Water Resources Research

RESEARCH ARTICLE
10.1029/2019WR024897

Special Section:
Big Data & Machine Learning in Water Sciences: Recent Progress and Their Use in Advancing Science

Data Mining to Uncover Heterogeneous Water Use Behaviors From Smart Meter Data

A. Cominola[1], K. Nguyen[2,3], M. Giuliani[4], R. A. Stewart[2,3], H. R. Maier[5], and A. Castelletti[4]

[1]Chair of Smart Water Networks, Technische Universität Berlin - Einstein Center Digital Future, Berlin, Germany, [2]School of Engineering and Built Environment, Griffith University, Gold Coast, Australia, [3]Cities Research Institute, Griffith University, Gold Coast, Australia, [4]Department of Electronics, Information and Bioengineering, Politecnico di Milano, Milan, Italy, [5]School of Civil, Environmental and Mining Engineering, University of Adelaide, Adelaide, South Australia, Australia

Key Points:
- We investigate which water use behaviors can be identified from nonintrusive, single-point, smart meter data
- We identify and cluster primary water use behaviors of single-family households from disaggregated end use data
- We reveal the main water end uses driving different behaviors, usage patterns, and regularity, to support customized demand management

Correspondence to:
A. Cominola,
andrea.cominola@tu-berlin.de

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

Received 31 JAN 2019
Accepted 8 OCT 2019
Accepted article online 16 OCT 2019
Published online 19 NOV 2019
Corrected 17 FEB 2020

This article was corrected on 17 FEB 2020. See the end of the full text for details.

Abstract Knowledge on the determinants and patterns of water demand for different consumers supports the design of customized demand management strategies. Smart meters coupled with big data analytics tools create a unique opportunity to support such strategies. Yet, at present, the information content of smart meter data is not fully mined and usually needs to be complemented with water fixture inventory and survey data to achieve detailed customer segmentation based on end use water usage. In this paper, we developed a data-driven approach that extracts information on heterogeneous water end use routines, main end use components, and temporal characteristics, only via data mining existing smart meter readings at the scale of individual households. We tested our approach on data from 327 households in Australia, each monitored with smart meters logging water use readings every 5 s. As part of the approach, we first disaggregated the household-level water use time series into different end uses via Autoflow. We then adapted a customer segmentation based on eigenbehavior analysis to discriminate among heterogeneous water end use routines and identify clusters of consumers presenting similar routines. Results revealed three main water end use profile clusters, each characterized by a primary end use: shower, clothes washing, and irrigation. Time-of-use and intensity-of-use differences exist within each class, as well as different characteristics of regularity and periodicity over time. Our customer segmentation analysis approach provides utilities with a concise snapshot of recurrent water use routines from smart meter data and can be used to support customized demand management strategies.

1. Introduction

After more than 20 years since the first experimental developments of smart water metering (Mayer et al., 1999), the opportunities offered by digital technologies and augmented data mining capabilities, combined with the threats posed by worsening climatic, infrastructural, and societal challenges, are increasingly pushing the water utility sector to transition to the digital age (Beal & Flynn, 2015; Cheong et al., 2016; Turner & White, 2017). There is now a rich literature reporting on the actual and potential financial, operational, and environmental benefits that digital technologies can bring to the water utility sector. Mining of enhanced sensor data collected with fine spatiotemporal resolution improves utilities' understanding of the evolving status of their network assets (Sensus, 2012), as well as their ability to design both supply- and demand-side management (DSM) strategies (Cominola et al., 2015; Escriva-Bou et al., 2015; Stewart et al., 2010, 2013). In this context, a wide range of contributions in the scientific and technical literature envisions the digital transformation of water utilities and the adoption of digital technologies for the development of smart water networks (Tsakalides et al., 2018). Yet the lack of strong business cases supporting such a digital transformation, along with several research and development priorities that need to be addressed at the levels of *strategy, information*, and *technology*, hinder the realization of this vision (Stewart et al., 2018).

Among the primary R&D priorities for the digitalization of water utilities, data accessibility and the integration of new sensor data with suitable data management and analytics are key. Extracting information on water use behavior from individual or groups of water consumers based on smart meter data is becoming increasingly important to effectively support demand modeling and the design of customized DSM strategies (Cominola et al., 2015). There is growing evidence about the potential of using smart meter-based feedback

mechanisms to manage water demand: experimental studies demonstrate that water demand reductions between 2.5% and 28.6% can be obtained with consumption feedback in near real time, as consumers are better informed about their water consumption habits and conservation efforts (Sønderlund et al., 2016). Further studies indicate that both alternative demand management targets (e.g., demand peak shifting; Beal et al., 2016) and tailored price- and non-price-based incentives schemes (Novak et al., 2018; Rougé et al., 2018) can be informed via the discovery of the key determinants and patterns of water demand from individual customer data. Better data tracking and reporting would also enable utilities to monitor behavioral responses to demand management and rebound effects in the long term (Gonzales & Ajami, 2017). Finally, analysis of smart meter traces of individual customers supports prompt detection of leakages (Luciani et al., 2019) and other types of anomalies in residential and nonresidential accounts (Patabendige et al., 2018).

The need to identify water use components, demand drivers, and water use patterns of individual consumers to inform customized DSM strategies has fostered the development of several recent works mainly looking at either (i) disaggregating single end use appliance use patterns from aggregate household water consumption meter data, taken individually or coupled with energy use information (Nguyen et al., 2015; Vitter & Webber, 2018), or (ii) modeling water demand patterns to identify time-of-use characteristics, anomalous behaviors, demand-driving factors, and forecast actual or statistically generated demand patterns over time (Blokker et al., 2010; Cominola, Giuliani, et al., 2018; Creaco et al., 2016; Duerr et al., 2018; Kofinas et al., 2018). Only a few recent studies have looked at segmenting water consumers into groups showing similar water demand characteristics (e.g., demand volumes and patterns). Data-driven customer segmentation has been widely adopted in urban studies to characterize groups of customers with heterogeneous recurring or irregular behavioral patterns of transportation network utilization and other socioeconomic activities, such as working, sleeping, and leisure activities (e.g., Cardell-Oliver & Povey, 2018; Poussevin et al., 2016; Yuan et al., 2014) and electricity consumption (e.g., Espinoza et al., 2005; Kwac et al., 2014; Nambi et al., 2016). Yet this kind of analysis has been tested only recently in the water sector. Building on previous research on pattern identification from smart meter data (Cardell-Oliver, 2013a; 2013b), Cardell-Oliver et al. (2016) developed a five-stage customer segmentation method aimed to identify "regular high-magnitude behaviors", that is, water use behaviors that occur at regular intervals, from smart metered data collected at hourly time intervals. Regular high-magnitude behaviors reveal who uses water, as well as when and how they use it and, in combination with survey data on household characteristics, enable the interpretation of predominant water use behaviors. This is particularly informative for targeting water conservation efforts to relevant customer groups: for instance, Cardell-Oliver et al. (2016) found that only 12–15% of the households from two case studies in Western Australia caused over 80% of the peak-hour demand. Similarly, previous work by the authors (Cominola, Spang, et al., 2018) contributed a three-phase customer segmentation analysis to discriminate among heterogeneous water-electricity demand routines and provide insights for coordinated water-energy DSM, based on water and electricity usage data metered with hourly sampling frequency and survey information on objective and subjective sociopsychographic information. In addition, Laspidou et al. (2015) applied Kohonen Self-Organized Maps to quarterly water-billing data from heterogeneous water accounts, demonstrating the potential to discriminate both among different categories of accounts (residential/nonresidential) and different types of businesses among nonresidential accounts. Finally, Candelieri (2017) recently contributed a clustering approach to identify a variety of anomalies, such as smart meter faults, frauds, cyberphysical attacks, or meaningful changes in water consumption habits, from individual customer hourly water consumption data.

All of the above studies contributed customer segmentation methods to identify water consumption patterns and gain knowledge of the key determinants of water demand. While the benefits of using information on water consumption patterns and determinants to formulate DSM strategies have not yet been quantified, such information can be used to design customized DSM strategies and test their effectiveness in comparison with less customized approaches. However, these previous approaches only used aggregate household water consumption data; thus, the information content of smart meter data is not fully mined at the level of end use components. Consequently, identifying heterogeneous segments of water consumers, that is, groups of water consumers that share similar characteristics of water consumption but differ from the consumers in other groups, based on end use traces is still an open research challenge. While some previous studies have considered the cross correlation of end use component information together with water consumers' characteristics (Horsburgh et al., 2017; Quesnel & Ajami, 2018; Willis et al., 2013), they have looked at investigating the potential determinants of specific residential or nonresidential water end uses, often monitored

with dedicated meters, rather than identifying clustered routines and categories of water consumers based on their end use demand patterns. Identification of the above determinants traditionally requires the collection of additional information via (i) intrusive submetering campaigns performed to gather ground-truth training data for disaggregation algorithms (Cominola et al., 2015) and (ii) survey campaigns to gather complementary data on water users' demographics and psychographics (Russell & Fielding, 2010). However, both submetering and survey campaigns have a number of disadvantages, including sensor costs and the intrusiveness of submetered ground-truth data collection, time and resources consuming expert-based data processing in the absence of automated data mining, and privacy issues when running audits and campaigns to gather water fixture inventory and survey data. All of the above may limit the acceptance, technical feasibility, and ability to update behavioral studies. Consequently, developing customer profiling and segmentation methods that do not rely on intrusive survey campaigns and that can be scaled at the utility level also presents a research challenge.

In this paper, we demonstrate a data mining procedure to address the two research challenges described above. First, we aim to extract information on the existence of heterogeneous water use routines at the level of end use components (e.g., shower, tap, and irrigation), along with their behavioral regularity and temporal characteristics. Second, we only rely on data mining of smart meter data at the scale of individual households without requiring any other information. Our approach relies on a completely data-driven, end use-based, water consumer segmentation analysis to (i) identify recurring water end use routines, based on water use time series broken down at the end use level, (ii) group water use accounts that reveal similar water end use profiles, and (iii) characterize these profiles in terms of primary end use determinants, behavioral regularity, and periodicity over time, ultimately generating customized feedback to inform demand management and water conservation interventions.

To the authors' knowledge, no research currently exists that exploits solely smart meter data to couple nonintrusive end use disaggregation of water consumption with customer segmentation for understanding the primary behavioral routines of water consumers and identifying groups with heterogeneous behaviors across an ensemble of households. Our methodology relies uniquely on subminute resolution water use data collected via single-point household water meters and couples pretrained state-of-the-art end use disaggregation algorithms with customer segmentation based on eigenbehavior analysis. The main advantage of our customer segmentation analysis is that it enables a characterization of water consumers from smart meter data, at the level of individual end uses and without the need for additional data that are more difficult, more time consuming and more expensive to collect (e.g., survey data). This is likely to provide a stronger business case for increased usage of metering technology.

We demonstrated the capabilities of our customer segmentation analysis working on data from 327 single-family households located in Southeast Queensland (SEQ) and Melbourne (Australia), each monitored with smart meters logging water use data with a sampling frequency interval of 5 s for a time span of about 10 months in 2010. Revealing the heterogeneous water use behaviors and primary routines contained in these data enabled us to identify behavioral characteristics that are key to formulating customized demand management strategies and informing tailored conservation feedback while avoiding intrusive and expensive data collection campaigns.

The rest of the paper is organized as follows. In section 2 we introduce the end use-based water consumer segmentation analysis approach, followed by a description of the study site and the experimental settings we considered in this study in section 3. In section 4 we report and discuss the numerical results. We make final remarks and present opportunities for further research in the last section.

2. Material and Methods

The proposed end use-based water customer segmentation analysis approach uses data mining to aid the discovery of residential water end use behaviors from water use data metered at subdaily sampling resolution by means of single-point intelligent meters, without the need for complementary information on end use fixtures and household demographics. In keeping with the objective mentioned in section 1, we investigated which water end use behaviors can be identified by expanding the information content of aggregate water use data sampled with high temporal frequency, along with their main end use components, behavioral regularities, and temporal characteristics. The consumer segmentation analysis was performed via the method conceptualized and detailed in Figure 1.

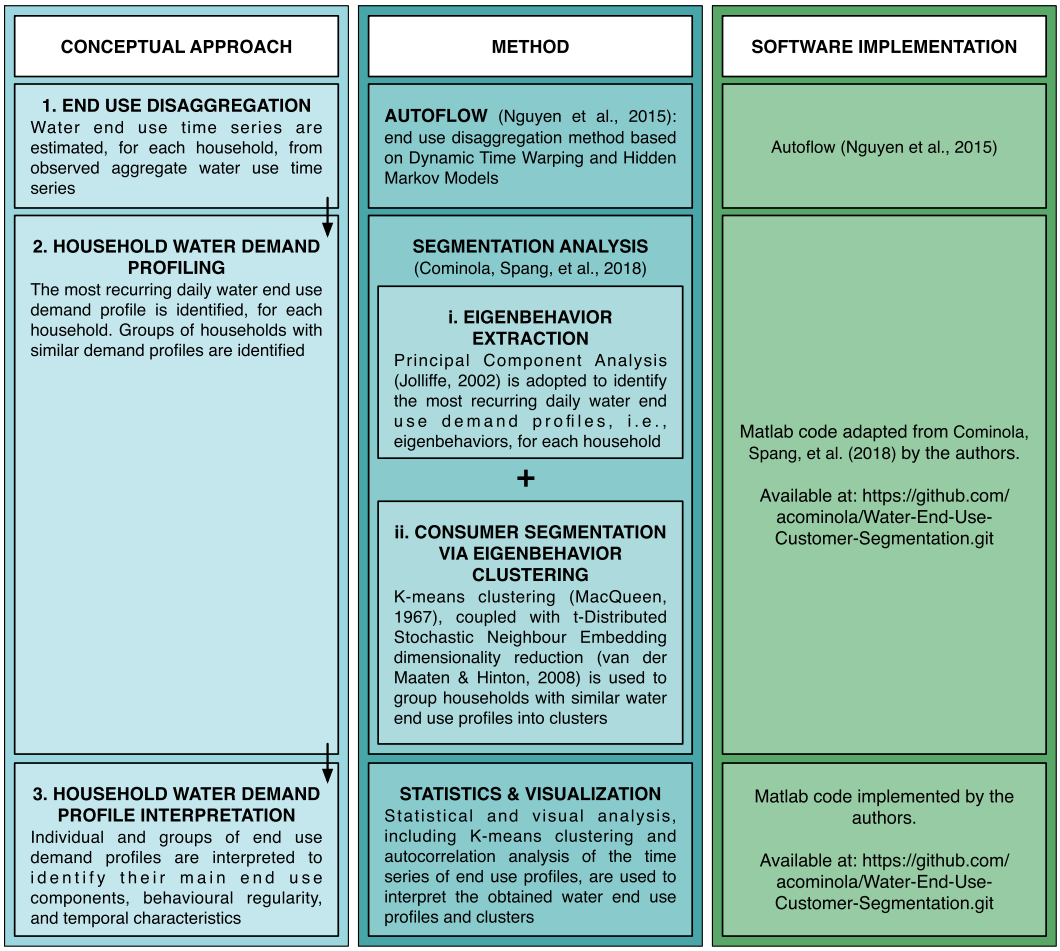

Figure 1. Flowchart of the end use-based customer segmentation analysis. Conceptual, methodological, and software implementation levels are visualized.

At the conceptual level (left column in Figure 1), we first disaggregated the water use time series obtained for each household with the aid of single-point smart meters into candidate contributing end uses (e.g., tap, shower, and washing machine). Second, we modeled the most recurring daily water end use demand profiles of each household (i.e., routines represented by typical patterns of daily water end use) and identified groups of consumers with similar demand profiles. Finally, we interpreted the identified individual profiles and household groups in order to pinpoint predominant end use components and behavioral and temporal characteristics. From a methodological point of view (central column in Figure 1), each conceptual phase of customer segmentation analysis was implemented as follows:

1. *End use disaggregation*. We performed the end use disaggregation by means of *Autoflow*, the intelligent autonomous system for residential water end use classification developed by Nguyen et al. (2015). The algorithm combines hidden Markov models and artificial neural networks with dynamic time warping pattern recognition to disaggregate water end uses from an aggregate time series of water use metered at the household level. After learning from training end use data, the algorithm estimates the unobserved substates (i.e., the water use time series y_t^i of each fixture i) by solving a *blind identification* problem (Abed-Meraim et al., 1997) formulated as follows:

$$\bar{Y}_t = \sum_{i=1}^{N} y_t^i + e_t \tag{1}$$

where $\bar{Y}_t$ is the observed water use as metered at each time step via single-point smart meters, y_t^i is the water use of fixture i at time step t, N is the total number of fixtures, and e_t is the measurement noise. *Autoflow* has been demonstrated to achieve end use classification accuracies ranging from 85.9% to 96.1%

for single events (i.e., time periods when only one fixture at the time is used) and 81.8–91.5% for combined events (i.e., time periods when more than one fixture is used simultaneously), relying on water use data metered with subminute frequency (Nguyen et al., 2015).

2. *Household water demand profiling*. In the second phase of the consumer segmentation analysis, we adapted the customer segmentation method based on eigenbehavior modeling proposed in Cominola, Spang, et al. (2018) to account for heterogeneous household water end use profiles. The segmentation analysis phase is composed of two sequential steps.

   The first (i in Figure 1) consists of the so-called *eigenbehavior extraction* (Eagle & Pentland, 2009): we extracted the principal daily water demand profile of each household for each end use fixture. For each household j, such profiles are extracted via Principal Component Analysis (PCA; Jolliffe, 2002) performed on an input matrix Γ^j with dimension $D \times (U \times N)$, where D is the number of observed days in the water use time series of each household, U the number of water end uses and N the number of observations per day (e.g., 24, if data sampled at a 1-hr frequency are considered). As was done in Cominola, Spang, et al. (2018) and Giuliani and Herman (2018), the principal daily water end use profiles, called *eigenbehaviors*, were computed for each household as the eigenvectors w_q of the covariance matrix Q of Γ^j:

$$Q = \frac{1}{D}\sum_{d=1}^{D}(\Gamma^j)^\top \cdot \Gamma^j \tag{2}$$

   Those eigenbehaviors accounting for most of the variance in the input data are considered *routines*, that is, recurrent demand profiles representing frequent behaviors. The Γ^j input matrix can have high dimensionality when either the observed time series are long (i.e., D is high) or data are sampled with high frequency (i.e., N is high). The main benefit of applying eigenbehavior extraction on these water end use data is that it provides a reduced-dimension representation of the above input matrix, composed of a set of $1 \times (U \times N)$ vectors that explain the maximal variance in the original data. Moreover, a subset of these vectors is usually sufficient to explain a large part of the variance in the input data. Each of the variables in the input data matrix was centered on its mean value before performing PCA.

   The second step of the household water demand profiling is the *consumer segmentation via eigenbehavior clustering* (ii in in Figure 1). In this phase, we clustered the principal water end use profiles obtained for each household based on their similarity as measured by Euclidean distance. In keeping with the method in Cominola, Spang, et al. (2018), we sequentially used t-Distributed Stochastic Neighbor Embedding (t-SNE; van der Maaten, 2014; van der Maaten & Hinton, 2008) and K-means clustering (MacQueen, 1967) to obtain K heterogeneous clusters of similar end use profiles. t-SNE projects the eigenvectors into a two-dimensional space and facilitates the choice of the number K of cluster sets in K-means. The quality of the clustering outcome, that is, cluster separation and cohesion, can be sensitive to t-SNE and K-means parameter selection. In this work, we assessed the resulting cluster quality for different combinations of t-SNE parameterization and number of K-means clusters K by means of the silhouette coefficient (Rousseeuw, 1987). The silhouette coefficient evaluates cluster quality as a combination of cluster cohesion and separation, as follows:

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}} \tag{3}$$

   where $s(i)$ is the silhouette value of data point i, $a(i)$ the average distance between data point i and the other points in its cluster, and $b(i)$ the minimum average distance between i and all points belonging to other clusters (where i is not included). Based on the values assumed by the average of the silhouette coefficient over all data points, we set the best performing combination of t-SNE and K-means parameters.

3. *Household water demand profile interpretation*. Lastly, we interpreted the water end use profiles obtained from Phase 2 to answer the following three research questions: (Q1) Does any end use component emerge to characterize the main water end use profiles? (Q2) Do these profiles reveal behavioral differences due to, for instance, changing human activities between weekdays and weekends? (Q3) Do the main profiles show regular/periodic behaviors over time? First, we addressed Q1 via a mix of visual analysis and basic statistics of the results obtained from Phase 2. This enabled us to investigate how the end use profiles were heterogeneous in terms of daily water use patterns, peaks, and time of use of the main end use components. Second, we addressed Q2 by calculating how frequently households shifted between primary

and secondary behaviors during different days of the week. Finally, we addressed Q3 by looking at whether the autocorrelation of the main water end use profiles showed any periodicity over time.

Overall, we expected clustering, coupled with end use disaggregation and eigenbehavior extraction, to enable a fine-scale, end use based, customer segmentation analysis, which, in turn, allows the main types of water end use demand profiles across an ensemble of metered households to be revealed and characterized in terms of relative impact, time of use of each water end use, and adoption frequency. Interpretation of these outputs provided contextualized end use information on primary water use behaviors and time-of-use routines. Understanding to which extent our analysis allowed the above information to be extracted from smart meter data without other a priori information is key to support the design of customized feedback to improve consumers' perception of their water use, plan conservation actions, and facilitate the identification of anomalous behaviors repeated in time, due to, for instance, extreme water use behaviors, leakages, or incorrect meter reading.

3. Study Site and Experimental Settings

In this study, we applied the proposed end use-based customer segmentation analysis introduced in section 2 to anonymous water use data metered for 327 single-family households located in SEQ and Melbourne, Australia. The SEQ region of Australia has a subtropical climate and a population of approximately 3.5 million people. The average annual temperature is 19 °C, summer average temperature is 24 °C, 20 °C in spring and autumn, and 14 °C in winter. In 2009, SEQ emerged from one of its harshest and most protracted droughts on record. The variability of rainfall in the region, combined with high population growth and strong economic development, led to the implementation of various integrated urban water management strategies. In an attempt to improve water security, many state government agencies and water authorities in the region imposed water restrictions and water saving measures, to manage demand and ensure more considered use of water across residential, commercial, and industrial sectors.

Beal et al. (2011) conducted a comprehensive residential water end use study using a sample of 252 residential dwellings located in the study area. Existing standard water meters were replaced with high-resolution meters capable of providing 0.014-L/pulse outputs in 5-s intervals to wireless data loggers. A representative sample of received data was extracted from the database and disaggregated into all end use events associated with the sampled residential households (e.g., toilet, tap, and shower). This repository of residential water end use data underpinned the customer segmentation analysis study presented in this paper. Each household was monitored with smart meters logging water use data with a sampling frequency interval of 5 s for a time span of about 10 months from 15 January to 16 October 2010. A full description of the data set used as input data for this present study is provided in Beal et al. (2011). In addition, to improve the generality of our analysis and the results obtained, we selected a small subset of 75 homes from Melbourne to combine with the 252 homes from SEQ for this study. To ensure consistency for data analysis and consumption segmentation, high-resolution data in Melbourne were collected during the same period as in Queensland (February–December 2010), and these 75 homes were manually chosen to have very similar demographic background to the ones from SEQ in terms of number of occupants, age, and occupation. The primary water end use disaggregation tasks conducted have shown a high similarity in water consumption behavior between these two data sets, which has ensured consistency in the results for the subsequent household segmentation analysis.

Details of how the general method for customer segmentation analysis we presented in Figure 1 was applied to the case study region are given in Figure 2, showing the relationship between the data, the analysis methods, and the outcomes. Eight potential end use categories contribute to the total household water usage of each household, namely, tap, shower, washing machine, dishwasher, toilet, bathtub, irrigation, and evaporative cooler. We used the water use time series metered at 5-s sampling frequency for each household as input to the end use disaggregation phase of the customer segmentation analysis (top left corner in Figure 2). We then aggregated the obtained end use time series to hourly and 6-hr levels to feed into the household water demand profiling phase (top right corner in Figure 2). In this specific application, the aggregation of data to hourly or larger time intervals was needed by our approach to ensure the number of observations (rows of Γ^j) to be higher than the number of variables (columns of Γ^j), before performing the eigenbehavior extraction. This constraint was formulated as $D \geq U \times N$, where D is the number of observed days in the time series of each household, U the number of end uses, and N the number of observations per day. In our

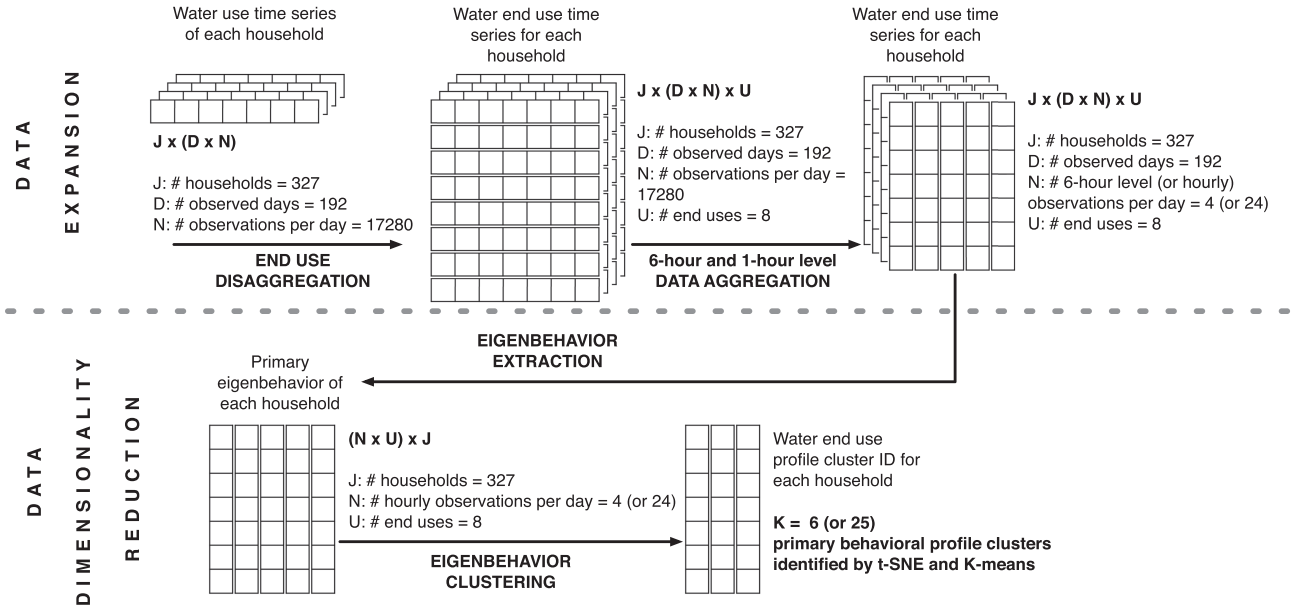

Figure 2. Detailed flowchart of the end use-based customer segmentation analysis applied onto water consumption data from 327 households located in Southeast Queensland and Melbourne, Australia. Data dimensions are highlighted across the different steps.

case study, the minimum value found for D is equal to 192 days and U is equal to 8; therefore, the number of observations per day N could not be higher than 24. We considered both 1- and 6-hr time aggregation intervals to evaluate how sensitive the segmentation outcome is to the choice of time boundaries, especially considering that some water usage activities (e.g., washing or irrigation) can last longer then 1 hr. In addition, we performed a two-step preprocessing analysis on the data: first, we removed the observations for incomplete days at the beginning and end of the time series. Second, we removed all anomalous meter readings, which are most likely due to erroneous meter readings (i.e., reading values larger than 1,000 L/h and the top 5% positive hourly water use readings of each user). With this procedure, we deleted less than 2% of available observations for each household.

It is worth noticing how we facilitated the discovery of water use behaviors from single-point smart meter data by sequentially performing data expansion and data dimensionality reduction. First, we expanded the amount of information available via end use disaggregation to estimate water end uses from aggregated meter readings. The amount of data was thus initially increased by a factor of 8, equal to the number of considered end uses. Then, we reduced the dimensionality of the expanded data set via PCA and clustering to obtain a concise snapshot representation of the main water use routines, defined as daily water end use patterns. The amount of data was thus reduced to K primary behavior profile clusters, each represented by 24 or 4 (data points per day) × 8 (end uses) routines if only the primary eigenbehavior was considered. In such a way, we combined a full exploitation of the information content of metered readings with a compact representation of primary behaviors that can be used to inform demand management programs at the utility/municipality level.

4. Results

In this section, we assess whether the end use-based segmentation analysis we developed can be used to extract information on heterogeneous water use routines at the level of end use components, their behavioral regularity, and temporal characteristics. We queried our results referring to the three questions we formulated in section 2:

Q1. Does any end use component emerge to characterize the primary water end use profiles?
Q2. Do these profiles reveal behavioral differences due to, for instance, changing human activities between weekdays and weekends?
Q3. Do the main profiles show regular/periodic behaviors over time?

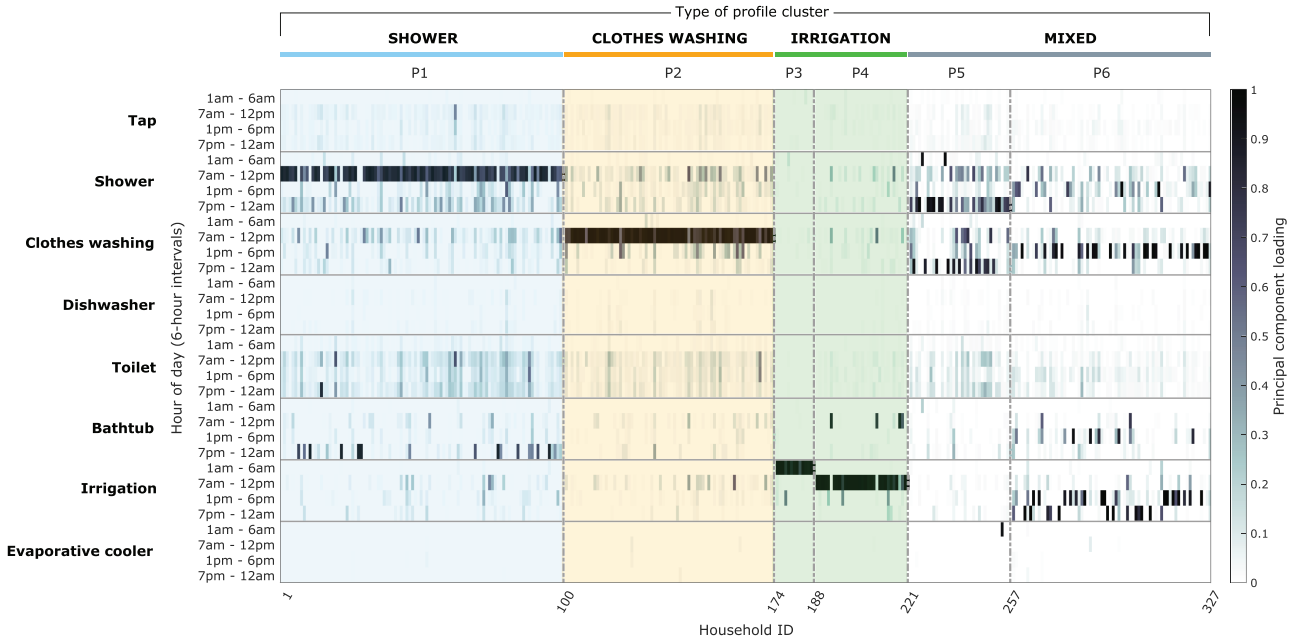

Figure 3. Daily water end use eigenbehaviors for the 327 single-family household in Southeast Queensland and Melbourne. Each column matrix represents the loadings of the first eigenbehavior for one household (each household is identified by a household ID, see *x* axis). Eigenbehavior loadings are reported for eight diverse water end uses over four daily periods of 6 hr (*y* axis). Matrix cell foreground color is proportional to principal component magnitude (see color bar). Eigenbehaviors are clustered in six clusters based on their similarity: profile clusters are identified by the Labels P1, P2, … , P6 on top of the figure and the vertical dashed lines. Three main types of profile clusters (characterized by shower, clothes washing, or irrigation end use) are identified with the light blue, orange, and green background color. A fourth profile cluster with mixed end uses is also reported. Negative loadings were filtered out before visualization.

4.1. Identification and Characterization of Primary Water End Use Profiles

To answer Q1, we assessed whether the household water end use profiles obtained via eigenbehavior extraction (see section 2) from the disaggregated water use observations revealed insights into primary end uses and routines that are informative for demand management. As described in section 2, a household water end use profile in our analysis represents frequently adopted daily water use routines across candidate indoor and outdoor residential water end uses, and is called eigenbehavior.

4.1.1. Identification of Primary Water End Use Profiles

The first eigenbehavior for every household, that is, the eigenbehavior that explained most of the variance in the input data, is visualized in Figure 3. Each column contains the principal component loadings for a single household. They weight the different end uses for aggregated subdaily periods of 6 hr ([1 a.m. to 6 a.m.], [7 a.m. to 12 p.m.], [1 p.m. to 6 p.m.], and [7 p.m. to 12 a.m.]); thus, each column contains four data points for each end use, and the color is proportional to the magnitude of the eigenbehavior loadings. The results show that a variety of eigenbehaviors is found across the observed 327 households. Our end use-based consumer segmentation analysis clustered the households into six heterogeneous profile clusters, labeled as P1, P2, … , P6 in the figure, according to the silhouette coefficient (see section 2 and Appendix A for further information on the selection of the number of clusters). Actually, we could observe that three main primary end uses are revealed, highlighted with different background colors in the figure. First, profile cluster P1 shows behavioral profiles characterized by *shower* end use, as eigenbehavior loadings assume high values only for the shower end use, mainly for hours of the day (*y* axis) falling in the morning (between 7 a.m. and 12 p.m.), meaning most of the variance in the data is explained by shower end use. Second, profile cluster P2 reveals behavioral profiles characterized by *clothes washing* end use, and third, profiles P3 and P4 by *irrigation* end use. In addition, profiles P5 and P6 appear to be *mixed*, in the sense that eigenbehavior loadings assume high values for more than a single end use: P5 is mainly characterized by shower and clothes washing end uses, while P6 shows scattered behaviors with an emphasis on shower, clothes washing, bathtub, and irrigation end uses.

Figure 4 is structured similarly to Figure 3, but represents the eigenbehavior as obtained from data aggregated to 1-hr intervals; thus, each column contains 24 data points for each end use. Three relevant aspects

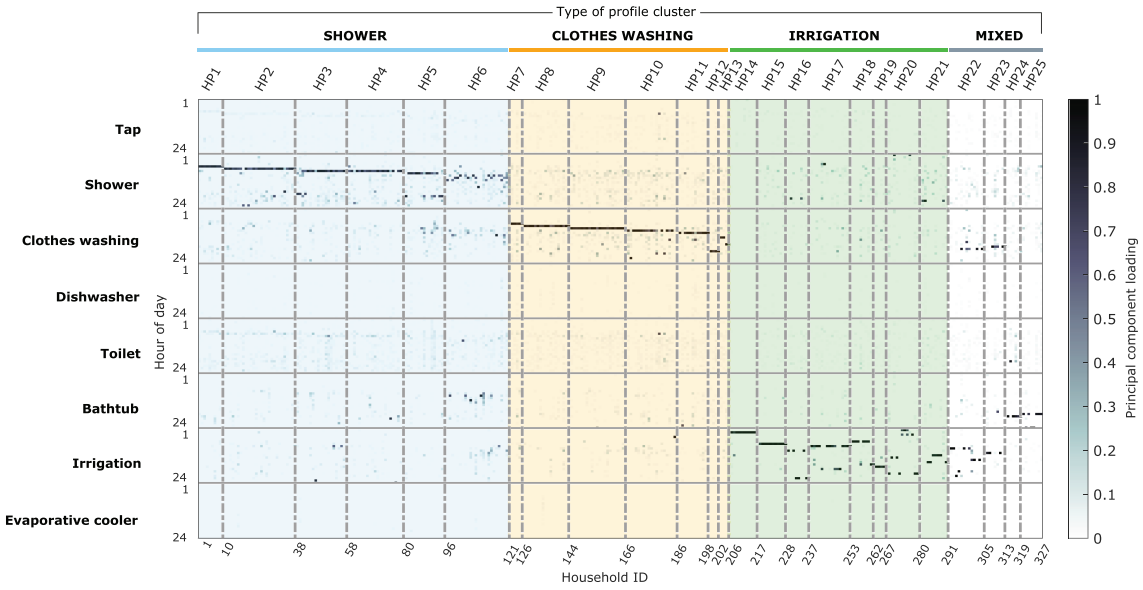

Figure 4. Daily water end use eigenbehaviors for the 327 single-family households in Southeast Queensland and Melbourne. Each column matrix represents the loadings of the first eigenbehavior for one household (each household is identified by a household ID, see *x* axis). Eigenbehavior loadings are reported for eight diverse water end uses over 24 hr (*y* axis). Matrix cell foreground color is proportional to principal component magnitude (see color bar). Eigenbehaviors are clustered in 25 clusters based on their similarity: hourly profile clusters are identified by the labels HP1, HP2, … , HP25 on top of the figure and the vertical dashed lines. Three main types of profile clusters (characterized by shower, clothes washing, or irrigation end use) are identified with the light blue, orange, and green background color. A fourth profile cluster with mixed end uses is also reported. Negative loadings were filtered out before visualization.

emerge from a joint analysis of Figures 3 and 4. First, the number of profile clusters found to best group the 327 households in Figure 4 amounts to 25. This is 4 times higher than the six profiles found in the previous analysis, because data with a sampling resolution that is higher than 6 hr enable time-of-use differences to be characterized better. Second, the three primary behavioral profile clusters identified in Figure 3 are still present: behavioral profiles characterized by shower, clothes washing, and irrigation end uses (see Figure 4). However, the distribution of households among the three main behavioral profile clusters changes between the analyses. Finally, the number and clarity of mixed profiles is reduced in Figure 4: only a few households belonging to HP22 to HP25 show more than one prevailing end use, or a primary end use (i.e., bathtub) different than shower, clothes washing, or irrigation. This means that time aggregation has an effect on how specific end uses emerge, which is likely because the duration of some end uses can span more than 1 hr and, thus, becomes visible only when the data are aggregated.

To better visualize the three types of primary behavioral profile clusters identified in the analysis with hourly data, that is, those characterized by shower, clothes washing, and irrigation end uses, Figure 5 shows the average principal component loading for each of these end uses, in addition to the mixed profile, over a 24-hr period. The figure confirms that, for each of the three types of routines, one specific water end use explains most of the variance in the data. Shower (top left subplot), clothes washing (top right subplot), or irrigation (bottom left subplot) are weighted, on average, with higher principal component loadings compared with other end uses, especially for some hours of the day or night. The additional contribution from the analysis of hourly data is twofold. First, demand peaks can be better identified as daily patterns become more detailed with higher time sampling frequency. Second, subclusters with routines that differ in terms of time of use of water during the day emerge within the same end use typology (e.g., HP1 to HP6 are all characterized by shower end use, but they differ in terms of showering time). Time of shower use differs only slightly across the profiles characterized by shower end use for different households, because most of them reveal overall peaks of shower use during morning hours, similar to findings from other studies (Beal & Stewart, 2014). Similarly, time of use gradually changes across household profiles characterized by clothes washing end use, with a minority of households (HP12 and HP13) using washing machines more frequently in the afternoon/late afternoon, while most of the other household clusters (HP7 to HP11) show preferred use in the morning hours. Even more heterogeneity is revealed by household profiles characterized by irrigation

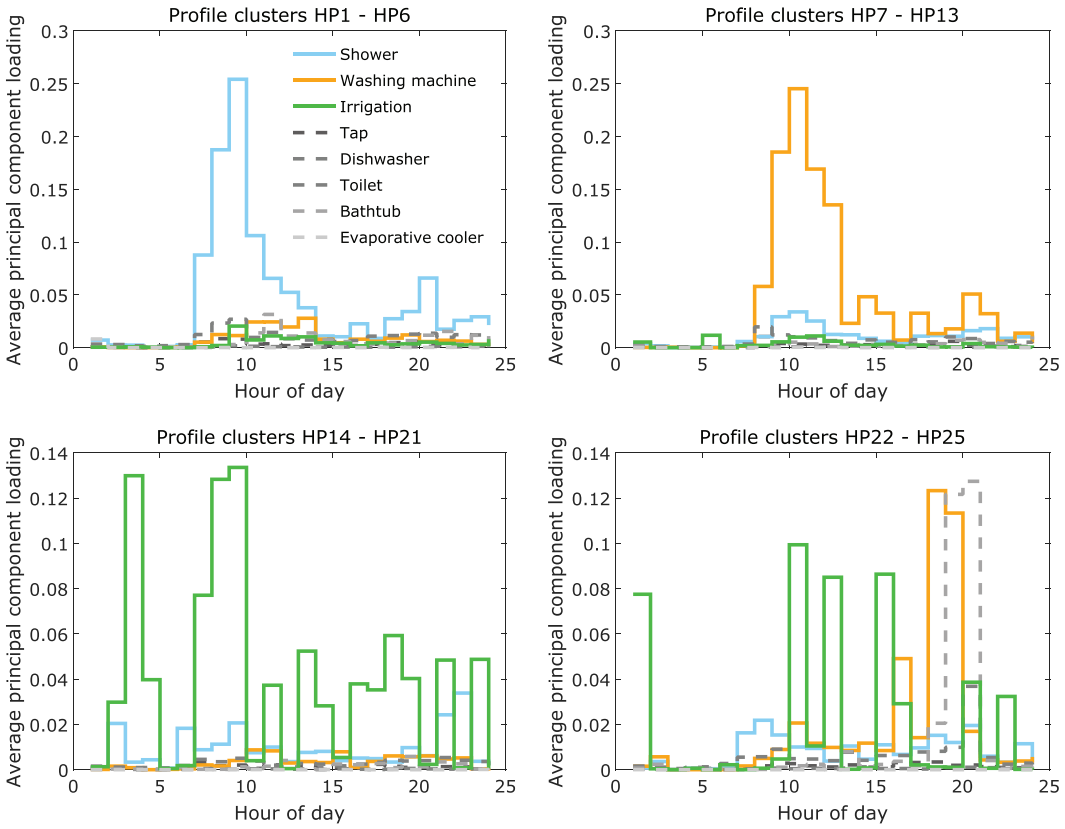

Figure 5. Average principal component loading, for four different types of profile clusters, eight end uses, and different times of day.

end use, as they reveal irrigation habits that span across most day and night hours, likely depending on whether timers for automatic watering control are used.

The results visualized in Figures 3–5 are particularly informative about the main water use behaviors across the observed community of water users, as well as their regularity. The relative weight of the eigenbehavior loadings across different end uses, considered together with their magnitude values (these are usually larger than 0.5 for primary end uses and time of use—see color bar values in Figure 4), suggests that the main regularities of household water use behavior can be captured by one, or a few, primary end uses and times of use. This implies that eigenbehavior analysis can be used in practical settings to tailor water use feedback for different categories of users, based on their revealed habits of water use across different end uses and demand patterns. For instance, customized feedback aimed at reducing the amount of water used for irrigation during day hours on a hot day can be formulated to target households characterized by irrigation end use and avoid inefficient water use at times of scarcity. As another example, customized feedback aimed at lowering the morning peak flow would likely target households characterized by shower end use to shift their water use from those hours to alternative times in the day, when peak consumption is lower.

Considering the overall consistency between the household profiles obtained from data aggregated to 1- or 6-hr intervals, in the following sections we follow up with the characterization of primary end use profiles referring only to those aggregated with subdaily periods of 6 hr.

4.1.2. Characterization of Primary Water End Use Profiles

As illustrated in the above examples, customized feedback can be better tailored to effectively influence the water use behavior of target users if we have a better understanding of the determinants of their water demand patterns. By using the proposed approach, this can be achieved by better characterizing the primary behavioral profiles identified by looking at simple aggregate statistics derived a posteriori from the water end use data of each profile cluster. In Figure 6, we visualize the relative share of each water end use for each profile cluster of Figure 3, averaged across the households that belong to it. The figure demonstrates that the end uses previously identified as primary across behavioral profiles, that is, shower, washing

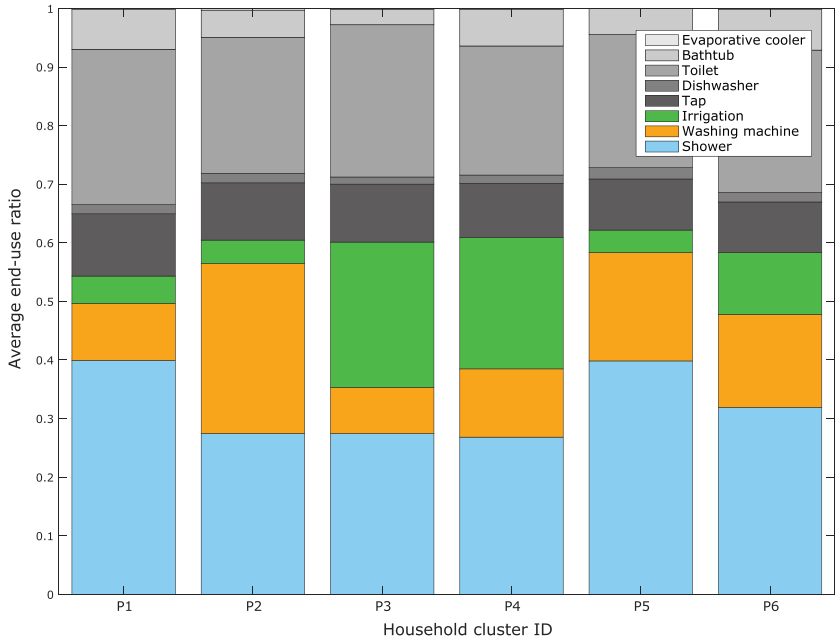

Figure 6. Average water end use share for each profile cluster P1, P2, … , P6. Profile clusters are sorted on the *x* axis as in Figure 3. The relative water end use ratio is reported in the *y* direction, calculated as an average across the households belonging to the same cluster. Different colors represent different residential water end uses.

machine, and irrigation, have a larger share of total water use for those profile clusters that exhibit higher eigehbehavior loadings related to those end uses, compared with other profiles, despite not always having the absolute largest share. For example, the share of water used for irrigation is generally larger for profile clusters characterized by irrigation end use, compared to all other profiles. Similar characteristics are exhibited by shower end use for the profile cluster characterized by shower end use, and by washing machine for the profile clusters characterized by clothes washing end use, which also emerge in mixed profiles P5 and P6. It is not surprising that the end uses that characterize the main behavioral profile clusters identified by our approach also account for the largest relative magnitude of water use.

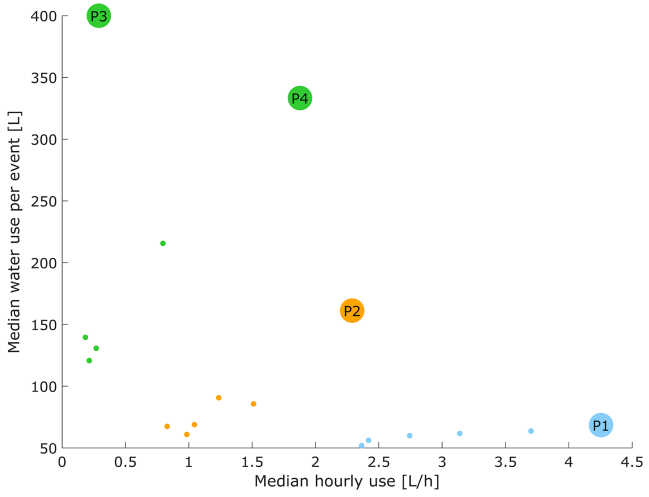

Figure 7. Scatter plot of median end use water consumption per consumption event (*y* axis) versus median hourly water use (*x* axis), for each household cluster. Each color represents a different end use: shower (green), washing machine (orange), and irrigation (light blue). Each small or big point in the scatter plot represents a household cluster. For each end use (i.e., shower, clothes washing, or irrigation), markers of profile clusters characterized by that specific end use in Figure 3 are represented with bigger size than other profile clusters, and labeled, for their primary end use, according to profile cluster labels P1–P4.

A more in-depth analysis of the statistical characteristics of these end uses suggests that different combinations of two factors contribute to increasing the magnitude of use of a specific end use, that is, (i) its frequency of use and (ii) the amount of water used per end use event. The distribution of the scatter elements in Figure 7 shows how frequency of use and amount of water used per event combine differently for different types of behavioral profile clusters. In the figure, the median hourly shower, washing machine, and irrigation water end uses and the median water use per end use event are represented, for each profile cluster. For a given level of median water used per end use event, the median hourly use increases if the specific end use occurs more frequently. Each profile cluster is further characterized in the next paragraphs, based on the information visualized in Figure 7.

- *Behavioral profile clusters characterized by irrigation end use*. These profile clusters (green markers in the scatter plot) assume high values of median water use per event but low values of median hourly water use. This can be explained by the fact that water is not used for irrigation purposes everyday; thus, it has a low frequency of use, but it involves a larger amount of water used each time, compared to other end uses. Moreover, the median water used per irrigation event assumes values in a wider range than that of other end uses, reaching up to 400 L/hr.

The likely reason for this is that the amount of water used per irrigation event depends on several factors, including landscape size, irrigation technology and efficiency, and duration. It should be noted that the median water use values for irrigation in Figure 7 are relatively low. This is because (i) the data were collected during a period of water scarcity, where water restrictions were imposed on many major end use categories, especially irrigation and (ii) the data were collected during winter, spring, and autumn (from February to December), when irrigation consumption is generally low due to relatively cool weather conditions.

- *Behavioral profile clusters characterized by shower end use.* The characteristics of the behavioral profile cluster characterized by shower end use are symmetrical: It exhibits high median hourly water use but low median water use per event. This is likely due to the higher frequency of the use of showers, which are usually taken on a daily basis, and, at the same time, the relatively low amount of water used per event. The spread in median hourly use for the behavioral profile cluster characterized by shower end use is most likely caused by different shower frequencies.
- *Behavioral profile clusters characterized by clothes washing end use.* Finally, the behavioral profile cluster characterized by clothes washing end use shows intermediate characteristics, which can be explained by the fact that a single washing machine run generally uses more water than a single shower, but less than an irrigation cycle. The spread in the values for median water per event is likely to be caused by the use of different washing programs.

Overall, these numerical results addressed Q1 and demonstrated that our end use-based customer segmentation analysis can identify heterogeneous water use behaviors and suggest time-related patterns and end use drivers. Therefore, our analysis, fed only with water use data sampled with high-frequency single-point meters, supports the discovery of water end use behaviors.

4.2. Behavioral Regularity

In the previous section, we analyzed only the first eigenbehavior of each household. This is in line with previous studies (e.g., Cominola, Spang, et al. 2018; Giuliani & Herman, 2018) and is supported by our numerical results: In our case study, we found that the first eigenbehavior could explain up to over 90% of the variance in the input data for a small fraction of households. In turn, the second eigenbehavior explained less than 20% of the variance for more than 90% of the households, and even less variance was accounted for by other secondary behaviors.

While secondary behaviors did not appear to contribute toward explaining the most frequent water use routines and the variance in the data, understanding how frequently households shifted between primary and secondary behaviors can be relevant for demand management. Behavior change actions for water conservation and demand management might intuitively produce more effective outcomes when targeted at users who exhibit stable behavior. Here, we identified more or less stable behaviors and thereby addressed Q2 by looking at the amount of variance explained by the primary eigenbehavior, as well as how frequently primary eigenbehaviors were adopted for different days of the week by the consumers of different households.

First, we considered the amount of variance explained by the first eigenbehavior for the six different profile clusters we identified with our customer segmentation analysis in Figure 3. The numerical results indicate that the first eigenbehavior explains more than 50% of the data variance, mainly for households belonging to profile Clusters P1, P2, and P3. Moreover, the households with the largest amount of data variance explained by the first eigenbehavior belong to profile clusters characterized by irrigation end use. This suggests that there is some regularity in these households' routines, which is largely captured with a single eigenbehavior. Water demand management actions designed for these households would likely foster more efficient watering habits and technologies, given that the impact caused by irrigation end use on the overall water use of these households is predominant and likely to be caused by very regular routines.

Second, a few interesting aspects related to behavioral regularity emerge from the results in Figure 8. The boxplots in the figure represent the number of occurrences of the first eigenbehavior over 25 weeks in the observation period, for different days of the week and behavioral profile clusters (households belonging to profile clusters characterized by shower end use are on top, clothes washing in the middle, and irrigation at the bottom). For each type of profile, two subclusters of households were identified by K-means and silhouette coefficient, represented in the figure in the right and left subplots. The fact that K-means and silhouette identified two subclusters for each type of profile suggests that households that are characterized by a similar eigenbehavior can differ in terms of how regularly they adopt such behavior.

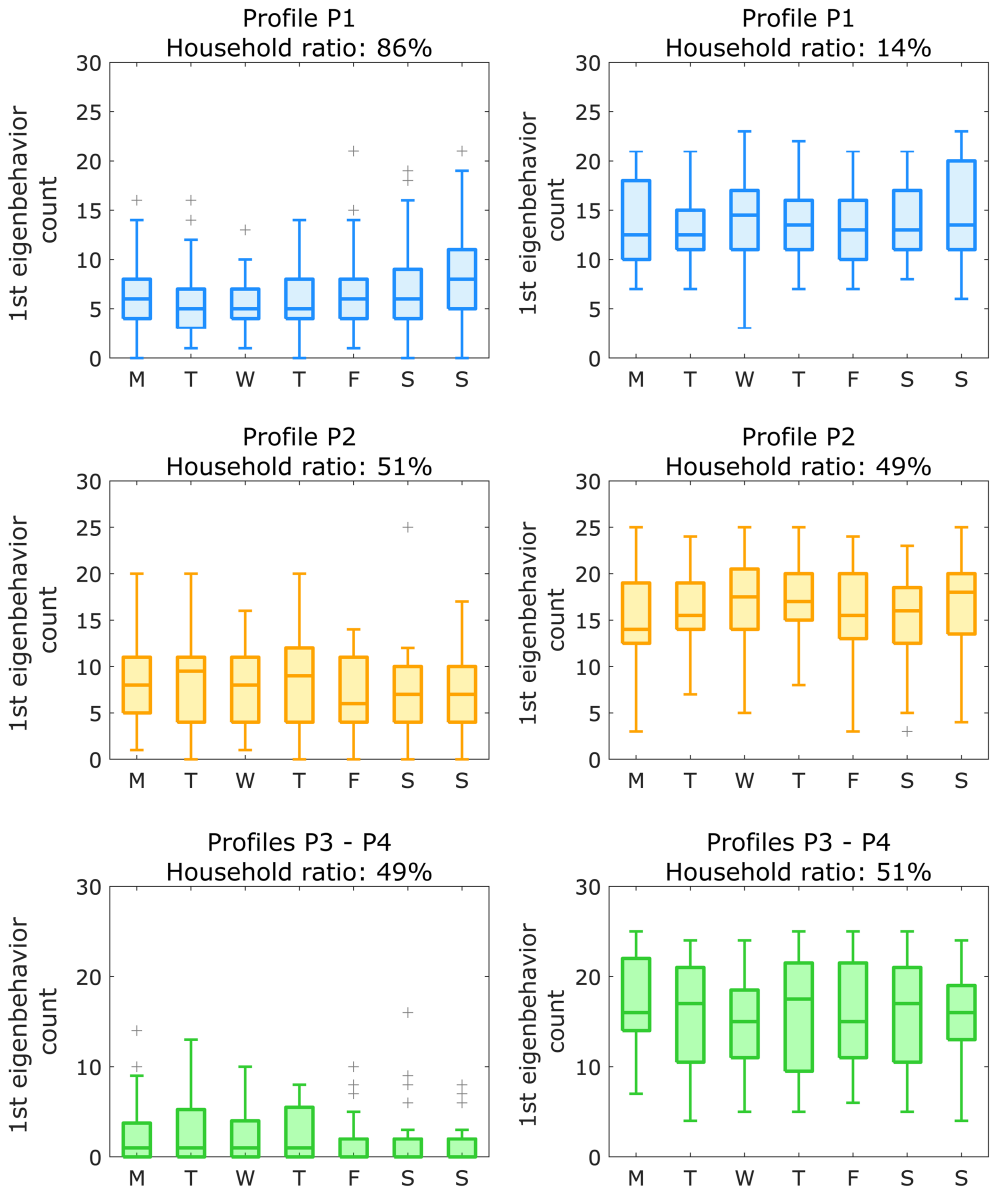

Figure 8. Boxplot of first eigenbehavior count for subclusters of households belonging to the three main profile clusters. For each type of profile cluster (figure rows), two subclusters (figure columns) were found by K-means clustering and silhouette coefficient. The two subplots in the first row refer to household profiles characterized by shower end use, the two in the second row to those characterized by clothes washing end use, and the two subplots in the third row to those characterized by irrigation end use, as the consistent use of light blue, orange, and green colors across figures suggests.

Overall, about half of the households belonging to profile clusters characterized by irrigation and clothes washing end uses show higher values of occurrences, while only 14% of the households belonging to the profile cluster characterized by shower end use have, on average, more than 10 occurrences over the 25 weeks. Some of them even reach the maximum value of 25, as shown by the bottom right subplot. This supports the hypothesis that more regular routines can be achieved for those activities that can be automatically scheduled, such as irrigation and clothes washing. For these types of profiles, no significant differences between weekdays and weekends can be identified from this analysis. Yet the oscillating median value and the different lengths of the boxes, especially for household profiles characterized by irrigation and clothes washing end uses, suggests that there is more regularity for some days of the week.

In contrast, and of particular interest for demand management actions based on behavioral feedback, is the fact that differences between weekday and weekend water use patterns emerged for the household profile

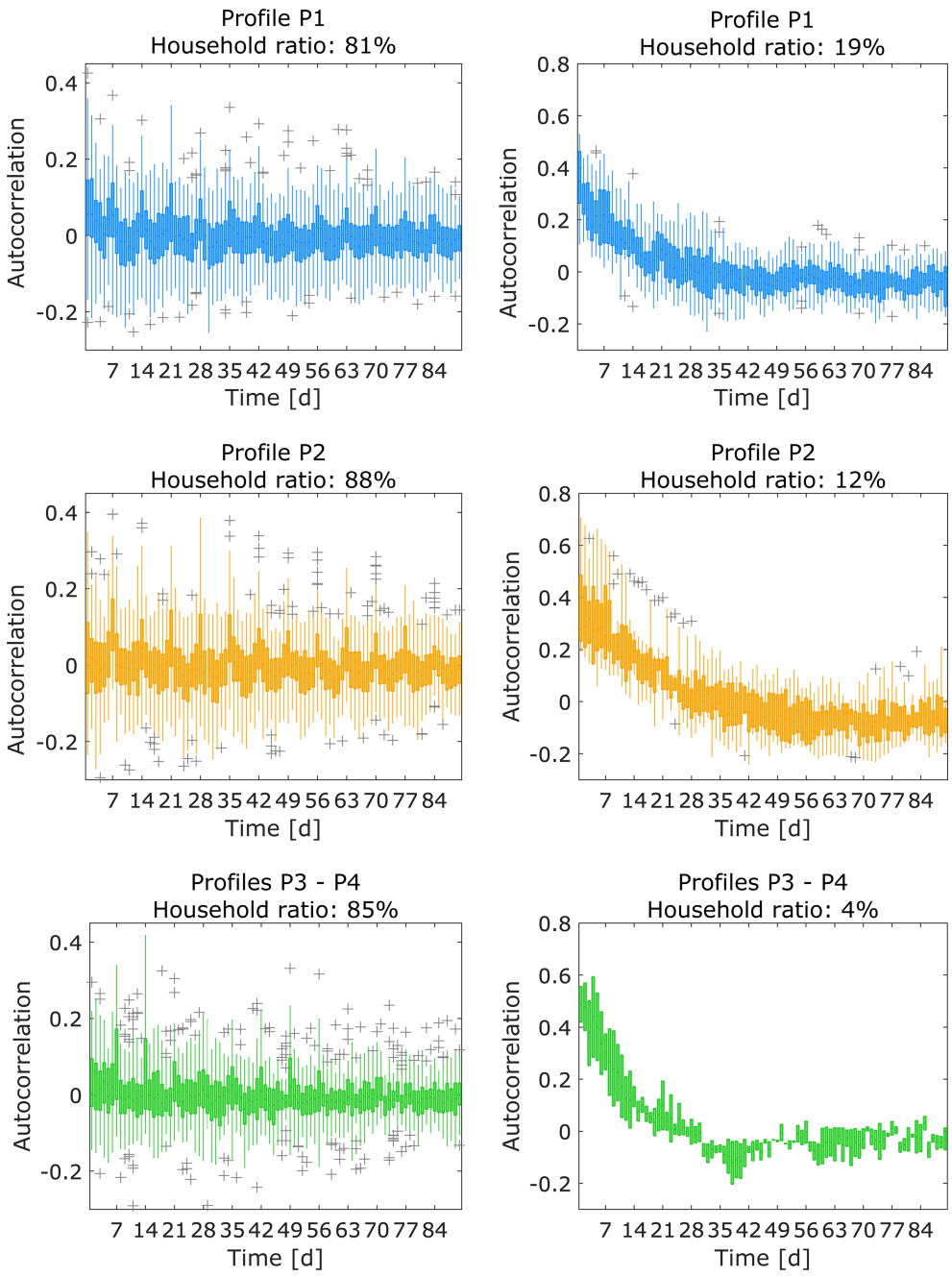

Figure 9. Boxplot of the eigenbehavior autocorrelation for subgroups of households belonging to the three main profile clusters. For each type of profile cluster (figure rows), two subclusters (figure columns) were found by K-means clustering and silhouette coefficient. The two subplots in the first row refer to household profiles characterized by shower end use, the two in the second row to those characterized by clothes washing end use, and the two in the third row to those characterized by irrigation end use.

characterized by shower end use, for which water use patterns (i.e., duration, total flow, and peak flow) largely depend on human habits and attitudes, rather than being filtered by programmable timers and usage (as it stands for irrigation and clothes washing usages). These results are contrary to findings in previous studies, because the behavioral profiles characterized by shower end use seem to be slightly more regular during weekends, rather than weekdays in this study (see top left subplot).

Apart from the above aspects, this specific case study does not reveal particularly significant patterns or differences for different days of the week.

4.3. Periodicity of Primary Water End Use Profiles

In this final section, we address Q3 by analyzing whether the main water end use profiles show any periodicity characteristics, to extend our analysis to focus on the discovery of medium- and long-term behavioral patterns. This was done by calculating the autocorrelation of the time series of the eigenbehaviors (primary or secondary) that achieved the highest score for each day of the time horizon. This analysis—*eigenbehavior autocorrelation analysis* from this point on—provided an indication of the degree of correlation of the assumed behaviors over time, as time lag increases. The most evident results are represented in Figure 9, where the autocorrelation plot for behavioral profile clusters characterized by shower, clothes washing, and irrigation end uses are represented from top to bottom. Lag times up to 90 days were included in this analysis (see *x* axis of the subplots in the figure). Considering the results obtained from the eigenbehavior analysis (Figure 3), only results from the eigenbehavior autocorrelation analysis for time interval [7 a.m. to 12 p.m.] are reported here.

Similar to the results presented in the previous section, two different subbehaviors emerge for each type of profile. For instance, looking at the two top subplots with household profiles characterized by shower end use, over 80% are characterized by an autocorrelation of the highest primary eigenbehavior with peaks of correlation repeated with a periodicity of 7. This means that weekly routines exist, and thus, medium-term behavioral patterns characterize the majority of these households. Even though the absolute value of the autocorrelation peaks is only around 0.2, such peaks remain evident over all 90 days considered as the lag time for autocorrelation evaluation. The same behavior can be observed for household profiles characterized by clothes washing end use, with similar patterns and household percentage. As can be seen, only a minority of households contribute to the subplots on the right representing "short-memory" households, which tend to have more consistent water usage habits in the range of a few days but do not show any medium- or long-term periodicity. Finally, while preserving correlation peaks at 7- and 14-day lag times, the autocorrelation periodicity of household profiles characterized by irrigation end use is less clear as longer-term seasonal patterns are more likely to characterize irrigation usage. These cannot be captured in the 10-month period of data used in this analysis (in addition, for this type of households, five households presented NaN values in the correlation analysis and are not represented in Figure 9. They constitute the missing 11%).

A comparison of the above results with those presented in Figure 8 suggests that behaviors are repeated over time for the majority of households with weekly routines. Therefore, even though primary behaviors are not always adopted, some behavioral consistency emerges in the medium term.

5. Discussion and Conclusions

In this paper, we explored which heterogeneous water use behaviors and consumer segments can be identified and characterized at the end use level only via data mining smart meter data. We contributed a data-driven, end use-based, water consumer segmentation analysis that expands the information content of water use data sampled with nonintrusive, single-point, smart meters to answer the above question and uncover heterogeneous water use behaviors. More specifically, our consumer segmentation identifies recurring water end use routines from disaggregated end use data, groups water use households in clusters according to similarities of water end use profiles, and characterizes the latter in terms of primary end use determinants, regularity, and periodicity over time. Finally, it helps formulate recommendations to design demand management and water conservation interventions.

In light of both the developed method and the numerical results, we claim that the benefit provided by our end use-based customer segmentation analysis approach to demand management is threefold. First, our end use-based customer segmentation analysis advances water demand modeling from smart meter data. Coupling end use demand modeling with segmentation analysis enables the aggregate data gathered via single-point smart meters to be disaggregated before performing the actual consumer segmentation. Thus, as long as single-point smart meters are sampled with a time resolution that is sufficiently frequent to enable a disaggregation/classification of end uses (e.g., subminute) with properly trained end use disaggregation algorithms, our customer segmentation analysis characterizes water use behaviors without requiring additional input data or a priori analysis. This implies that the intrusiveness and cost associated with survey campaigns or submetering installations are avoided (or only limited to the time of development of the end use disaggregation algorithm) and household behaviors are discovered automatically by our proposed data-driven segmentation method. Second, the numerical results presented in sections 4.1, 4.2, and 4.3

illustrative how the end use-based customer segmentation analysis we implemented is a powerful tool to identify heterogeneous water end use behaviors from time series of smart meter data, as they reveal the main drivers of such behaviors and characterize water users' behaviors with information about the regularity of their water use routines. Third, our approach performs data dimensionality reduction via PCA and clustering. Thus, it extracts a concise representation of the main water demand patterns from long time series of water use data, which can be handled easily as input to the design of demand management strategies. In addition to providing a deep understanding of the main individual and macro water use behaviors within a community of water consumers, these outcomes provide insights into user-specific characteristics that can be considered when designing customized water demand management actions. In fact, end use level routines and time of use are key to shaping and customizing water demand management programs to target specific water use patterns (e.g., for peak shifting) or end uses (e.g., to improve water use efficiency or identify and change anomalous behaviors).

Looking at the practical aspects and wider application of our customer segmentation analysis approach, we acknowledge that the availability of smart meter data and the capability to conduct this kind of analysis are fundamental requirements to obtain information about primary water use behaviors with high spatial and temporal resolution. One of the current requirements of our methodology, for instance, is the availability of subminute metered readings from households. These traces are needed in the end use disaggregation phase. Gathering data with such a high resolution, at the household level, is not currently done by utilities, primarily due to unclear benefits, high volumes of data to manage, and meter battery duration. These requirements might limit the current application of such analysis, as many utilities have not yet widely deployed smart metering infrastructure and analytic resources. Yet we believe that demonstrative applications like the one in this study contribute to showing which information can be extracted from smart meter data to gain knowledge on water demand patterns and determinants and potentially inform DSM strategies. Methods and findings from this study assist with facilitating a serious conversation on the benefits of smart metering technology and the development of business cases for its future deployment. If clear benefits emerge, utilities might consider increasing the temporal frequency of meter readings, at least as part of short-term campaigns to collect the data needed for profiling and segmentation analysis similar to the one proposed in this study, building baseline models of their water consumers, and monitoring behavior changes over time when DSM actions are implemented. Moreover, the fact that subminute resolution data are not currently collected on a routine basis does not necessarily imply that this will not be enabled in coming years by better and more cost-effective metering and data management technologies. The electricity sector and smart grids, for instance, have already been disrupted by digital technologies (Tuballa & Abundo, 2016). A similar phenomenon, powered by new metering technologies and digital networks, such as 5G, can happen in the water sector, enhancing water utilities' management decisions and their level of engagement with customers.

Overall, the contributions and limitations of our customer segmentation analysis offer opportunities for further research. The availability of longer time series of smart meter data would enable better evaluation of the generality and usability of our customer segmentation analysis approach across temporal and spatial scales. This would regard, first, assessing if any change in the main daily demand patterns occurs in the medium- and long-term in relation to seasonal effects (e.g., on the amount of water used for irrigation) or demand management interventions (e.g., water use restrictions). Second, the discovery of recurrent demand patterns that occur at temporal scales longer than 1 day would complement the periodicity analysis presented in section 4.3. Third, replicating our customer segmentation analysis with smart meter data from several households located in many heterogeneous climatic and economic areas would support comparative analysis, the identification of context-dependent behavioral differences and attitudes, and an overall assessment of the generality and scalability of our approach, along with the validity of the results obtained in this study. Further analysis with data from other contexts (e.g., Europe and the United States) would also enable better assessment of the ability of the Autoflow end use disaggregation algorithm to disaggregate smart meter data collected in contexts different from Australia, where a partial recalibration might be needed due to different water fixture characteristics. Fourth, assessing the sensitivity of our customer segmentation analysis approach to varying temporal resolution of its input data for different case studies would provide a robust evaluation of its usability in cases with reduced availability of high-resolution data from smart meters. Moreover, cross-correlating the end use demand patterns we identified with users' demographics and psychographics would enable a more detailed characterization of the profiles, as in Cominola, Spang, et al. (2018). Finally, coupling energy use information with the identified profiles can facilitate water-energy end

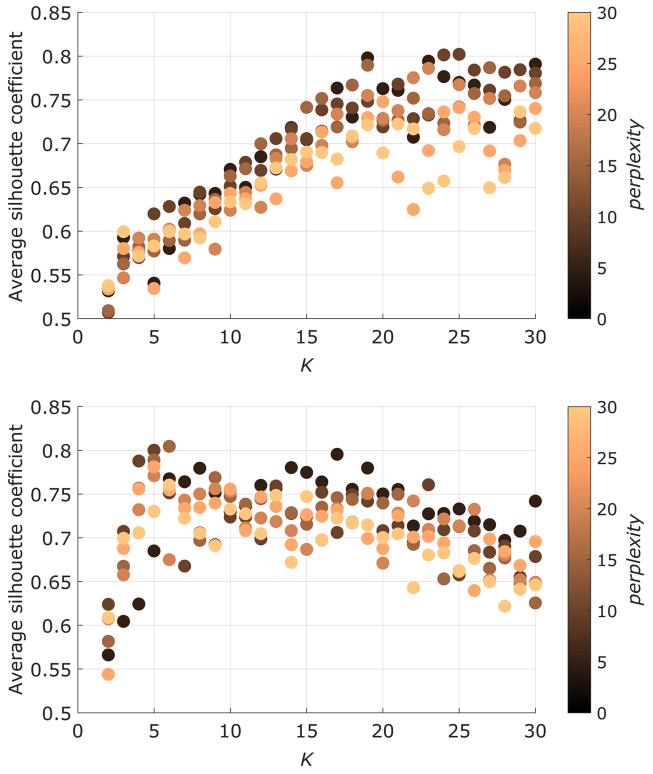

Figure A1. Average silhouette coefficient obtained for different combinations of t-SNE *perplexity* and K-means *K* parameter values. Time series of water end use aggregated both at hourly and 6-hr levels are considered as input to the household water demand profiling based on t-SNE and K-means: the subplot on top reports the results from hourly data, and the subplot on the bottom reports those from 6-hr aggregated data.

use disaggregation, enable a quantitative assessment of the water-energy nexus at the household level, and provide insights for coordinated water-energy demand management programs.

Appendix A: Silhouette Coefficient for t-SNE and K-Means Parameterization

The outcome of the *household water demand profiling* phase of the customer segmentation analysis approach presented in this paper can be sensitive to the values of the parameters of t-SNE (van der Maaten, 2014; van der Maaten & Hinton, 2008) and K-means (MacQueen, 1967). As explained in section 2, we first computed the average silhouette coefficient (Rousseeuw, 1987) to evaluate the quality of the cluster outcome obtained for different combinations of t-SNE *perplexity* parameter (which balances local and global data characteristics in the t-SNE algorithm) and K-means *K*, that is, the number of clusters in K-means. Second, we selected the best performing combination of *perplexity* and *K* to perform the *household water demand profiling*.

Figure A1 shows the average silhouette coefficient obtained for different combinations of *perplexity* and *K*. The cluster quality improves as the silhouette coefficient assumes values close to 1 and decreases as values get closer to −1. We evaluated the silhouette coefficient both for the *household water demand profiling* of end use time series aggregated both to hourly and 6-hr levels, as detailed in section 4.1. We considered values of *K* varying between 2 and 30, with Step 1, and *perplexity* values varying between 5 and 30, with Step 5.

The results visualized in Figure A1 show that silhouette values above 0.5 can be obtained already by splitting the considered data in two clusters, both for hourly and 6-hr level data. The best silhouette value obtained in both cases is approximately 0.8, suggesting overall a good cluster quality. Silhouette values for hourly data (top subplot) are sensitive to *K* and become relatively stable approximately for $K \geq 19$. They are less sensitive to changes in the *perplexity* value, even though, for a given *K*, slightly better silhouette values are obtained for *perplexity* values lower than 15. Similarly, the silhouette coefficient over data aggregated at 6-hr level (bottom plot) varies more in relation to changes in *K*, rather than changes in the *perplexity*. The main difference that emerges by comparing the results from the hourly and 6-hr level data is that a lower number

of clusters is needed as the data aggregation interval increases, as detailed in section 4.1. In fact, the best silhouette s is obtained for $K = 6$ ($s = 0.80$) for 6-hr level data, while the best value of s is obtained for $K = 25$ ($s = 0.80$) when hourly data are considered. In contrast, the *perplexity* values corresponding to the best silhouette values are very similar in the two cases: 15 for 6-hr level data and 10 for hourly data.

Acknowledgments
The SEQ end use data set was collected as part of the South East Queensland Urban Water Security Research Alliance. Particular thanks go to the Systematic Social Analysis Team, Allconnex Water, Urban Water Utilities, UnityWater, SEQREUS research team located at the Smart Water Research Centre. The original data are maintained and can be requested at this data repository (https://researchdata.ands.org.au/south-east-queensland-2010-read/9008; https://doi.org/10.4225/01/4F8CFA0465337. The Matlab code developed for the segmentation analysis and the water end use time series, aggregated at 1-hr time sampling resolution for sample residential accounts considered in this study, are available in this GitHub repository (https://github.com/acominola/Water-End-Use-Customer-Segmentation.git). Autoflow is an autonomous water end use disaggregation tool which was developed under an industry collaboration project (with three Melbourne water utilities). The release of the Autoflow software and its source code (either for download or purchase) is restricted under the agreement with water utilities.

References

Abed-Meraim, K., Qiu, W., & Hua, Y. (1997). Blind system identification. *Proceedings of the IEEE*, *85*(8), 1310–1322.

Beal, C. D., & Flynn, J. (2015). Toward the digital water age: Survey and case studies of australian water utility smart-metering programs. *Utilities Policy*, *32*, 29–37.

Beal, C. D., Gurung, T. R., & Stewart, R. A. (2016). Demand-side management for supply side efficiency: Modeling tailored strategies for reducing peak residential water demand. *Sustainable Production and Consumption*, *6*, 1–11.

Beal, C., & Stewart, R. (2014). Identifying residential water end uses underpinning peak day and peak hour demand. *Journal of Water Resources Planning and Management*, *140*(7).

Beal, C., Stewart, R. A., Huang, T., & Rey, E. (2011). South East Queensland residential end use study. *Journal of the Australian Water Association*, *38*(1), 80–84.

Blokker, E., Vreeburg, J., & van Dijk, J. (2010). Simulating residential water demand with a stochastic end-use model. *Journal of Water Resources Planning and Management*, *136*(1), 19–26.

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

Cardell-Oliver, R. (2013a). Discovering water use activities for smart metering. In *2013 IEEE Eighth International Conference on Intelligent Sensors, Sensor Networks and Information Processing*, IEEE, pp. 171–176.

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

Cardell-Oliver, R., & Povey, T. (2018). Profiling urban activity hubs using transit smart card data. In *Proceedings of the 5th Conference on Systems for Built Environments*, ACM, pp. 116–125.

Cardell-Oliver, R., Wang, J., & Gigney, H. (2016). Smart meter analytics to pinpoint opportunities for reducing household water use. *Journal of Water Resources Planning and Management*, *142*(6), 4016007.

Cheong, S.-M., Choi, G.-W., & Lee, H.-S. (2016). Barriers and solutions to smart water grid development. *Environmental Management*, *57*(3), 509–515.

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

Cominola, A, Giuliani, M, Piga, D, Castelletti, A, & Rizzoli, A. E. (2015). Benefits and challenges of using smart meters for advancing residential water demand modeling and management: A review. *Environmental Modelling & Software*, *72*, 198–214.

Cominola, A., Spang, E. S., Giuliani, M., Castelletti, A., Lund, J. R., & Loge, F. J. (2018). Segmentation analysis of residential water-electricity demand for customized demand-side management programs. *Journal of Cleaner Production*, *172*, 1607–1619.

Creaco, E., Kossieris, P., Vamvakeridou-Lyroudia, L., Makropoulos, C., Kapelan, Z., & Savic, D. (2016). Parameterizing residential water demand pulse models through smart meter readings. *Environmental Modelling & Software*, *80*, 33–40.

Duerr, I., Merrill, H. R., Wang, C., Bai, R., Boyer, M., Dukes, M. D., & Bliznyuk, N. (2018). Forecasting urban household water demand with statistical and machine learning methods using large space-time data: A comparative study. *Environmental Modelling & Software*, *102*, 29–38.

Eagle, N., & Pentland, A. S. (2009). Eigenbehaviors: Identifying structure in routine. *Behavioral Ecology and Sociobiology*, *63*(7), 1057–1066.

Escriva-Bou, A., Lund, J. R., & Pulido-Velazquez, M. (2015). Modeling residential water and related energy, carbon footprint and costs in California. *Environmental Science & Policy*, *50*, 270–281.

Espinoza, M., Joye, C., Belmans, R., & De Moor, B. (2005). Short-term load forecasting, profile identification, and customer segmentation: A methodology based on periodic time series. *IEEE Transactions on Power Systems*, *20*(3), 1622–1630.

Giuliani, M., & Herman, J. D. (2018). Modeling the behavior of water reservoir operators via eigenbehavior analysis. *Advances in Water Resources*, *122*, 228–237.

Gonzales, P., & Ajami, N. (2017). Social and structural patterns of drought-related water conservation and rebound. *Water Resources Research*, *53*, 10,619–10,634. https://doi.org/10.1002/2017WR021852

Horsburgh, J. S., Leonardo, M. E., Abdallah, A. M., & Rosenberg, D. E. (2017). Measuring water use, conservation, and differences by gender using an inexpensive, high frequency metering system. *Environmental Modelling & Software*, *96*, 83–94.

Jolliffe, I. (2002). *Principal component analysis* (2nd ed.). New York: Springer.

Kofinas, D. T., Spyropoulou, A., & Laspidou, C. S. (2018). A methodology for synthetic household water consumption data generation. *Environmental Modelling & Software*, *100*, 48–66.

Kwac, J., Flora, J., & Rajagopal, R. (2014). Household energy consumption segmentation using hourly data. *IEEE Transactions on Smart Grid*, *5*(1), 420–430.

Laspidou, C., Papageorgiou, E., Kokkinos, K., Sahu, S., Gupta, A., & Tassiulas, L. (2015). Exploring patterns in water consumption by clustering. *Procedia Engineering*, *119*, 1439–1446.

Luciani, C., Casellato, F., Alvisi, S., & Franchini, M. (2019). Green smart technology for water (GST4Water): Water loss identification at user level by using smart metering systems. *Water*, *11*(3), 405.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations, *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1, pp. 281–297). Oakland, CA, USA.

Mayer, P. W., DeOreo, W. B.,Optiz, E. M.,Kiefer, J. C.,Davis, W. Y.,Dziegielewski, B., & Nelson, J. O. (1999). *Residential end uses of water*. Denver, CO: American Water Works Association.

Nambi, S., Pournaras, E., & Prasad, R. V. (2016). Temporal self-regulation of energy demand. *IEEE Transactions on Industrial Informatics*, *12*(3), 1196–1205.

Nguyen, K. A., Stewart, R. A., Zhang, H., & Jones, C. (2015). Intelligent autonomous system for residential water end use classification: Autoflow. *Applied Soft Computing*, *31*, 118–131.

Novak, J., Melenhorst, M., Micheel, I., Pasini, C., Fraternali, P., & Rizzoli, A. E. (2018). Integrating behavioural change and gamified incentive modelling for stimulating water saving. *Environmental Modelling & Software*, *102*, 120–137.

Patabendige, S., Cardell-Oliver, R., Wang, R., & Liu, W. (2018). Detection and interpretation of anomalous water use for non-residential customers. *Environmental Modelling & Software*, *100*, 291–301.
Poussevin, M., Tonnelier, E., Baskiotis, N., Guigue, V., & Gallinari, P. (2016). Mining ticketing logs for usage characterization with non-negative matrix factorization. In M. Atzmueller et al. (Eds.), *Big Data Analytics in the Social and Ubiquitous Context. SENSEML 2015, MUSE 2014, MSM 2014. Lecture Notes in Computer Science*. Cham: Springer.
Quesnel, K. J., & Ajami, N. K. (2018). Large landscape urban irrigation: A data-driven approach to evaluating conservation behavior. *Water Resources Research*, *55*, 771–786. https://doi.org/10.1029/2018WR023549
Rougé, C., Harou, J. J., Pulido-Velazquez, M., Matrosov, E. S., Garrone, P., Marzano, R., et al. (2018). Assessment of smart-meter-enabled dynamic pricing at utility and river basin scale. *Journal of Water Resources Planning and Management*, *144*(5), 4018019.
Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, *20*,53–65.
Russell, S., & Fielding, K. (2010). Water demand management research: A psychological perspective. *Water Resources Research*, *46*, W05302. https://doi.org/10.1029/2009WR008408
Sensus (2012). *Water 20/20: Bringing smart water networks into focus*. London: Sensus.
Sønderlund, A. L., Smith, J. R., Hutton, C. J., Kapelan, Z., & Savic, D. (2016). Effectiveness of smart meter-based consumption feedback in curbing household water use: Knownsand unknowns. *Journal of Water Resources Planning and Management*, *142*(12), 4016060.
Stewart, R., Giurco, D., & Beal, C. (2013). Age of intelligent metering and bigdata: Hydro informatics challenges and opportunities. *JournaloftheInternational Association for Hydro-environment Engineering and Research*, *2*, 107–110.
Stewart, R. A., Nguyen, K., Beal, C., Zhang, H., Sahin, O., Bertone, E., et al. (2018). Integrated intelligent water-energy metering systems and informatics: Visioning a digital multi-utility service provider. *Environmental Modelling & Software*, *105*, 94–117.
Stewart, R. A., Willis, R., Giurco, D., Panuwatwanich, K., & Capati, G. (2010). Web-based knowledge management system: Linking smart metering to the future of urban water planning. *Australian Planner*, *47*(2), 66–74.
Tsakalides, P., Panousopoulou, A., Tsagkatakis, G., & Montestruque, L. (Eds.) (2018). *Smart water grids*. Boca Raton, FL: CRC Press. https://doi.org/10.1201/b21948
Tuballa, M. L., & Abundo, M. L. (2016). Are view of the development of smart grid technologies. *Renewable and Sustainable Energy Reviews*, *59*, 710–725.
Turner, A., & White, S. (2017). Urban water futures: Trends and potential disruptions.
van derMaaten, L. (2014). Accelerating t-SNE using tree-based algorithms. *Journal of Machine Learning Research*, *15*(1), 3221–3245.
van derMaaten, L., & Hinton, G. (2008). Visualizing data usingt-SNE. *Journal of Machine Learning Research*, *9*(Nov), 2579–2605.
Vitter, J. S., & Webber, M. (2018). A non-intrusive approach for classifying residential water events using coincident electricity data. *Environmental Modelling & Software*, *100*, 302–313.
Willis, R. M., Stewart, R. A., Giurco, D. P., Telebpour, M. R., & Mousavinejad, A. (2013). End use water consumption in households: Impact of socio-demographic factors and efficient devices. *Journal of Cleaner Production*, *60*, 107–115.
Yuan, N. J., Zheng, Y., Xie, X., Wang, Y., Zheng, K., & Xiong, H. (2014). Discovering urban functional zones using latent activity trajectories. *IEEE Transactions on Knowledge and Data Engineering*, *27*(3), 712–725.

Erratum

In the originally published version of this article, Figures 1 and 2 as published were older versions. The figures have since been updated and this version may be considered the authoritative version of record.