldsymbol{\mu}$ is shown to illustrate the difference of $\mathbf{x}^*$ and $\boldsymbol{\mu}$. It can be seen that the SDTW clustering approach is capable of segmenting the daily demand patterns according to the employment status of the inhabitants. Furthermore, the barycenters show the expected water-use behavior of the user groups. The users within the home cluster use water over the whole day, whereas users in the work cluster have almost zero consumption while their residents are at work.

Second, a cluster analysis is performed to see if the correct number of patterns can be identified. The distinct number of patterns is two (work and home) because there is no other high-level information contained in the data. The results of the cluster analysis with the SDTW method are presented in Fig. 3(a) and compared with the benchmark algorithms. Additionally, the individual silhouette coefficients $S(\mathbf{y}_l)$ are shown for the correct number of clusters $k = 2$ [Fig. 3(b)]. The maximum $\overline{S}$ value indicates the most probable number of clusters in the data set. The SDTW and Euclidean clustering approach are capable of finding the correct number of clusters (maximum at $k = 2$), whereas the simple algorithm overestimates the number of clusters. Fig. 3(c) shows a comparison of the three clustering algorithms with respect to the SRs, where the SDTW algorithm clearly performs best.

Simulated Data Set with Multiple-Person Households

The purpose of this experiment is to apply the SDTW algorithm on a more complex data set. Fig. 4 shows the data set. On average,

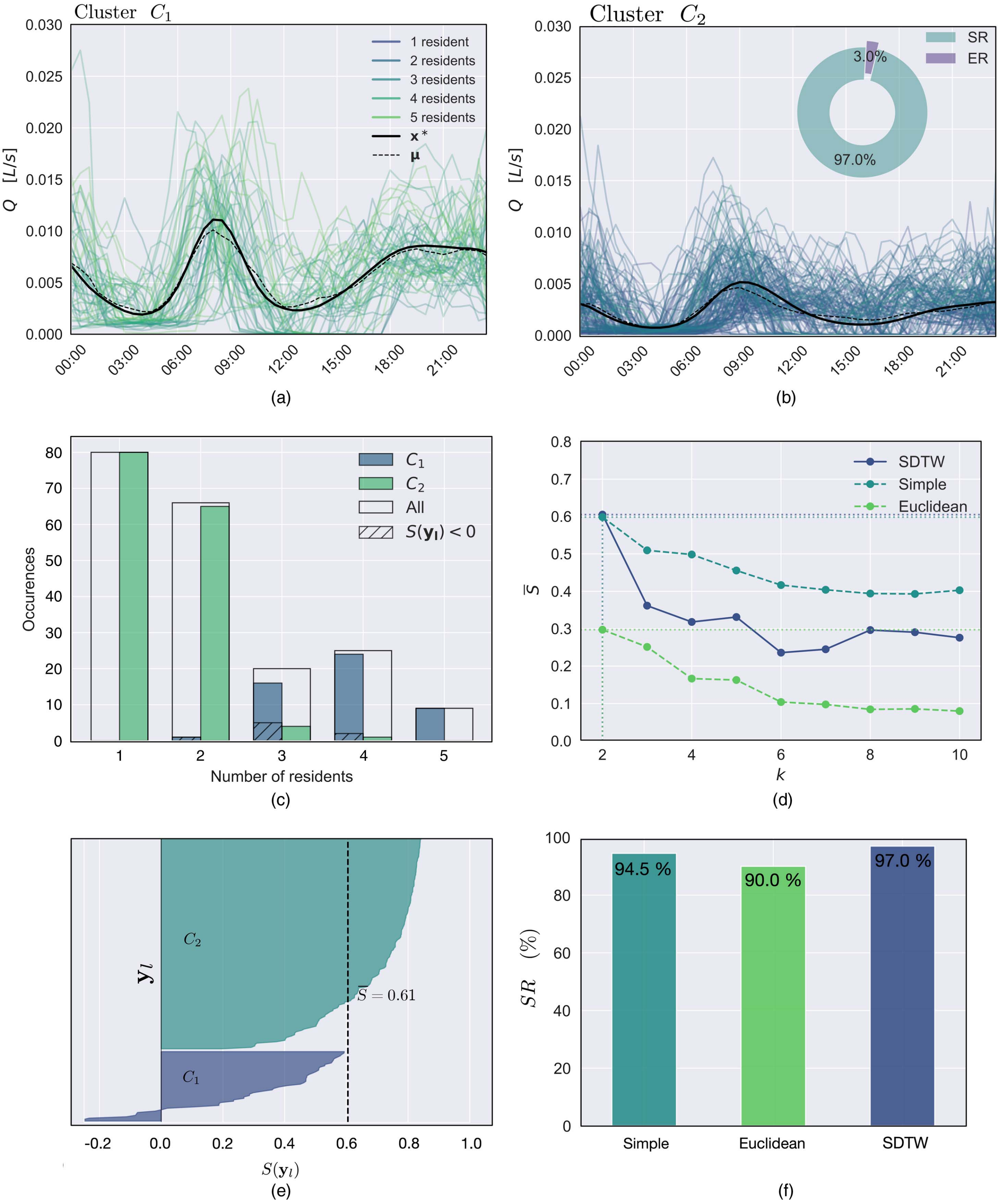


Fig. 5. Clustering results for SIMDEUM multiperson households obtained with SDTW clustering. SDTW: (a) family household cluster; (b) one- and two-person home cluster, all shown with barycenters $\mathbf{x}^*$, cluster means $\boldsymbol{\mu}$, and number of residents; (c) histograms for the resident distribution within the clusters; (d) cluster analysis plots; (e) silhouette plots for $k = 2$ for SDTW clustering; and (f) comparison of clustering performance with respect to SRs.

the consumption grows linearly with the household size or the type of household (one-person, two-person, or family). Hence, this high-level information (type of households and number of users) is supposed to be influential on the pattern shape. Furthermore, the growing variance in the data leads to a lot of consumption overlaps between households of different resident numbers, making it difficult to segment the data by consumption only. First, a cluster analysis will be performed to identify the number of distinct patterns in the data set, followed by a closer look on the cluster's barycenters. Second, an attempt will be made to identify the most influential high-level information. Again, the performance of the SDTW method is compared with simple and Euclidean clustering.

The doughnut charts in Fig. 5(b) depict the SR and ER in segmenting the daily patterns in one- and two-person homes versus family homes. Fig. 5(c) uses bars to show the total number of homes, and results with negative silhouette coefficients are also highlighted. The cluster analysis is shown in Fig. 5(d) together with the individual silhouette coefficients for SDTW clustering for $k = 2$ [Fig. 5(e)]. The average silhouette value is $\overline{S} = 0.61$. All clustering algorithms identified two distinct clusters. Clustering results for $k = 2$ are depicted in Fig. 5. Cluster C_1 contains mostly family households, and C_2 contains one- and two-person homes. It is assumed that the algorithm segments the daily patterns into family and nonfamily homes (one-person and two-person households). This consideration is taken into account to compute the success rate, which equals $SR = 97\%$ [Fig. 5(b)]. A comparison with the benchmark algorithms shows again that SDTW clustering has the highest SRs [Fig. 5(f)]. Fig. 5(c) shows a histogram of the cluster members in dependency on the resident numbers. The clusters are well separated for one and two persons as well as for four and more persons. Three-person households are present in both clusters with a much higher probability of being a member of C_1 (family households).

Measured Data Set (Milford, Ohio)

The same technique are now applied to the Milford, Ohio, data set. Subsequently, a closer look at the clustering results and individual silhouette coefficients is used to identify possible outliers within the clusters. Fig. 6 shows the data set. The increase in consumption by household size is not as prominent as for the SIMDEUM simulations (Fig. 4). Furthermore, the variance is high, leading to overlaps between households of all different resident numbers. Hence, the number of residents will not play a big role in segmenting the patterns as other information, e.g., work schedules.

The results of the cluster analysis can be found in Fig. 7(d). The average silhouette coefficients $\overline{S}$ show a very distinct maximum for $k = 2$ for all clustering approaches, resulting in the assumption that two distinct patterns are present in the data set. Figs. 7(a–c) present clustering results for SDTW for $k = 2$. Additionally, Fig. 7(c) shows the total number of homes, and results with negative silhouette coefficients $[S(\mathbf{y}_l) < 0]$ are highlighted. The doughnut charts depict the SR and ER in segmenting the daily patterns in work and home patterns. The clusters are not connected to the number of residents as for the SIMDEUM multiperson homes, but are shown to be dependent on the residents' work schedules. The barycenters $\mathbf{x}^*$ clearly show that cluster C_1 [Fig. 7(a)] is the work cluster, and C_2 represents the home cluster [Fig. 7(b)]. Thus, the work schedules are identified as the most influential high-level information with a $\mathrm{SR} = 95\%$. SDTW clustering results again in the highest SRs [Fig. 7(f)]. Additionally, Fig. 7(c) shows a histogram of the cluster members in dependency of their work schedules. The clusters are well separated in the work and home cluster.

Only two patterns are misclassified. These special cases are depicted in Fig. 8. The weekday pattern of Home 11 has a different shape from all other patterns with three peaks. It is marked as a work pattern through expert opinion, but is a cluster member of the home cluster. The weekend pattern of Home 15 is the other way around, i.e., classified as work pattern, but marked as home pattern (all weekend patterns are supposed to be home patterns). A closer look reveals a shape that is between the shape of work and home patterns, showing distinct morning and evening peaks, but no prominent valley during the day. Furthermore, by looking at the negative silhouette coefficients, two dissimilar patterns are identified (weekend patterns of Home 3 and 6). These patterns have a pre-eminent morning peak, but are missing a peak in the evening.

Additional to the results presented in this section, further numerical studies on the same data sets can be found in the Supplemental Materials. The additional results contain comparisons between

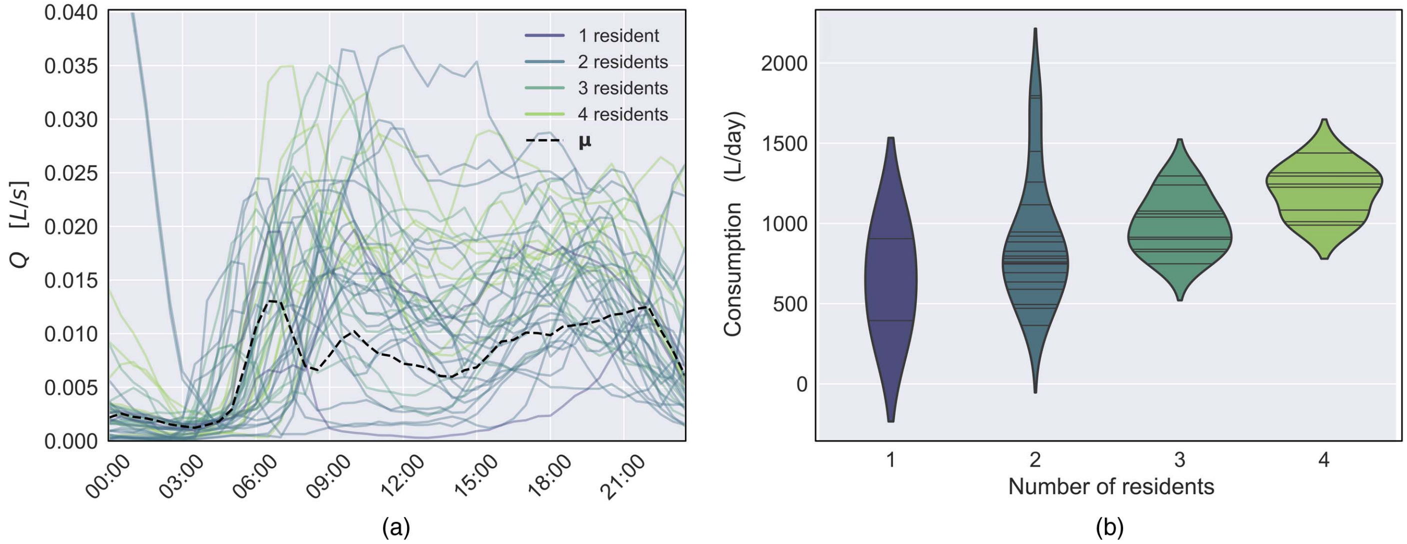


Fig. 6. Representation of the Milford (Ohio), data set consisting of 21 households and divided into weekday and weekends: (a) daily weekday and weekend demand pattern of the households according to the number of residents, with mean $\boldsymbol{\mu}$ depicted as a dashed line; and (b) violin plots for the average daily consumption as a function of the resident number.

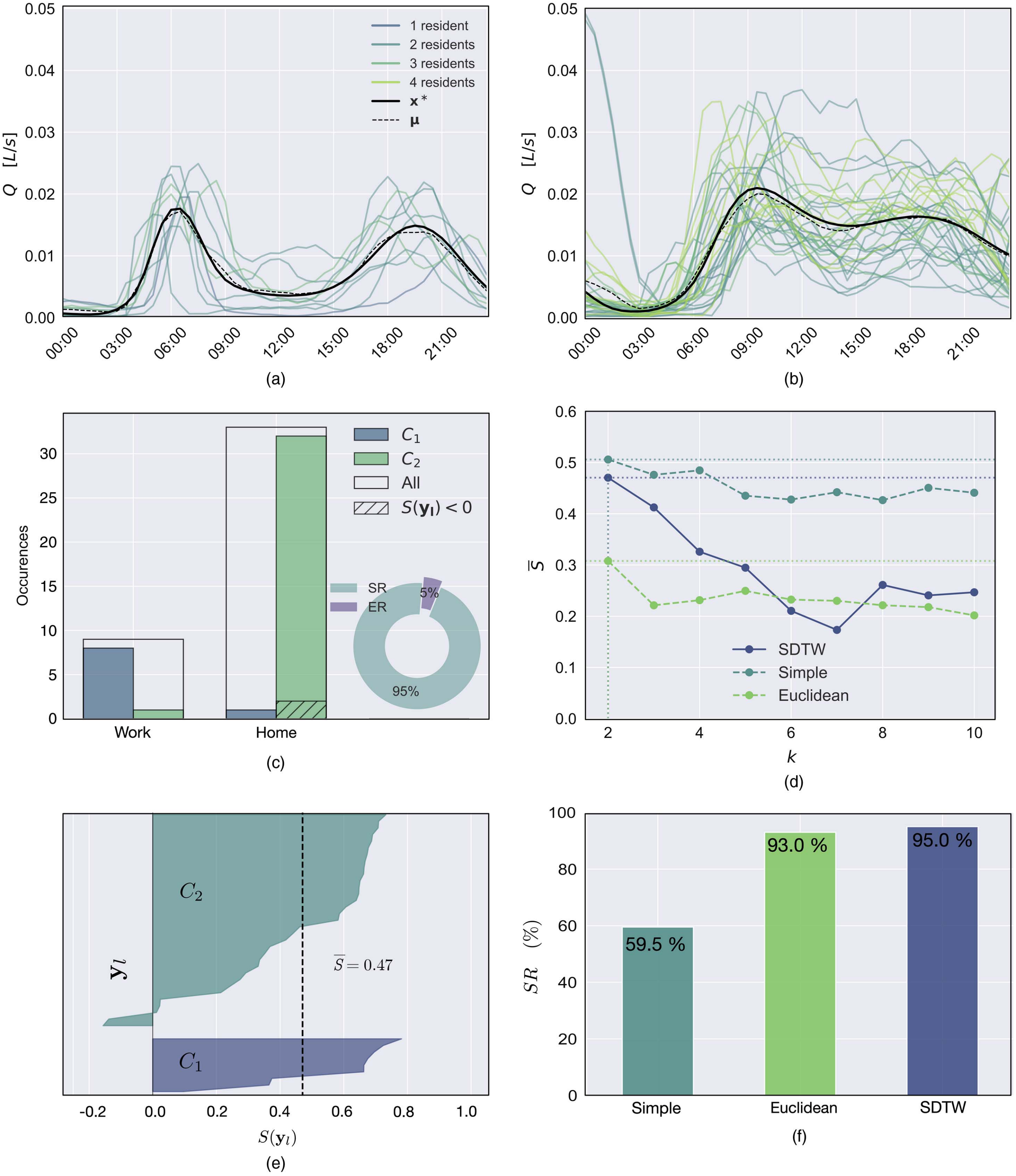


Fig. 7. Clustering results for Milford (Ohio), data set showing (a) work cluster; and (b) home cluster, all shown with barycenters $\mathbf{x}^*$, cluster means $\boldsymbol{\mu}$, and number of residents; (c) histograms for the resident distribution within the clusters; (d) cluster analysis plots; (e) silhouette plots for $k = 2$ for SDTW clustering; and (f) comparison of clustering performance with respect to SRs.

SDTW algorithm with a clustering based on the original DTW score, and time-invariance robustness tests of Euclidean clustering compared with the SDTW algorithm. Both tests were performed on the first data set. In both cases, the SDTW algorithm showed to be superior compared with DTW and Euclidean clustering. Moreover, the Supplemental Materials contain more details on the benchmark tests comparing the performance of SDTW clustering with the simple as well as the Euclidean clustering algorithm.

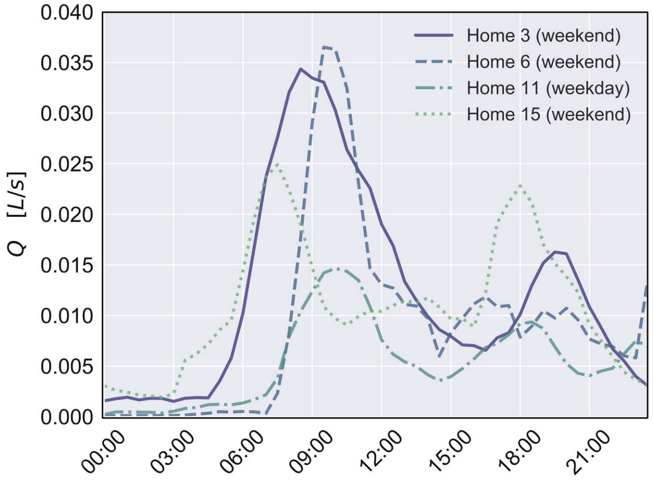


Fig. 8. Special demand patterns identified in the Milford (Ohio), data set.

Discussion

As outlined in the "Introduction," the first question this study sought to answer was the number of distinct demand patterns in a specific SWM data set. Cluster analysis was used to answer this question. The SDTW clustering method was able to identify the correct number of clusters ($k = 2$) for the simulated single-person households (Fig. 3), which correspond to the work schedules of the residents. Cluster analysis on the SIMDEUM multiple-person households clearly revealed the presence of two different demand pattern types [Figs. 5(d) and 7(d)]. The clusters of the SIMDEUM multiperson household are connected to the type of houses (family versus one- and two-person houses) or to the number of residents, respectively. Although the number of residents ranges from one to five, it was assumed herein that the household's average consumption and its variance lead to substantial overlaps of patterns with different resident numbers, making it difficult for clustering algorithms to disaggregate the data set into five separated groups. None of the benchmarking algorithms was able to identify more than two clusters in this data set, as well.

For the Milford data set, two different clusters have been found, which are linked to the residents' work schedules. By looking at the individual silhouette coefficients in the Milford data set, additional outlier patterns were identified with different shapes (e.g., a missing evening peak) (Fig. 8; Home 3 and 6). Identification of outlier patterns can be done in an automatic way. Benchmark tests with other clustering algorithms clearly showed a better performance of the SDTW algorithm with respect to success rates over all data sets (Supplemental Materials). However, the better clustering performance of the SDTW method comes with higher computational times introduced by DTW distance computations. Nonetheless, the computation time grows only linearly with the length of the time series. For the data analyzed in this paper, the clustering analysis took less than 1 min on a common personal computer. Therefore, computational complexity does not pose a problem for real-world applications.

The second question addressed the underlying high-level information responsible for the different pattern shapes. For the SIMDEUM single-person household data set, the differences in the patterns were caused by the residents' employment status (work or home). In this case, the shapes of the SDTW barycenters can be intuitively linked with the employment characteristics, resembling qualitatively better the expected behavior (e.g., work patterns have low valleys in consumption while the residents are at work). For the SIMDEUM multiperson households, the clustering approach resulted in clusters based on household type (one- and two-person versus family) with a high accuracy of 97%.

For the Milford data set, barycenters were linked to the residents' work schedules. Application of the clustering algorithms resulted in barycenters of a home cluster with high consumption during the day and a work cluster in which the consumption is low during work hours. Reasons for the low but nonzero consumption of the work cluster could be (1) that the households are multiple-person households in contrast to single-person households in the first SIMDEUM data set and, hence, some of the inhabitants stay at home during the day, or (2) that the inhabitants have different daily schedules on different days of the week, e.g., four-day jobs. In summary, it can be said that the out-of-home activities (e.g., work or school) and the number of household residents are the most important high-level information revealed by the automatic clustering algorithm.

In future, the authors plan to focus on analyzing more complex data sets (1) with other important high-level information of (e.g., household income, age, and gender distribution), (2) from different countries, (3) with different end-use devices, or (4) disaggregated by end uses. Besides the clustering of consumption patterns, the proposed method is additionally valuable for (1) finding outlier patterns between different customers (e.g., unusual high water consumption) or (2) identifying changes in patterns over time (e.g., changing daily routines caused by illness or unemployment). The next research steps will concentrate on parameterizing stochastic end-use models based on this approach to provide more realistic demand simulation tools.

Conclusion

Because water demand is shaped by socioeconomic characteristics, knowledge of these characteristics and how they are connected with the dynamics of water consumption is highly valuable for WDS modeling. This work shows how data science algorithms can be used to link SWM data to high-level information. The novel SDTW clustering technique is capable of finding similarities in daily demand patterns even when they have similar features shifted in time. In this paper, the technique is tested on simulated and measured SWM data sets. It is shown for the data set where the ground truth is known that SDTW clustering is able to classify patterns accurately as well as to identify the correct number of patterns. It is shown that SDTW clustering outperforms commonly used clustering algorithms (e.g., Euclidean clustering). Furthermore, the shape of the cluster's barycenters can be linked to user characteristics. Employment status and number of household residents is identified as the most important underlying high-level information. Additionally, the methodology presented in this work can be used to identify outlier within demand patterns.

Generally, the findings of this study clearly demonstrated that socioeconomic characteristics manifest themselves in the shapes of water-usage patterns and, hence, these characteristics can be identified from the data sets through the proposed clustering approaches. Because demand patterns can be linked to high-level information, this information can be used to infer and simulate water consumption at unmeasured points in a WDS either by using directly the daily demand patterns in hydraulic simulation software, or by using a SIMDEUM model parameterized by customers' socioeconomic data. For example, in the Netherlands, socioeconomic data are freely available at a neighborhood (postcode) level from the national statistical agency. This offers the opportunity of complementing data gaps in hydraulic models and, hence, the possibility of reducing model uncertainties.

Data Availability Statement

Some or all data, models, or code generated or used during the study are available in a repository or online in accordance with funder data retention policies (https://github.com/steffelbauer/swm_sdtw).

Acknowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie Grant Agreement No. 707404. The opinions expressed in this document reflect only the authors' view. The European Commission is not responsible for any use that may be made of the information it contains.

Supplemental Materials

Tables S1–S4 and Figs. S1–S11 are available online in the ASCE Library (www.ascelibrary.org).

References

Aksela, K., and M. Aksela. 2011. "Demand estimation with automated meter reading in a distribution network." *J. Water Resour. Plann. Manage.* 137 (5): 456–467. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000131.

Alvisi, S., M. Franchini, and A. Marinelli. 2007. "A short-term, pattern-based model for water-demand forecasting." *J. Hydroinf.* 9 (1): 39–50. https://doi.org/10.2166/hydro.2006.016.

Arthur, D., and S. Vassilvitskii. 2007. "K-means++: The advantages of careful seeding." In *Proc., 18th Annual ACM-SIAM Symp. on Discrete Algorithms, SODA 2007*, 1027–1035. Philadelphia: Society for Industrial and Applied Mathematics.

Blokker, E. J. M., J. H. G. Vreeburg, S. G. Buchberger, and J. C. van Dijk. 2008. "Importance of demand modelling in network water quality models: A review." *Drinking Water Eng. Sci.* 1 (1): 27–38. https://doi.org/10.5194/dwes-1-27-2008.

Blokker, E. J. M., J. H. G. Vreeburg, and J. C. van Dijk. 2010. "Simulating residential water demand with a stochastic end-use model." *J. Water Resour. Plann. Manage.* 136 (1): 19–26. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000002.

Boyle, T., D. Giurco, P. Mukheibir, A. Liu, C. Moy, S. White, and R. Stewart. 2013. "Intelligent metering for urban water: A review." *Water (Switzerland)* 5 (3): 1052–1081.

Buchberger, S. G., J. Carter, Y. Lee, and T. G. Schade. 2003. *Random demands, travel times, and water quality in deadends*. Denver: American Water Works Association Research Foundation.

Candelieri, A. 2017. "Clustering and support vector regression for water demand forecasting and anomaly detection." *Water (Switzerland)* 9 (3): 224.

Cardell-Oliver, R. 2013a. "Discovering water use activities for smart metering." In *Proc., 2013 IEEE 8th Int. Conf. on Intelligent Sensors, Sensor Networks and Information Processing*, 171–176. New York: IEEE.

Cardell-Oliver, R. 2013b. "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.

Cardell-Oliver, R., J. Wang, and H. Gigney. 2016. "Smart meter analytics to pinpoint opportunities for reducing household water use." *J. Water Resour. Plann. Manage.* 142 (6): 04016007. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000634.

Cheifetz, N., Z. Noumir, A. Samé, A.-C. Sandraz, C. Féliers, and V. Heim. 2017. "Modeling and clustering water demand patterns from real-world smart meter data." *Drinking Water Eng. Sci.* 10 (2): 75–82. https://doi.org/10.5194/dwes-10-75-2017.

Cominola, A., M. Giuliani, A. Castelletti, D. Rosenberg, and A. Abdallah. 2018. "Implications of data sampling resolution on water use simulation, end-use disaggregation, and demand management." *Environ. Modell. Software* 102 (Apr): 199–212. https://doi.org/10.1016/j.envsoft.2017.11.022.

Cominola, A., M. Giuliani, D. Piga, A. Castelletti, and A. 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.

Cominola, A., K. Nguyen, M. Giuliani, R. A. Stewart, H. R. Maier, and A. Castelletti. 2019. "Data mining to uncover heterogeneous water use behaviors from smart meter data." *Water Resour. Res.* 55 (11): 9315–9333.

Cuturi, M., and M. Blondel. 2017. "Soft-DTW: A differentiable loss function for time-series." In Vol. 70 of *Proc., 34th Int. Conf. on Machine Learning*, edited by D. Precup and Y. W. Teh, 894–903. Cambridge, UK: Proceedings of Machine Learning Research.

Díaz, S., and J. González. 2020. "Analytical stochastic microcomponent modeling approach to assess network spatial scale effects in water supply systems." *J. Water Resour. Plann. Manage.* 146 (8): 04020065. https://doi.org/10.1061/(ASCE)WR.1943-5452.0001237.

Dürrenmatt, D. 2011. "Data mining and data-driven modeling approaches to support wastewater treatment plant operation." Doctoral thesis, Institute of Environmental Engineering, ETH Zürich.

Dürrenmatt, D. J., D. D. Giudice, and J. Rieckermann. 2013. "Dynamic time warping improves sewer flow monitoring." *Water Res.* 47 (11): 3803–3816. https://doi.org/10.1016/j.watres.2013.03.051.

Espinoza, M., C. Joye, R. Belmans, and B. De Moor. 2005. "Short-term load forecasting, profile identification, and customer segmentation: A methodology based on periodic time series." *IEEE Trans. Power Syst.* 20 (3): 1622–1630. https://doi.org/10.1109/TPWRS.2005.852123.

Garcia, D., D. González Vidal, J. Quevedo, V. Puig, and J. Saludes. 2015. "Water demand estimation and outlier detection from smart meter data using classification and big data methods." In *Proc., 2nd New Developments in IT & Water Conf.*, 1–8. London: IWA Publishing.

Gurung, T. R., R. A. Stewart, A. K. Sharma, and C. D. Beal. 2014. "Smart meters for enhanced water supply network modelling and infrastructure planning." *Resour. Conserv. Recycl.* 90 (Sep): 34–50. https://doi.org/10.1016/j.resconrec.2014.06.005.

Huang, P., N. Zhu, D. Hou, J. Chen, Y. Xiao, J. Yu, G. Zhang, and H. Zhang. 2018. "Real-time burst detection in district metering areas in water distribution system based on patterns of water demand with supervised learning." *Water (Switzerland)* 10 (12): 1765.

Hutton, C. J., Z. Kapelan, L. Vamvakeridou-Lyroudia, and D. A. Savić. 2014. "Dealing with uncertainty in water distribution system models: A framework for real-time modeling and data assimilation." *J. Water Resour. Plann. Manage.* 140 (2): 169–183. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000325.

Kwac, J., J. Flora, and R. Rajagopal. 2014. "Household energy consumption segmentation using hourly data." *IEEE Trans. Smart Grid* 5 (1): 420–430. https://doi.org/10.1109/TSG.2013.2278477.

Lin, J., S. Williamson, K. Borne, and D. DeBarr. 2012. "Pattern recognition in time series." Chap. 1 in *Advances in machine learning and data mining for astronomy*, 617–645. Boca Raton, FL: CRC Press.

Lloyd, S. P. 1982. "Least squares quantization in PCM." *IEEE Trans. Inf. Theory* 28 (2): 129–137. https://doi.org/10.1109/TIT.1982.1056489.

McKenna, S. A., F. Fusco, and B. J. Eck. 2014. "Water demand pattern classification from smart meter data." *Procedia Eng.* 70 (2014): 1121–1130. https://doi.org/10.1016/j.proeng.2014.02.124.

Monks, I., R. A. Stewart, O. Sahin, and R. Keller. 2019. "Revealing unreported benefits of digital water metering: Literature review and expert opinions." *Water* 11 (4): 838. https://doi.org/10.3390/w11040838.

Mounce, S. R., W. R. Furnass, E. Goya, M. Hawkins, and J. B. Boxall. 2016. "Clustering and classification of aggregated smart meter data to better understand how demand patterns relate to customer type."

In *Proc., 14th Int. Conf. of Computing and Control for the Water Industry—CCWI 2016*, 1–9. London: IWA Publishing.

Nambi, S. N., E. Pournaras, and R. V. Prasad. 2016. "Temporal self-regulation of energy demand." *IEEE Trans. Ind. Inf.* 12 (3): 1196–1205. https://doi.org/10.1109/TII.2016.2554519.

Nguyen, K. A., R. A. Stewart, and H. Zhang. 2014. "An autonomous and intelligent expert system for residential water end-use classification." *Expert Syst. Appl.* 41 (2): 342–356. https://doi.org/10.1016/j.eswa.2013.07.049.

Nguyen, K. A., R. A. Stewart, H. Zhang, O. Sahin, and N. Siriwardene. 2018. "Re-engineering traditional urban water management practices with smart metering and informatics." *Environ. Modell. Software* 101 (Mar): 256–267. https://doi.org/10.1016/j.envsoft.2017.12.015.

Nizar, A. H., and Z. Y. Dong. 2009. "Identification and detection of electricity customer behaviour irregularities." In *Proc., 2009 IEEE/PES Power Systems Conf. and Exposition*. New York: IEEE.

Nocedal, J., and S. J. Wright. 2006. *Numerical optimization*. 2nd ed. New York: Springer.

Rousseeuw, P. 1987. "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis." *J. Comput. Appl. Math.* 20 (1): 53–65.

Sakoe, H., and S. Chiba. 1978. "Dynamic programming algorithm optimization for spoken word recognition." *IEEE Trans. Acoust. Speech Signal Process.* 26 (1): 43–49. https://doi.org/10.1109/TASSP.1978.1163055.

Shafiee, M. E., A. Rasekh, L. Sela, and A. Preis. 2020. "Streaming smart meter data integration to enable dynamic demand assignment for real-time hydraulic simulation." *J. Water Resour. Plann. Manage.* 146 (6): 06020008. https://doi.org/10.1061/(ASCE)WR.1943-5452.0001221.

Shumway, R. H., and D. S. Stoffer. 2010. *Time series analysis and its applications: With R examples*. 4th ed. Basel, Switzerland: Springer.

Stewart, R. A., et al. 2018. "Integrated intelligent water-energy metering systems and informatics: Visioning a digital multi-utility service provider." *Environ. Modell. Software* 105 (Jul): 94–117. https://doi.org/10.1016/j.envsoft.2018.03.006.

Walski, T., D. Chase, D. Savic, W. Grayman, S. Backwith, and E. Koelle. 2003. *Advanced water distribution modeling and management*. 1st ed. Waterbury, CT: Haestead.

Wang, J., R. Cardell-Oliver, and W. Liu. 2015. "Efficient discovery of recurrent routine behaviours in smart meter time series by growing subsequences." Vol. 9078 of *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)*, 522–533. New York: Springer.

Witten, I. H., E. Frank, and M. A. Hall. 2011. *Data mining: Practical machine learning tools and techniques*. 3rd ed. San Francisco: Morgan Kaufmann.

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 (Switzerland)* 10 (9): 1221. https://doi.org/10.3390/w10091221.

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