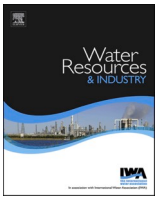
Water Resources and Industry 36 (2026) 100380



Contents lists available at ScienceDirect

Water Resources and Industry

journal homepage: www.elsevier.com/locate/wri



Non-residential water-consumption characterization and clustering: evidences from service-industry users in Northern Italy

Jacopo Lodi *, Vittorio Micai, Valentina Marsili, Filippo Mazzoni, Stefano Alvisi

Department of Engineering, University of Ferrara, Ferrara, Italy

ARTICLE INFO

Keywords:
Smart metering
Non-residential activities
Daily consumption
Daily profiles
Unsupervised machine learning

ABSTRACT

Non-residential water consumption can represent a significant share of urban water demand, making the characterization of users belonging to the service industry (SI) pivotal for urban water management. However, the relationship between water consumption and SI user categories remains underexplored, limiting accurate modeling and the reliability of leakage assessments. This study aims to fill this gap by analysing hourly water-consumption data collected through smart meters over a one-year period for 44 labelled SI users located in Northern Italy and grouped into six categories: shops, catering services, offices, wellness services, recreational and training activities and police stations. The methodological framework relies on four phases: (*i*) data preprocessing, including the removal of periods affected by long-term closure or post-meter leakages; (*ii*) characterization of daily water consumption for each user; (*iii*) derivation of seasonal and annual average daily consumption profiles; and (*iv*) profile clustering based on the application of K-means and Gaussian Mixture Models (GMMs) and clustering performance evaluation by adopting the Adjusted Rand Index (ARI) and the Normalized Mutual Information (NMI). Results reveal substantial variability in both daily consumption and profiles, even among users belonging to the same SI category. Moreover, user categories do not closely correspond to the obtained clusters (ARI $\approx 0.10 \div 0.20$ and NMI $\approx 0.30 \div 0.40$), suggesting that user category alone may not be sufficient to describe actual water-consumption behaviours. The dataset is provided as an open-access library, supporting applications such as preliminary water-demand estimations, water-balance assessments, or modelling applications in contexts where detailed consumption data are unavailable.

1. Introduction

Population growth, rapid urbanization, and climate change effects are increasingly raising concerns about water availability worldwide [1,2]. Consequently, effective planning and management of drinking water supply systems are essential to mitigate stress conditions exacerbated by the demographic and environmental drivers [3]. In this context, strategic decision-making should be based on a detailed knowledge of where, when, and how water is used [4], making the spatiotemporal characterization of users' water consumption a research priority for decades [5].

Traditionally, water consumption has relied on manual readings of mechanical meters, collected monthly or annually for billing purposes. Despite low temporal resolution, these records supported early scientific investigations of water consumption [6,7]. Since

* Corresponding author.
E-mail address: jacopo.lodi@unife.it (J. Lodi).

https://doi.org/10.1016/j.wri.2026.100380
Received 16 March 2026; Received in revised form 29 May 2026; Accepted 21 June 2026
Available online 22 June 2026
2212-3717/© 2026 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

the late 1990s, technological advancements and the introduction of smart meters have enabled water-consumption monitoring at higher temporal resolutions (i.e. ranging from daily to 1-sec time step) and with the possibility of automatically collecting and transmitting water-consumption data [8]. The installation of these tools is currently increasing worldwide [9], thereby contributing to the ongoing transition toward the digitization of water distribution systems [10]. However, from an operational standpoint, issues of different nature still hinder their scalability to large-scale continuous operative smart meter installations [8]. These include aspects related to investment costs, privacy regulations, and operational factors such as battery life or data transmission issues [11]. Despite these practical challenges, the availability of high-resolution data offers significant potential applications. By way of example, it enables the analysis of water-consumption patterns at the single-user scale with the possibility to disaggregate water consumption into individual end uses [12], or identifying extreme values, i.e. peak values [13,14] or the Minimum Night Consumption, a crucial parameter for the estimation of water losses through the Minimum Night Flow method [15,16].

In light of the benefits related to the use of smart meters, extensive research has focused primarily on residential water consumption (e.g., Refs. [17–20]). However, although residential users generally represent the majority of users in a water distribution network (WDN), a significant fraction of urban water consumption is attributable to non-residential (NR) activities [21].

Given the potential impact of the non-residential sector on WDN operation in urban contexts, characterizing the related water consumption has recently received increasing attention in scientific research [22,23]. While – similar to the residential sector – early investigations on non-residential water consumption focused on characterizing water consumption on an annual or, at most, daily basis using billing data (e.g., Refs. [24–27]), the advent of smart meters has enabled sub-daily monitoring, fostering recent studies on the topic.

Notably, attention has been paid primarily on large non-residential activities belonging to the industrial sector, whose water consumption is typically monitored by water utilities using flow meters [28]. Indeed, the volumes withdrawn by a large industrial user can affect a WDN comparably to a small hydraulic district, justifying water utility's interest in performing high-resolution monitoring of these highly demanding users.

However, the research gap is still pronounced for *small* non-residential users belonging to the service industry (hereinafter denoted as *SI users*), consuming less than 10,000 m^3/year [29] and spanning a wide range of activities, from commercial businesses to catering services, office/administrative activities, wellness services, training activities, and educational and recreational programs.

Existing literature on smart metering of SI users can be grouped into two categories. The first category includes studies focusing on very specific categories, such as schools and university campuses [30,31], hotels and tourism-related businesses [32–34], assisted-care facilities [23,32], office buildings [32,35], and manufacturing facilities [23]. Small groups of guest rooms and restaurants were also considered for modelling purposes and with a specific focus on hot water consumption [36,37]. Although relying on very high-resolution data (i.e. 5-min or finer), the aforementioned studies typically consider a limited sample size and cover relatively short periods (ranging from a few days to some weeks or, at best, some months). Such limitations hinder both the definition of representative consumption profiles per SI user category and the characterization of the related seasonal variations in water consumption.

A second category of studies on water consumption in the SI focuses on extensive datasets involving a large number of SI users, with the aim of clustering water-consumption patterns. Notable is the study by McKenna et al. [38] who proposed an approach combining Gaussian Mixture Models with K-means clustering applied to 85 residential and commercial users. More recently, Cheifetz et al. [39] developed a data-driven methodology to cluster water-consumption profiles observed from over 10,000 residential and non-residential users. By extracting seasonal components of water consumption and applying two clustering techniques, they identified two non-residential profiles, tied to commercial and office/industrial water consumption, respectively. However, the lack of user labelling in both studies, either separating residential from commercial users or identifying distinct non-residential categories, prevented validation of the methodologies. As a result, a significant gap persists for research that integrates smart meter water-consumption data from a statistically significant number of users of known category over a multi-seasonal period.

To advance the understanding of SI water use, this study analyses a one-year dataset of (hourly) smart meter readings from 44 labelled SI users in Northern Italy, capturing both weekly and seasonal variations of water-consumption profiles. This also allows user-level and category-level characterization. Specifically, the proposed methodology is structured in four phases. *Phase I* involves data pre-processing, including the identification of periods affected by gaps, post-meter leakages or prolonged business inactivity. In *Phase II,* daily consumption is assessed and analysed throughout the week to identify days of activity, which are considered in *Phase III* to derive water-consumption magnitude, seasonal daily profiles, and average daily profiles for individual SI users and different user categories. Lastly, *Phase IV* applies unsupervised Machine Learning (ML) to cluster dimensionless average daily profiles. Overall, the outcomes of this work constitute an open-access dataset of non-residential water consumption representing SI users from different categories.

This paper is structured as follows. Section 2 (*Materials and Methods*) details the methodology adopted for profile analysis and classification and introduces the monitored user sample and the associated water-consumption datasets. Section 3 (*Results and Discussion*) presents and discusses the results obtained from applying the methodology to the field data. Section 4 (*Study Limitations and Future Developments*) underlines the main limitations of the study and discusses future research perspectives. Lastly, in Section 5 (*Conclusions*) the main findings are summarized.

2. Materials and methods

This section presents the methodological framework developed to analyse water consumption in the SI, together with the specific datasets analysed referring to a significant number of SI users of known category across different urban areas in Northen Italy.

2.1. Methodology

The methodology proposed in the study, the layout of which is represented in Fig. 1, is applied to water-consumption time series collected with sub-hourly or hourly step at the level of the individual SI user. Overall, the method includes four phases: data pre-processing (*Phase I*), analysis of daily consumption (*Phase II*), investigation of daily profiles of water consumption (*Phase III*), and profile clustering (*Phase IV*). Details on each phase are provided in the following sub-sections.

2.1.1. Phase I: water-consumption data pre-processing

In *Phase I*, regularly sampled (sub-hourly or hourly) water-consumption time series for each SI user are obtained and pre-processed. In particular, after being derived from cumulative volumes, water-consumption data are subdivided into weekly periods (i.e. from Monday to Sunday) to facilitate pre-processing. Overall, data pre-processing includes a gap analysis aimed at detecting and managing missing values and a subsequent analysis of users' water-consumption patterns aimed at detecting post-meter leaks and/or extended periods of inactivity due to long-term (i.e., seasonal) closures. Based on this three-filter stage, individual weeks related to water-consumption data affected by not-solvable anomalies are excluded to ensure the integrity of the analysis.

Let us consider that for each SI user u (with $u = 1, \ldots, U$ and U being the total number of SI users), cumulative-volume data $v_{cum,u}(t_i)$ (with $i = 1, \ldots, I$ and I the total number of meter readings), are available over a data-collection period of, at least, one year. The water-consumption time series, q_u, is derived from the difference between consecutive cumulative-volume readings (e.g. $v_{cum,u}(t_{i+1})$ and $v_{cum,u}(t_i)$) as follows (Equation (1)):

$$q_{u,i} = \frac{v_{cum,u}(t_{i+1}) - v_{cum,u}(t_i)}{\Delta t} \tag{1}$$

Δt being the time step related to the regular sampling (e.g., hourly).

For each SI user u, the resulting water-consumption time series is subdivided into w weekly time series (i.e. 7-day periods from Monday to Sunday, with $w = 1, 2, \ldots W$ and W the total number of full weeks making up the 1-year period) and subject to gap analysis. Only weekly time series in which at least 90% of data is available (e.g., 152 out of 168 values, in case of hourly data) are retained. In addition, gaps lasting no longer than 1 h are closed using linear interpolation, whereas weekly time series characterized by gaps exceeding this threshold are excluded from the analysis.

Following gap detection and filling, water-consumption time series are analysed with the aim of identifying anomalous patterns in water consumption, such as prolonged period characterized by non-zero consumption (i.e. revealing potential post-meter leakages) or prolonged periods without water consumption (i.e. potentially related to business inactivity due to seasonal closure). In relation to the first type of anomaly, given the characteristics of SI users, it is assumed that consumption should approach zero at least once a week, most likely during nighttime or specific closing days. From an operational standpoint, for each SI user, the identification of the periods with potential post-meter leakages is performed by means of a sliding-window-based approach. Specifically, a sliding window of weekly length is applied to identify the periods with continuous water consumption, i.e. all periods during which the water-consumption time series is always greater than the accuracy of smart meter (e.g., 1 L/h). Therefore, all weeks showing the occurrence of a potential post-meter leakage are excluded from the analysis.

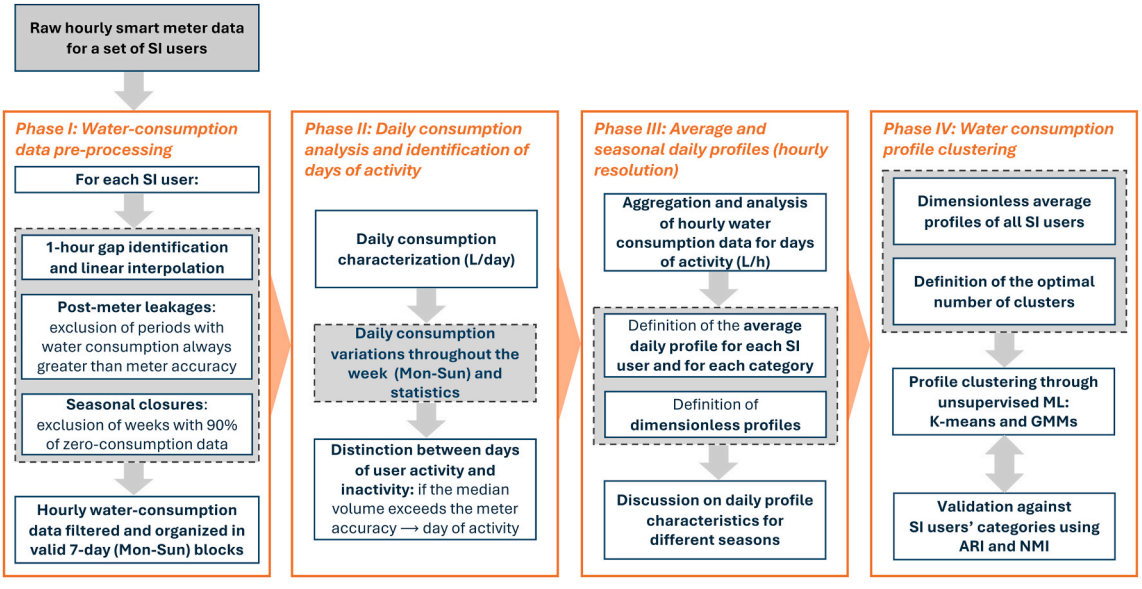


Fig. 1. Layout of the multi-phase methodology designed for non-residential water-consumption characterization in the SI and profile clustering.

The second type of anomaly, consisting of prolonged periods without water consumption, includes periods of user's inactivity, such as summer breaks in mid-August or winter holidays around December. In this regard, the methodology distinguishes between *seasonal closures*, corresponding to long-term periods of inactivity (i.e., several days, like summer or winter breaks), and *days of inactivity*, which represent routine days of very low or nil water consumption due to business weekly rest, a key-aspect for the subsequent phase of the methodology. Specifically, to remove all periods of potential long-term closure from the analysis only weeks containing at least 10% of non-zero values are maintained.

Overall, the filtering process aimed at discarding weekly periods affected by data anomalies results in specific sets of $\widetilde{W}_u$ weeks (with $\widetilde{W}_u \leq W$) with no potential post-meter leakages or periods of long-term closure which varies for each individual SI user u. These weeks will be hereinafter denoted to as *valid weeks*.

2.1.2. Phase II: daily consumption analysis and identification of days of activity

The second phase of the methodology focuses on the analysis of daily consumption (L/day) of each SI user. It is worth noting that, while transitioning to daily resolution might appear a simplification of the information provided by the dataset due to decreased resolution, this step is fundamental for characterizing user behaviour according to the specific day of the week, based on which consumption profiles are analysed.

In greater detail, by evaluating the daily consumption of each SI user, *Phase II* allows for the reconstruction of the working schedule for every SI user, discriminating between *days of user activity* and *days of user inactivity*. This step is needed due to the fundamental differences between residential and SI water-consumption characteristics. On the one hand, for residential users, daily consumption trends are typically defined by the distinction between weekdays and weekends which remains overall fixed for all users [40]. On the other hand, the analysis of SI users requires a different criterion for day type classification. First, since a business's specific day-off can fall on any day of the week, identifying a SI user's functional schedule becomes critical to ensure accurate water-consumption characterization. Second, non-residential behaviour in terms of water consumption can shift between days of full operational activity (e.g., when a shop is open for business and, consequently, consumes water), days of total inactivity (e.g., when a shop is unoccupied and water consumption is nil), but also in-between scenarios characterized by intermediate states (e.g., days when, although closed for customers, a SI user may still consume water due to routine maintenance, cleaning services or related activities). In this context, days of user activity remain of high interest as they represent integral component of the SI user's overall water consumption.

Operationally, to identify days of user activity, the daily consumption (L/day) related to the SI user u over the d-th day of the week ($d = 1,\ldots,D$, with D the total number of weekdays, i.e. 7) making up the valid week $\widetilde{w} = 1, \ldots, \widetilde{W}_u$ is first evaluated as follows (Equation (2)):

$$V_u^{d(\widetilde{w})} = \sum_{j=1}^{J} q_{uj}^{d(\widetilde{w})} \tag{2}$$

with j the j-th water-consumption value of the d-th day of week $\widetilde{w}$ and J the total number of water-consumption data related to the same day (e.g., 24 in case of hourly resolution). Once the daily consumption is assessed for all valid weeks $\widetilde{W}_u$, this information is grouped by the day of the week, from Monday to Sunday. To evaluate the operational status of each day, the median of the daily consumption distribution for each specific weekday is compared against smart-meter accuracy (e.g., 1 L/h). If the median exceeds this threshold, the weekday is classified as day of activity for that SI user; otherwise, it is designated as a day of inactivity. As a result of this process, each SI user is associated with a specific set of $\widetilde{D}_u$ days of activity (with $\widetilde{D}_u \leq D$), reflecting its individual operational schedule.

After the identification of the weekly days of activity for each user, the daily consumption delivered over each of these days is considered and subjected to statistical evaluation.

2.1.3. Phase III: average and seasonal daily profile (hourly resolution)

Phase III involves the analysis of hourly water-consumption data of each SI user, in relation to the respective days of activity. The final objective of this stage is the obtainment of a representative daily profile for each SI user across the four seasons of the year and in relation to the overall yearly horizon. This allows the discussion on how SI consumption characteristics (i.e. entity, daily trend) change in relation to different periods of the year, along with the related average patterns.

Water-consumption data of each user u are first aggregated on an hourly basis (if not originally collected at this resolution) according to Equation (3):

$$q_{u,h}^{\tilde{d}(\widetilde{w})} = CF \sum_{j=1}^{J_h} q_{u,j}^{\tilde{d}(\widetilde{w})} \tag{3}$$

where: j denotes the j-th water-consumption value associated to the h-th hour of active day $\widetilde{d}$ in valid week $\widetilde{w}$, with $\widetilde{w} = 1,\ldots,\widetilde{W}_u$, and $\widetilde{W}_u$ representing the total number of valid weeks for user u; J_h denotes the number of water-consumption data collected during the same hour; CF denotes the conversion factor between water-consumption data prior to aggregation (e.g., L/s or L/15 min) and water-consumption data aggregated on an hourly basis (i.e., L/h).

Subsequently, for each SI user u, the average daily water-consumption profile is derived for each season. All valid weeks are categorized into $S = 4$ blocks based on season, being $s = \{winter, spring, summer, autumn\}$. This subdivision may result in a different

number of valid weeks per season depending on data availability and the application of the pre-processing stage. For each seasonal block of valid weeks $\widetilde{W_u}^s$, the related days of activity are considered to assess the seasonal average profile of water consumption based on Equation (4):

$$\overline{q_{u,h}}^s = \frac{1}{\widetilde{W_u}^s} \frac{1}{\widetilde{D}_u} \sum_{\tilde{w}=1}^{\widetilde{W_u}^s} \sum_{\tilde{d}=1}^{\widetilde{D}_u} q_{u,h}^{\tilde{d}(\tilde{w})} \tag{4}$$

Once the average seasonal daily profiles, $\overline{q_{u,h}}^s$, are obtained for each user, they are combined to obtain an average profile associated with the overall yearly period, thus expressing the average water consumption expected during days of activity. The average profiles for each user, subsequently used in *Phase IV* for clustering, are computed as defined in Equation (5). To prevent the resulting profiles from reflecting unbalanced samples due to the availability of different numbers of valid weeks across seasons, seasonal averages are first calculated and then combined to obtain the overall average profile. This approach ensures that each season contributes equally to the representative profile of each SI user.

$$\overline{q_{u,h}} = \frac{1}{S} \sum_{s=1}^{S} \overline{q_{u,h}}^s \tag{5}$$

Lastly, the normalization of the overall average profile was carried out with the aim of ensuring a direct comparison of hourly fluctuations of water consumption across the SI users independently of the magnitude. This step includes the obtainment of 24-hourly consumption coefficients for each user, derived as shown in Equation (6):

$$C_{u,h} = \frac{\overline{q_{u,h}}}{\frac{1}{24} \sum_{h=1}^{24} \overline{q_{u,h}}} \tag{6}$$

2.1.4. Phase IV: water-consumption profile clustering

The final phase of the methodology focuses on identifying recurrent patterns in water consumption by clustering the dimensionless average daily profiles of all SI users. To this end, two clustering methods are considered: K-means and Gaussian Mixture Models (GMMs). These algorithms are selected to compare the results of a hard clustering approach (i.e., K-means), where each user is associated with a single cluster, with a soft clustering approach (i.e., GMMs), which relies on a probabilistic cluster allocation.

- The K-means algorithm [41] is employed to perform a hard partitioning of the dimensionless profiles into K distinct classes treated like objects within a 24-dimensional space where each dimension represents an hourly step. This algorithm is selected due to its computational efficiency, simplicity of implementation and direct interpretability [42]. The determination of the optimal number of clusters is guided by a preliminary analysis of the silhouette curve [43]. In greater detail, the squared Euclidean distance is adopted as the metric to evaluate proximity between individual objects and cluster centroids. The K-value is selected by identifying the highest average silhouette parameter, ensuring a partitioning that maximizes both internal cluster cohesion and external separation. This process yields a set of cluster centroids, each representing the fingerprint of its respective class.
- The Gaussian Mixture Models (GMMs) [44] assume that the data are generated from a mixture of a finite number of Gaussian distributions, with each distribution representing a specific cluster. By utilizing the Expectation-Maximization (EM) algorithm, the model parameters are iteratively optimized. This procedure computes the posterior probability of each user profile belonging to each cluster, corresponding to a soft assignment.

As a final step, the clusters obtained from the application of K-Means and GMMs are validated against the known SI user categories. This comparison is performed to verify that the identified classes reflect the operational realities of the SI categories, linking distinct water consumption profiles to specific user's categories. To evaluate the performance of the clustering-based classification system with respect to SI daily profiles, a $K \times C$ contingency matrix M is computed for each clustering approach, where K denotes the number of optimal clusters identified through the preliminary analysis, and C represents the number of SI user categories included in the user sample considered. Therefore, each matrix entry M_{ij} represents the number of daily profiles belonging to j-th SI user category that are assigned to the i-th cluster.

Based on the contingency matrix obtained for each clustering method, the correspondence between the identified clusters and user categories is evaluated using the Adjusted Rand Index (ARI) [45] and the Normalized Mutual Information (NMI) [46], which are invariant to label permutation.

On the one hand, ARI measures the similarity between the predicted clusters and the real categories by evaluating pairs of data points, mathematically adjusted for the chance of random grouping. This index generally ranges from 0 to 1, where a value of 1 means that clusters match perfectly with real categories and 0 corresponds to random clustering.

On the other hand, NMI represents an information-theoretic metric that quantifies the amount of shared information between the true and predicted partitions, normalized by the geometric mean of the respective entropies to account for the impact of the number of clusters. The values of NMI span from 0 to 1, where 0 indicates a statistical independence between the clusters and the categories and 1 means that the clusters totally agree with the categories.

From an operational standpoint, the clustering is implemented using the MATLAB® programming software (version R2024a) [47]

and, specifically, based on the following functions: the *kmeans* function to perform K-means hard clustering; and the *fitgmdist* function coupled with the *posterior* function to carry out GMMs soft clustering. For both the *kmeans* and the *fitgmdist* functions, the maximum number of iterations is set to 1,000,000 (*MaxIter* option), and the clustering process is repeated with 100,000 different random initializations (*Replicates* option) to prevent the algorithms from converging to local minima. Furthermore, for the *fitgmdist* function a regularization value (*RegularizationValue* option) of 0.1 is added to the diagonal of the covariance matrices to ensure the stability of the EM iterations. Finally, regarding cluster validation, the two selected metrics (i.e., ARI and NMI) are calculated by applying their mathematical definitions to the obtained contingency matrices M.

2.2. Materials

The methodology for SI water-consumption characterization and clustering investigation is applied to a sample of $U = 44$ SI users located in Northern Italy, subject to hourly cumulative water consumption monitoring for a period of one year, thus allowing for the analysis of daily variations, weekly routines and seasonal variability. The monitored SI users are distributed across different urban contexts, ranging from small towns of about 1000 inhabitants to medium-sized cities exceeding 50000 inhabitants. This variety ensures that the study captures SI water-consumption behaviours typical of small-to-medium urban settlements.

Details on the monitored SI users are reported in Table 1. A key strength of the dataset considered in this study, overcoming the limitations of previous research on the clustering of SI activities (e.g., Refs. [38,39]), is the availability of verified category labels for each SI user. Overall, the sample comprises users belonging to the following $C = 6$ categories, representing a broad spectrum of urban

Table 1
Characteristics of the SI users monitored.

User ID	Category	User description	Yearly billed consumption (m^3/year)
S1	Shops	Boat shop and shipyard	70.9
S2	Shops	Supermarket	66.2
S3	Shops	Pharmacy	43.4
S4	Shops	Beverage shop	53.1
S5	Shops	Furniture shop	53.9
S6	Shops	Auto parts store	27.8
S7	Shops	Appliance repair	21.5
S8	Shops	Auto body shop	29.4
S9	Shops	Building materials store	15.8
S10	Shops	Tobacco shop	37.6
S11	Shops	Furniture shop	10.9
S12	Shops	Minimarket	8.0
S13	Shops	Electricity supplier	7.4
S14	Shops	Florist	13.2
S15	Shops	Appliance retailer	9.2
C1	Catering services	Café and restaurant	948.2
C2	Catering services	Restaurant	713.3
C3	Catering services	Café and restaurant	240.0
C4	Catering services	Restaurant	351.4
C5	Catering services	Café, restaurant and pizzeria	235.7
C6	Catering services	Ice-cream shop	647.2
C7	Catering services	Café and restaurant	65.5
C8	Catering services	Café/recreational club	39.6
C9	Catering services	Restaurant	49.6
O1	Offices	Industrial appliances supplier	169.1
O2	Offices	Institutional office	108.9
O3	Offices	Consulting office	48.2
O4	Offices	Technical office	28.7
O5	Offices	Bank office	23.2
O6	Offices	Bank office	9.5
O7	Offices	Technical office	12.0
O8	Offices	Technical office	13.1
W1	Wellness services	Hairdresser	100.5
W2	Wellness services	Hairdresser	65.6
W3	Wellness services	Gym	72.5
W4	Wellness services	Hairdresser	31.9
W5	Wellness services	Beauty salon	48.6
W6	Wellness services	Beauty salon	16.6
R1	Recreational and training activities	Training centre	506.4
R2	Recreational and training activities	Parish	78.1
R3	Recreational and training activities	Boating club	31.6
R4	Recreational and training activities	Driving school	8.5
P1	Police stations	Police station	130.4
P2	Police stations	Police station	103.4
Average			**121.9**

SI:

- *Shops* (users S1-S15), including includes local shops and commercial businesses such as minimarkets, florists, pharmacies, furniture shops, appliance suppliers, etc.
- *Catering services* (users C1-C9), including cafes, restaurants, and ice cream shops.
- *Offices* (users O1-O8), including legal, technical, and real estate agencies.
- *Wellness services* (users W1-W6), including personal care and wellness establishments such as hair salons, beauty centres, gyms.
- *Recreational and training activities* (users R1-R4), including training centres and parish recreational areas.
- *Police stations* (users P1-P2), including two police facilities staffed by officers.

As shown in Table 1, the average billed consumption is 121.9 m^3/year, ranging from a minimum of 7.4 m^3/year (user S13) to a maximum of 948.2 m^3/year (user C1). Consequently, consistent with the previous definition proposed by Baggelaar and Geudens [29], all 44 SI users are classified as *small* users.

Although all SI users were monitored over a one-year period, the monitored year corresponds to a different yearly time window for each user due to the progressive implementation of smart monitoring systems in the water distribution networks considered as a case study. In addition, different metering and data-acquisition technologies were employed, ranging from Sensus iPerl® meters equipped with a radio transmitter using wireless M-Bus communication, whose data are periodically collected onsite by an operator in walk-by

Table 2
Results of data pre-processing: excluded and valid weeks of data for each SI user.

User ID	Number of excluded weeks			Number of valid weeks			
	Data gaps	Long-term inactivity	Post-meter leakages	Winter	Spring	Summer	Autumn
S1	-	-	-	13	13	13	12
S2	2	-	-	11	13	13	13
S3	2	-	-	11	12	13	13
S4	1	1	-	13	13	12	12
S5	4	2	-	9	13	12	12
S6	3	2	-	10	11	12	13
S7	1	-	-	13	13	11	13
S8	1	3	-	12	13	11	12
S9	-	2	-	11	13	13	13
S10	3	-	3	11	11	10	13
S11	1	2	-	10	13	13	13
S12	18	1	-	4	4	12	13
S13	-	22	-	6	6	8	10
S14	22	4	-	9	6	4	6
S15	-	3	-	12	12	12	13
C1	-	-	-	13	13	13	13
C2	5	-	6	12	12	5	11
C3	30	-	-	6	7	5	4
C4	21	-	-	4	13	9	4
C5	4	-	-	12	13	13	9
C6	2	5	10	9	6	9	10
C7	2	-	7	12	12	13	6
C8	1	-	-	13	13	12	12
C9	-	2	2	13	12	12	10
O1	23	1	-	9	5	6	8
O2	17	-	-	8	8	8	11
O3	2	2	-	10	13	12	13
O4	3	2	-	10	11	12	13
O5	-	-	-	13	13	13	13
O6	-	15	-	7	8	9	13
O7	-	4	-	11	13	12	12
O8	4	1	-	9	13	12	13
W1	21	2	-	4	10	10	4
W2	3	2	-	10	11	12	13
W3	4	3	-	11	13	10	11
W4	-	5	-	12	12	9	13
W5	-	1	-	13	13	12	13
W6	3	3	-	10	11	12	12
R1	7	-	-	12	13	12	8
R2	2	1	-	12	12	12	13
R3	-	-	-	13	13	13	12
R4	-	20	-	6	8	9	9
P1	12	-	-	8	9	13	10
P2	4	-	3	13	11	11	9
Average	**5.2**	**2.5**	**0.7**	**10.2**	**11.1**	**10.9**	**11.0**

mode, to Kamstrup FlowIQ® meters that automatically transmit data to a cloud platform via the NB-IoT communication protocol. All devices provide hourly averaged information, and the recorded data are characterized by a volumetric accuracy of 1 L.

3. Results and discussion

The results obtained from the application the methodology to the water-consumption dataset are presented and discussed below for each phase.

The implementation of the initial pre-processing phase, aimed at filling 1-h gaps, identifying post-meter leakages, and detecting seasonal closures, results in the extraction of water-consumption time series over valid weekly windows (i.e. 7-day periods from Monday to Sunday) for each SI user. As far as missing data are concerned, the majority of SI users (approximately 70%, i.e., 31 out of 44) is affected by gaps in the water-consumption time series data: on average, 5.2 weeks per user are removed due to gaps, with a maximum of 30 weeks excluded for user C3. In addition, long-term inactivity periods affect nearly 60% of users (26 out of 44), which leads to the exclusion of an average of about 2.5 weeks per user, in line with typical holiday periods for most activities from the SI in Italy. Finally, the possible presence of post-meter leakages affects approximately 14% of SI users of the sample (i.e., 6 out of 44), with a peak of 10 excluded weeks for user C6. Overall, at least 4 *valid* weeks for each season of the year are available for all SI users. As reported in Table 2, the average number of valid weeks in the SI user sample considered ranges between 10.2 (winter) and 11.1 (spring). In summary, data pre-processing lead to the exclusion of an average of 8.4 weeks of data per user per year, representing more than 16% of the annual records.

Regarding the identification of the days of user activity, the application of *Phase II* leads to the examples illustrated in Fig. 2, where box-whiskers plots are displayed for a group of six SI users, one per category (user S11 for shops, C1 for catering services, O2 for offices, W1 for wellness services, R1 for recreational and training activities, P1 for police stations). In greater detail, the statistical distribution of daily water consumption (L/day) across the seven days of the week is illustrated for a selection of six SI users. For example, user S11 (Fig. 2a) shows a nearly constant daily consumption from Monday to Friday, followed by a slight reduction on Saturdays and a considerable drop on Sundays, with these latter classified as a day of inactivity. The same applies to user O2 (Fig. 2c), whose water consumption is nearly zero over the weekend as a consequence of office closure, and user W1 (Fig. 2d), which exhibits a drop in water consumption on Mondays (weekly closure) and increased consumption on Saturdays likely due to an increase in costumers during leisure days. On the other hand, the water consumption of users that are continuously active or occupied during the week – such as those related to the food and beverage sectors (Fig. 2b), or police stations that are continuously staffed by officers (Fig. 2f) – remains more stable throughout the week, sometimes with no identified weekly days of inactivity.

Overall, the application of *Phase II* to each of the 44 SI users allows identifying the days of user activity (on average, about 6.1 per week), which are considered for the investigation of daily profiles of water consumption (*Phase III*) and their classification (*Phase IV*).

The results obtained with *Phase II* – i.e., subsequently to the identification of the days of activity for each SI user and statistical analysis of the related daily consumption (L/day), including median values and variability – are summarized in Fig. 3. This figure includes a set of box-whisker plots (one for each of the 44 SI users, grouped by category) that statistically summarize the distribution of daily consumption in relation to days of activity, thus excluding weekly closures.

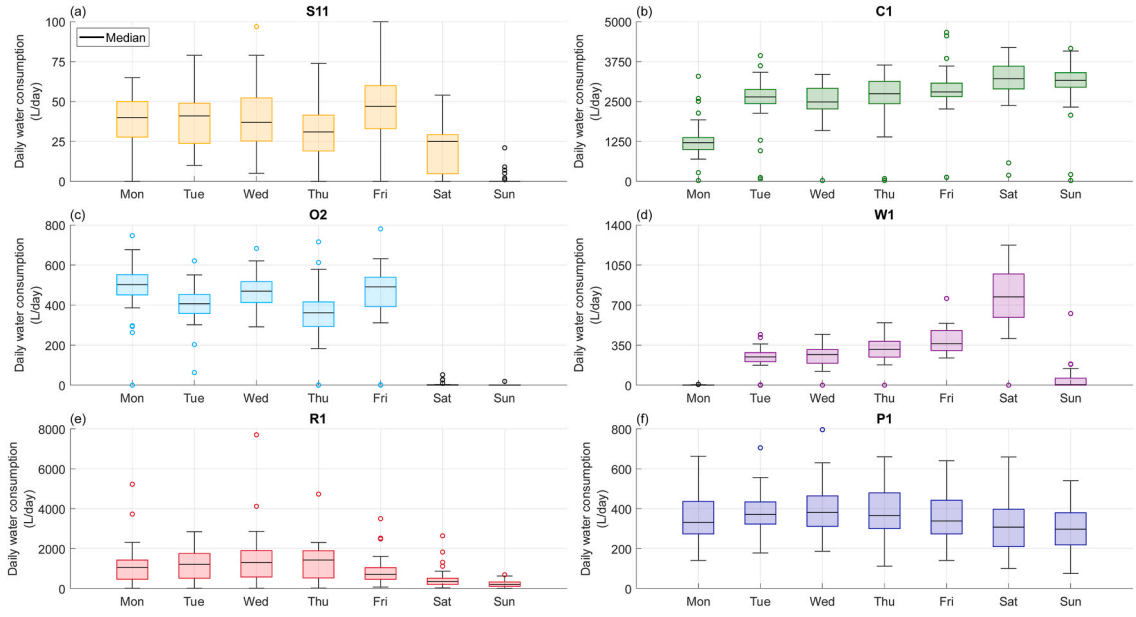


Fig. 2. Identification of days of user activity: box-whisker plots of daily consumption (L/day) for a set of six SI users (one example per category), illustrating the consumption distributions across active and inactive days.

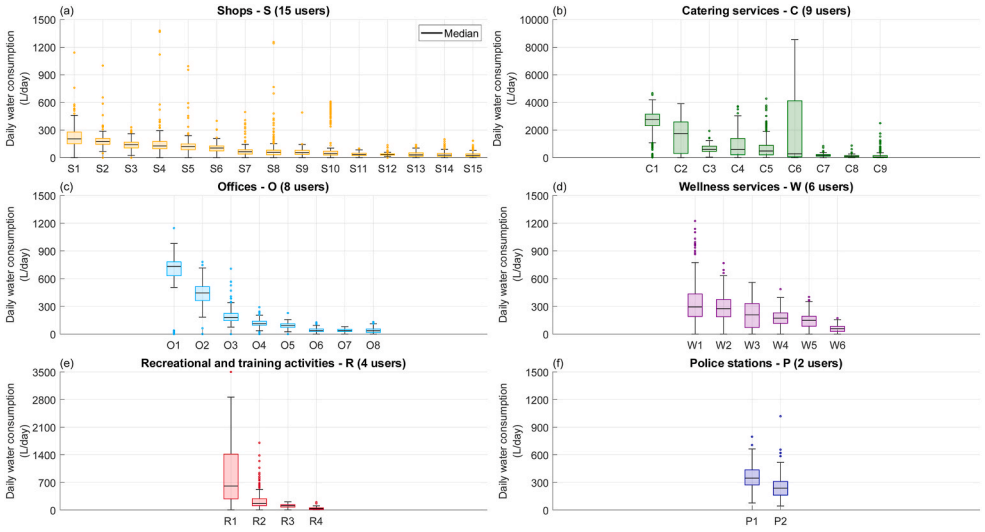


Fig. 3. Box-whisker plots of daily water consumption (L/day) of the 44 monitored SI users grouped by category, in relation to days of activity.

On the one hand, the figure reveals that catering facilities (Fig. 3b) exhibit the highest consumption, with interquartile ranges approaching or exceeding 1000 L/day for most users of the category. Conversely, the statistical distribution of daily consumption for other categories reveals lower values, typically below 1000 L/day. This is particularly evident in the case of shops (Fig. 3a), wellness services (Fig. 3d) and police stations (Fig. 3f), but also applies to the majority of offices (Fig. 3c) and recreational and training activities (Fig. 3e). However, despite the highest values observed for catering services, which suggests that cafés, restaurants, and ice-cream shops use larger amounts of water compared to other SI categories during routine activity days, the wide boxes and whiskers for users C1–C9 reveal substantial variability in daily consumption. This likely reflects pronounced fluctuations in operational routines throughout the year. This is particularly evident for user C6, an ice-cream shop whose productivity significantly increases in spring and summer, when customer demand is high and production intensifies, and decreases in winter, when, although the business remains open, ice-cream production is more limited. Conversely, box-and-whisker plots for shops are considerably narrower, with interquartile ranges typically below 100 L/day and non-outlier data ranges generally below 300 L/day (Fig. 3a). A similar, although slightly more variable, concentration of daily consumption around the median is observed for offices (Fig. 3c), wellness services (Fig. 3d) and police stations (Fig. 3f).

Generally, the analysis highlights significant differences in daily water consumption across categories. Some users, such as shops and police stations, exhibit homogeneous daily consumption, while others show much wider dispersion, reflecting operational routines that might fluctuate more throughout the year. Specifically, catering facilities display the greatest data dispersion, with substantially wider interquartile ranges compared to other categories. Beyond catering services, also offices (Fig. 3c) and recreational and training

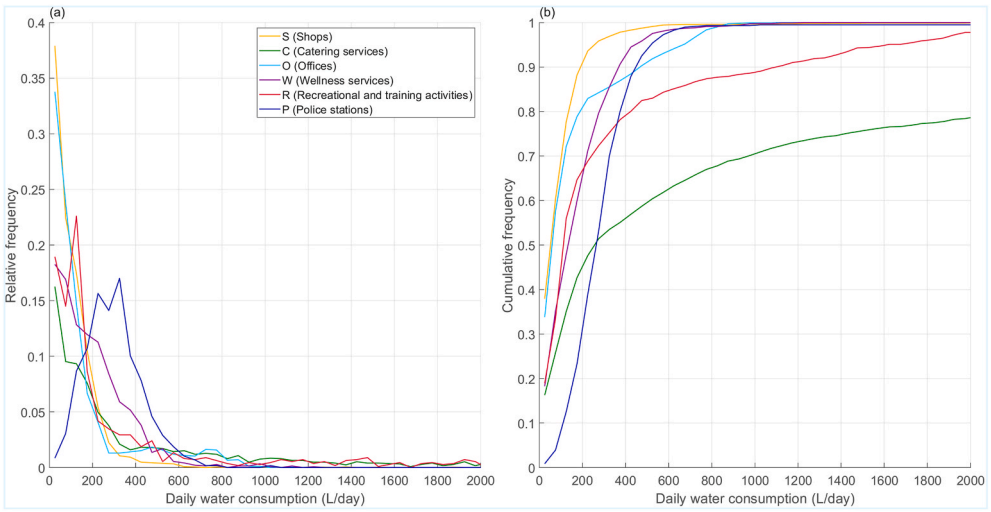


Fig. 4. Relative frequency and cumulative frequency distributions of daily water-consumption values by SI user category, grouped into 50 L/day-width class intervals.

activities (Fig. 3e) show large heterogeneity among users: for instance, the median daily consumption of user O1 and O2 are significantly larger compared to that of users O3–O8, and similarly, user R1 differs substantially from users R2–R4. In addition, for many SI categories (i.e. shops, offices, wellness service and recreational/training activities) a positive correlation emerges between median daily consumption and consumption variability: as the median daily consumption increases, the spread of daily values also tends to increase, indicating that higher average daily consumption may be associated with more variable daily distributions. This relationship does not appear for other categories, such as catering services.

The distribution of daily consumption delivered to all SI users within each category over all active days is shown in Fig. 4a. The distribution is normalized by the total number of active days across all users in each category and therefore represents the empirical relative frequency of daily water consumption by category. Specifically, daily consumption values are grouped into classes of 50 L/day. The corresponding cumulative frequency distribution is presented in Fig. 4b. Overall, across all six categories, most daily consumption values fall within the range of 0–400 L/day (Fig. 4a). On the one hand, four categories – namely shops, catering services, offices, and wellness services – exhibit a decreasing relative frequency distribution. On the other hand, recreational/training activities and police stations show an initial increasing trend, with peaks occurring at approximately 150 L/day and 300 L/day, respectively. In addition, the cumulative frequency distribution (Fig. 4b) reveals that, for nearly all SI user categories, 80% of daily consumption values are below 400 L/day. For shops and offices, 80% of values are even below 200 L/day. Catering services represent the only exception: about half of daily consumption values are above 400 L/day, while approximately 30% exceed 1000 L/day and more than 20% exceed 2000 L/day, confirming the higher water consumption of this category compared to other ones.

With regard to daily water-consumption profiles, the results of *Phase III* are presented in Fig. 5 andFig. 6, where each panel shows the average hourly distribution of water consumption observed on active days across different seasons of the year, along with the corresponding yearly average. Similarly, Table 3 reports the average hourly consumption per SI user and season, facilitating a more detailed comparison of water-consumption variability across seasons.

Concerning the average profiles, it is worth noting that several users exhibit consumption peaks during time windows when operational activity is expected to be negligible. For instance, evening consumption peaks are observed for users S4 (beverage shop; Fig. 5d) and O6 (bank office; Fig. 6f), whereas a morning peak at 7:00 occurs for user S11 (furniture shop; Fig. 5k). These behaviours, apparently anomalous compared to the rest of the analysed sample, demonstrate that peak water consumption does not necessarily coincide with the user's activity or maximum customer attendance. Probably, these occurrences are attributable to automated processes scheduled during specific time windows, or cleaning operations conducted outside of standard opening hours.

Regarding the analysis of seasonal profiles and their variations, it is primarily observed that the majority of monitored users maintain similar consumptions habits throughout the year, both in terms of profile shape and overall magnitude. Consequently, it can be asserted that seasonality generally appears to have a negligible impact on consumption variability for users belonging to certain categories. For instance, shops and offices are expected to operate uniformly during the year, resulting in water-consumption profiles that do not exhibit significant seasonal fluctuations. However, certain exceptions are observed, such as users S13 (electricity supplier; Fig. 5m), S14 (florist; Fig. 5n) and O1 (industrial appliances supplier; Fig. 6a).

Conversely, for a limited number of users, significant seasonal variability is observed both in terms of average hourly consumption and the specific trend of the consumption profile. More specifically, water consumption of catering services category is significantly influenced by seasonality: for example, user C6 (ice-cream shop) shows a limited water consumption in autumn and winter, while it considerably increases in spring and summer (Fig. 5u). Similarly, catering services C2 (restaurant; Fig. 5q), C3 (restaurant; Fig. 5r), and

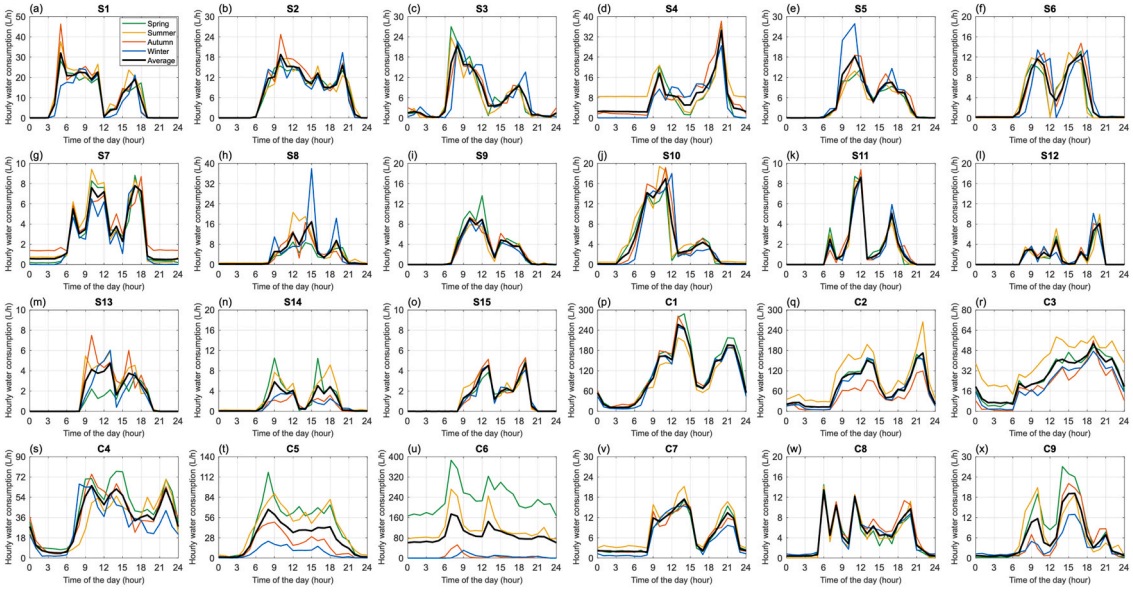


Fig. 5. Daily profiles of water consumption (L/h) for each user: seasonal profiles and yearly average profile (part I).

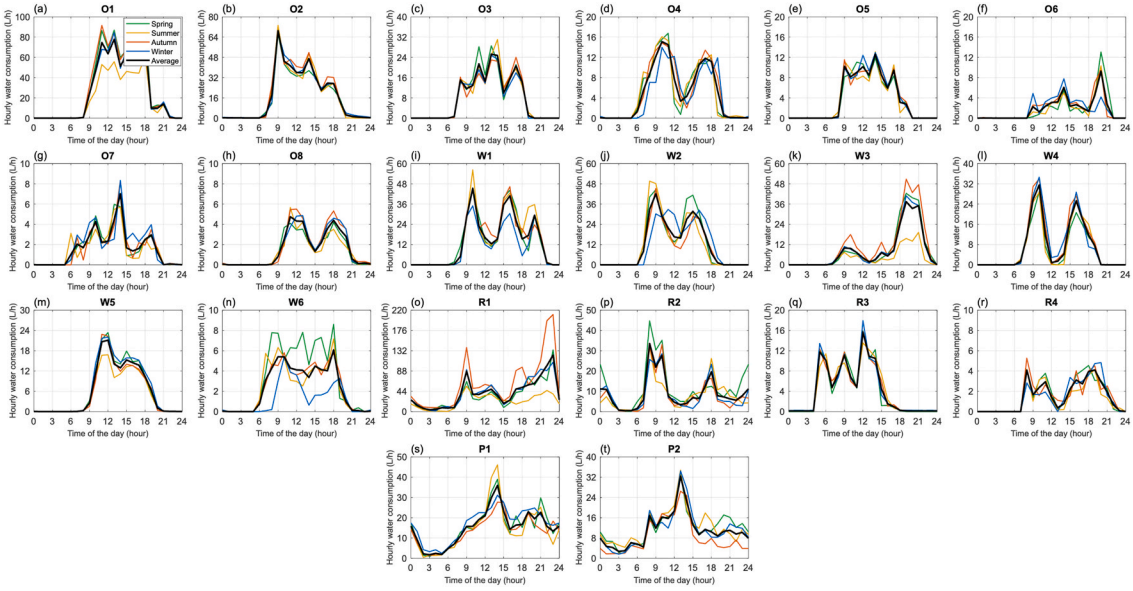


Fig. 6. Daily profiles of water consumption (L/h) for each user: seasonal profiles and yearly average profile (part II).

C5 (café, restaurant, and pizzeria; Fig. 5t) show a consumption that peaks in the spring and summer, while it reduces during the cold season. For the aforementioned users, the yearly average profile is considerably different from those related to individual seasons; therefore, the application of their yearly average profile for evaluations and assessments tied to specific seasons (i.e., water-demand modelling during summer) would be misleading.

Finally, a limited number of users exhibit shape variations of their consumption profiles across different seasons. In particular, winter profiles diverge significantly from those of other seasons for users S5 (furniture shop; Fig. 5e) and S8 (auto body shop; Fig. 5h), whereas for user R1 (training centre; Fig. 6o) a variation occurs in autumn. Instead, for user W6 (beauty salon; Fig. 6n) spring and winter profiles deviate from the summer and autumn ones, with the latter remaining more consistent with the annual average profile.

Overall, the profiles shown in Fig. 5 and Fig. 6 (hourly coefficients of which are included in Table S1 of the *Supplementary Materials*) display substantial variability in shape, peak timing and magnitude across SI users, even within the same category. A similar pattern emerges from Fig. 7, which compares the dimensionless yearly average profiles by category. Despite the heterogeneity observed within the different categories, some common patterns can be identified. For instance, shops typically exhibit consumption peaks in the mid-morning and mid-afternoon, with a reduction around lunchtime due to temporary closure; conversely, catering facilities and police stations display more stable consumption throughout the day, while offices tend to concentrate water use around mid-day. In line with these observations, the box-and-whisker plot of the daily water-consumption profiles (Fig. S1 of the Supplementary Materials) confirms that consumption variability is concentrated during daylight hours. In contrast, during nighttime hours, SI users' behaviour tends to be more consistent, with significantly reduced consumption.

Finally, with regard to *Phase IV*, a preliminary silhouette analysis is performed to identify the optimal number of clusters. Different configurations are tested by progressively increasing the number of clusters (K) from a minimum of 2 to a maximum of U, where $U = 44$ corresponds to the number of monitored SI users. For each configuration, the K-means algorithm is applied 100 times to classify normalized profiles into clusters. The resulting silhouette values are then averaged to evaluate the overall performance of the method for each possible value of K (Fig. S2 of the *Supplementary Materials*).

Overall, the analysis reveals that: (*i*) the average of silhouette values shows a monotonically increasing trend as a function of the number of clusters, without clear local maxima; (*ii*) silhouette values remain relatively low (below 0.5) for K-values lower than $U/2$; and (*iii*) silhouette values exceed 0.5 only for very high K-values (up to the 1 for $K = U$), suggesting overfitting of the classification problem. In light of these results, and following the heuristic approach proposed by Han et al. [48], clustering is performed assuming $K \cong \sqrt{U/2}$, corresponding to five clusters.

Applying the K-means algorithm with $K = 5$ leads to the partition of the dimensionless average profiles into the groups illustrated in Fig. 8. As illustrated in the figure and further detailed in the contingency matrix relating clusters to user categories (Table 4), five distinct consumption patterns are identified.

- *Cluster 1* (Fig. 8a), comprising 14 users, is the largest group and includes users that typically reduce their activity in the early afternoon (approximately 13:00–14:00). A bimodal distribution of consumption is observed, with two distinct peaks at 10:00 and 16:00 separated by a local minimum around 13:00. Confirming intra-cluster heterogeneity, this group includes users from five different categories, predominantly shops (six users) and wellness services (four users).

Table 3
Average water consumption (L/h) of SI users, per season.

User ID	Average hourly water consumption (L/h)			
	Winter	Spring	Summer	Autumn
S1	7.7	9.3	11.1	10.3
S2	7.3	7.3	7.4	8.2
S3	6.3	6.2	5.5	5.9
S4	6.6	6.1	10.5	8.8
S5	6.6	4.6	4.3	5.5
S6	4.0	3.7	3.8	4.4
S7	2.4	2.8	3.2	3.4
S8	4.9	3.0	4.9	3.4
S9	2.3	3.0	2.4	2.5
S10	4.2	4.2	5.0	4.7
S11	1.5	1.6	1.4	1.5
S12	1.4	1.4	1.4	1.6
S13	1.3	1.0	1.8	1.9
S14	1.1	2.3	2.6	1.0
S15	1.2	1.2	1.3	1.3
C1	103.4	118.6	96.9	111.5
C2	66.1	71.5	101.0	44.9
C3	21.4	27.5	39.8	20.6
C4	30.2	42.3	32.7	36.8
C5	8.0	43.6	42.7	18.2
C6	5.6	235.2	119.4	6.6
C7	6.0	6.9	8.8	6.9
C8	4.6	4.3	4.3	5.2
C9	3.8	8.0	6.4	5.6
O1	27.2	29.9	19.3	30.9
O2	18.1	16.3	17.5	19.1
O3	7.4	8.2	8.0	7.6
O4	4.5	4.9	5.2	5.3
O5	3.9	3.6	3.5	3.8
O6	1.8	1.8	1.6	2.0
O7	1.9	1.6	1.5	1.6
O8	1.7	1.4	1.4	1.8
W1	10.8	14.1	15.2	14.5
W2	11.8	12.6	11.7	10.3
W3	8.9	9.5	4.4	11.9
W4	7.8	6.0	6.4	8.0
W5	6.7	6.4	5.1	5.8
W6	1.3	3.4	2.4	2.4
R1	41.5	35.9	23.2	60.7
R2	8.2	11.8	8.4	8.1
R3	4.3	4.1	4.3	4.3
R4	1.5	1.6	1.3	1.6
P1	16.8	15.3	14.4	13.2
P2	11.4	12.2	12.3	8.7
Average	**11.5**	**18.6**	**15.6**	**12.3**

- *Cluster 2* (Fig. 8b) groups users exhibiting relatively continuous consumption throughout the working day (approximately 08:00–19:00). This cluster is mainly composed of offices (five users) and shops (four users).
- *Cluster 3* (Fig. 8c) includes profiles with consumption peaking around midday (13:00) and extending into the late evening (up to approximately 23:00), with minimum values occurring at night but not always reaching zero. This pattern is consistent with SI users that remain active into the late evening or early night, such as cafés and restaurants. In fact, this cluster includes six of the nine monitored catering facilities, together with both police stations (P1 and P2), which maintain relatively constant consumption throughout the day due to their continuous staffing and operation.
- *Cluster 4* (Fig. 8d) groups users characterized by higher water consumption in the morning and lower consumption in the afternoon. This pattern is associated with SI activities concentrated in the morning or requiring substantial water use early in the day due to cleaning operations or other water-related processes. Specifically, the cluster includes a mix of categories, comprising three shops, two catering facilities and two recreational activities, with some users that show significant demand in the early morning hours (between 05:00 and 07:00). A clear example of this is provided by the two catering facilities included in this group (C5 and C8). Since they offer a café service, their significant water consumption is well justified by the preparation of breakfast services and the arrival of the early customers.
- *Cluster 5* (Fig. 8e) includes four users characterized by an evening peak in consumption (approximately 19:00–21:00). The cluster comprises two shops, one office, and one wellness-related activity. The evening peak is likely due to cleaning operations in the case

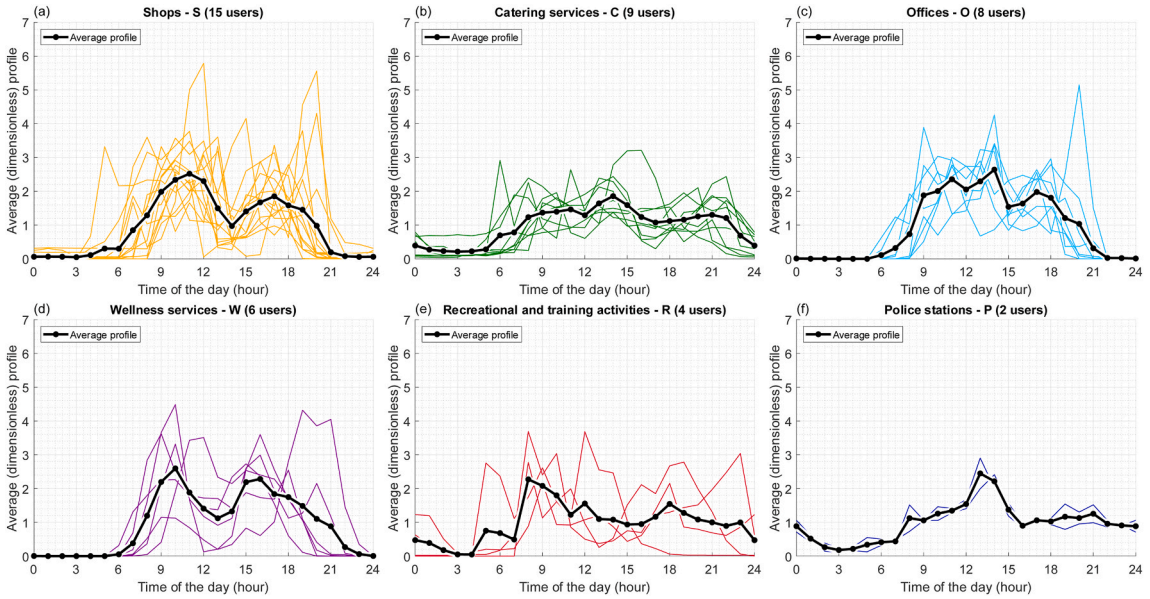


Fig. 7. Water-consumption daily profiles by category. Thin, coloured lines refer to individual annual average dimensionless profiles, while thick, black lines denote the average profile for each category.

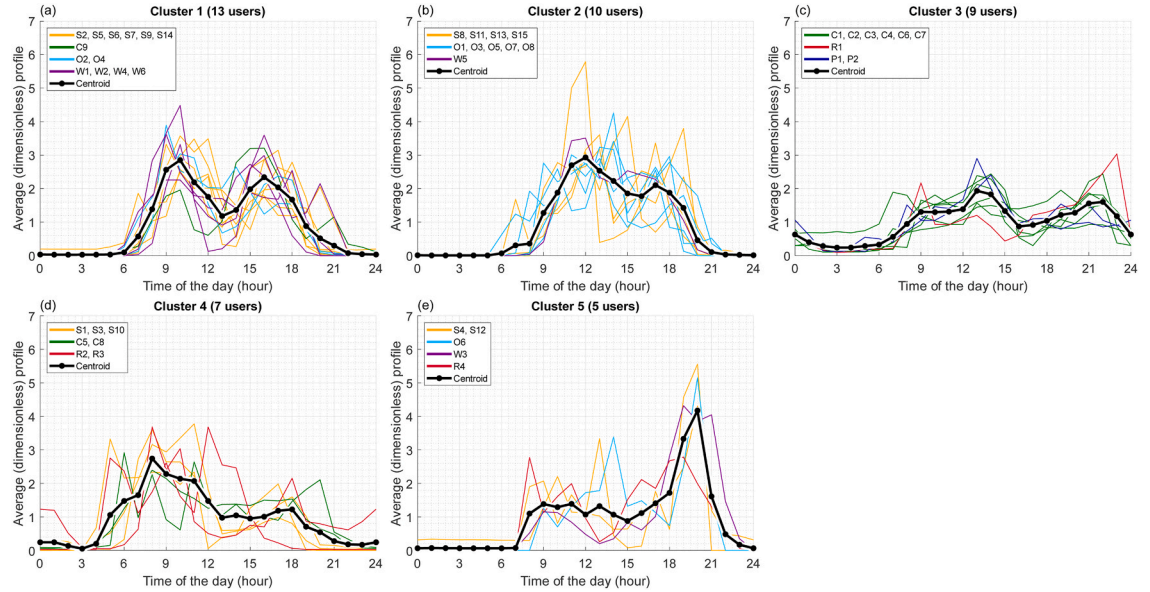


Fig. 8. Daily dimensionless water consumption profiles grouped by K-means clustering ($K = 5$). Thin, coloured lines represent individual annual average dimensionless profiles assigned to each cluster, while the thick, black line denotes the cluster's centroid.

Table 4
Contingency matrix relating K-means clusters and user categories.

Cluster	User category					
	S	C	O	W	R	P
1	6	1	2	4	1	0
2	4	0	5	1	0	0
3	0	6	0	0	1	2
4	3	2	0	0	2	0
5	2	0	1	1	0	0

of shops and offices, whereas it likely reflects actual water use by customers for user W3 which, being a gym, probably concentrates classes and activities in the late afternoon and evening hours.

As indicated by the low silhouette values (regardless of the number of clusters) and the results reported in Table 4 in relation to $K = 5$ clusters, the clustering performed via K-means on the monitored SI users leads to a partition characterized by limited geometric separation among the groups. This outcome further confirms the heterogeneity of SI user profiles, highlighting differences in behaviour even within the same category.

The results obtained using the K-means algorithm are similar to those related to the application of GMMs, this latter implemented using the same number of clusters previously considered for K-means (i.e., $K = 5$). Regarding the GMMs application, the algorithm yields a 100% posterior probability of belonging to a specific cluster for 43 out of the 44 monitored SI users, while for a single user – corresponding to user S5 – this probability was 99.9%. However, since the residual probability for alternative clusters of this single case is extremely low (i.e., 0.1%), each user is assigned to a single cluster, making the GMMs partition converge into a hard clustering approach. As far as the evaluation metrics are concerned, K-means achieves an ARI of 0.13 and an NMI of 0.34, while GMMs displays an ARI of 0.19 and an NMI of 0.38. The values of the selected performance metrics indicate that in this case different classification approaches lead to similar results, providing similar contingency matrices. Overall, the relatively low values of the selected performance metrics highlight the moderate agreement between the identified clusters and the SI user categories. In order to interpret this result, it is worth noting that the limited sample size considered (i.e., 44 users distributed across six SI categories) may affect the ability of the clustering algorithms to define robust cluster structures. However, it is also worth noting that the actual heterogeneity of water-consumption profiles within each SI category (as shown in Fig. 6 and Fig. 7) and the broad definition adopted for some SI categories in this study, which may include heterogeneous activities characterized by different water-consumption profiles, both contribute to reducing the separability among clusters.

Overall, the findings suggest that, from the perspective of daily profiles, water consumption is also shaped by operational schedules and activity patterns rather than only depending on user category per se. While the deterministic patterns reported in the literature for predefined SI categories (such as the average profiles proposed by Walski et al. [49]) provide a valuable baseline, these findings suggest that general consumption trends should be interpreted with flexibility, accounting for intra-category behavioural variability.

4. Study Limitations and Future Developments

This study offers numerous insights into the water consumption of SI users. Nevertheless, some limitations should be acknowledged, coupled with a discussion of specific future developments to address them.

Firstly, a limitation is related to the generalizability of the results, as the analysed dataset is based on a sample of 44 SI users located exclusively in Northern Italy. Water-consumption profiles in other areas may differ significantly from the ones presented in this study due to different climates (e.g., where higher temperatures require an increased water usage for cooling or for outdoor purposes), cultures (e.g., different working schedules or holiday periods) and socio-economic contexts. For these reasons, expanding the dataset with the aim to include a significant number of SI users across other categories, diverse geographical contexts and different operational frameworks, will be crucial to generalize these water usage patterns. Furthermore, to improve the interpretability of the findings, additional explanatory variables, such as floor area, number of employees or customers and specific operational characteristics, should be included in future analyses.

Another notable aspect is that the proposed method assumes that the weekly activity pattern remains constant across seasons for each SI user. Although this is reasonable for commercial and administrative SI users, other sectors may be strongly affected by seasonality. For instance, tourism-related activities or businesses located in seasonal holiday destinations (e.g., coastal resorts) may exhibit shifting operational schedules season by season. Further studies should incorporate methods to accurately capture variations in water-consumption routines throughout the year.

5. Conclusions

This study is carried out with the aim of analysing water use data from service-industry (SI) users, to contribute to classifying water consumption and identifying possibly recurrent profiles. To this aim, a multi-phase methodology is applied to a dataset of 44 service-industry (SI) users belonging to small-to-medium urban contexts in Northern Italy, for which information on their category was available. These users were monitored using smart meters with an hourly resolution and consumption data were collected over a one-year period.

The analysis provides several insights into the behaviour of SI users.

- Water consumption does not always align with opening hours; many users exhibit distinct days of activity and inactivity that reflect actual water use rather than business operation alone. In some cases, peaks occur outside customer service hours, often associated with cleaning or maintenance operations.
- Among the monitored categories, catering facilities show the highest average daily consumption and the greatest statistical variability, reflecting highly diverse operational routines both across users and depending on the day of the week or the period of the year.

- While seasonal effects are generally limited for most users, some catering activities exhibit strong seasonality, with annual average profiles that differ substantially from seasonal ones. For such users, relying solely on an annual average profile may lead to oversimplifications for modelling purposes.
- Unsupervised clustering of the average dimensionless profiles showed limited performance in partitioning the profiles and matching the identified clusters with user categories. While some categories tend to be associated with certain clusters, other clusters display marked heterogeneity, since users from different categories can exhibit similar daily profiles. This was quantitatively confirmed by the low values obtained for both the Adjusted Rand Index ($\approx 0.10 \div 0.20$) and Normalized Mutual Information ($\approx 0.30 \div 0.40$).

In conclusion, the findings of this study reveal that classifying SI water-consumption profiles is challenging and cannot rely solely on user category. The variability observed within the same SI category, along with the results obtained from clustering, suggests that category-based classifications alone are not sufficient to describe actual water-consumption behaviours. Nevertheless, in the absence of high-resolution water-consumption data for the specific SI user, category-based average daily profiles may still represent a useful tool for preliminary water-demand estimations, water-balance assessments, or modelling applications. From a practical perspective, the proposed dataset can provide a first benchmark for the characterization of SI water-consumption profiles across different categories, supporting applications related to preliminary estimation of SI consumption, the evaluation of the variability of water-consumption profiles within a given SI category, and non-residential water-demand modelling.

CRediT authorship contribution statement

Jacopo Lodi: Data curation, Formal analysis, Methodology, Software, Writing – original draft, Writing – review & editing. **Vittorio Micai:** Data curation, Formal analysis, Methodology, Software, Writing – original draft, Writing – review & editing. **Valentina Marsili:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Resources, Software, Validation, Visualization, Writing – review & editing. **Filippo Mazzoni:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Resources, Software, Validation, Visualization, Writing – review & editing. **Stefano Alvisi:** Conceptualization, Methodology, Project administration, Resources, Supervision, Validation, Visualization, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.wri.2026.100380.

Data availability

The authors confirm that the data supporting the findings of this study are available within the article and its supplementary materials.

References

[1] F.J. Suero, P.W. Mayer, D.E. Rosenberg, Estimating and verifying United States households' potential to conserve water, J. Water Resour. Plann. Manag. 138 (2012) 299–306, https://doi.org/10.1061/(ASCE)WR.1943-5452.0000182.
[2] K.A. Nguyen, R.A. Stewart, H. Zhang, C. Jones, Intelligent autonomous system for residential water end-use classification: autoflow, Appl. Soft Comput. 31 (2015) 118–131, https://doi.org/10.1016/j.asoc.2015.03.007.
[3] N. Avni, B. Fishbain, U. Shamir, Water consumption patterns as a basis for water demand modeling, Water Resour. Res. 51 (2015) 8165–8181, https://doi.org/10.1002/2014WR016662.
[4] R. Cardell-Oliver, J. Wang, H. Gigney, Smart meter analytics to pinpoint opportunities for reducing household water use, J. Water Resour. Plann. Manag. 142 (2016) 04016007, https://doi.org/10.1061/(ASCE)WR.1943-5452.0000634.
[5] G.M. Sanchez, J.W. Smith, A. Terando, G. Sun, R.K. Meentemeyer, Spatial patterns of development drive water use, Water Resour. Res. 54 (3) (2018) 1633–1649, https://doi.org/10.1002/2017WR021730.
[6] L.E. Danielson, An analysis of residential demand for water using micro time-series data, Water Resour. Res. 15 (4) (1979) 763–767, https://doi.org/10.1029/WR015i004p00763.
[7] S.A. Tanverakul, J. Lee, Residential water demand analysis due to water meter installation in California, Proc World Environ Water Resour Congr (2013) 936–945, https://doi.org/10.1061/9780784412947.090.
[8] A. Cominola, M. Giuliani, D. Piga, A. Castelletti, A.E. Rizzoli, Benefits and challenges of using smart meters for advancing residential water demand modeling and management: a review, Environ. Model. Software 72 (2015) 198–214, https://doi.org/10.1016/j.envsoft.2015.07.012.
[9] D. Fiorillo, G. Galuppini, E. Creaco, F. De Paola, M. Giugni, Identification of influential user locations for smart meter installation to reconstruct the urban demand pattern, J. Water Resour. Plann. Manag. 146 (8) (2020) 04020070, https://doi.org/10.1061/(ASCE)WR.1943-5452.0001269.
[10] M. Arnell, M. Miltell, G. Olsson, Making waves: a vision for digital water utilities, Water Res. X 19 (2023) 100170, https://doi.org/10.1016/j.wroa.2023.100170.

[11] F. Abate, M. Carratù, C. Liguori, A. Pietrosanto, Smart meters and water leakage detection: a preliminary study, in: Proceedings of the IEEE International Instrumentation Measurement Technology Conference, 20–23 May 2019, https://doi.org/10.1109/I2MTC.2019.8827020. Auckland, New Zealand.
[12] F. Mazzoni, E.J.M. Blokker, S. Alvisi, M. Franchini, An enhanced method for automated end-use classification of household water data, J Hydroinformatics 26 (2) (2024) 408–423, https://doi.org/10.2166/hydro.2024.125.
[13] C. Beal, R. Stewart, Identifying residential water end uses underpinning peak day and peak hour demand, J. Water Resour. Plann. Manag. 140 (2013), https://doi.org/10.1061/(ASCE)WR.1943-5452.0000357.
[14] G. Cole, R.A. Stewart, Smart meter enabled disaggregation of urban peak water demand: precursor to effective urban water planning, Urban Water J. 10 (3) (2013) 174–194, https://doi.org/10.1080/1573062x.2012.716446.
[15] S. Alassio, V. Marsili, F. Mazzoni, S. Alvisi, Exploring residential minimum night consumption in a real water distribution network based on smart-meter data, Discov Water 4 (2024) 88, https://doi.org/10.1007/s43832-024-00150-5.
[16] C. Giudicianni, F. Mazzoni, V. Marsili, S. Alvisi, E. Creaco, Modelling minimum night consumption as a function of long-run average outflow to users and applications to real world case studies, Front. Water 7 (2025), https://doi.org/10.3389/frwa.2025.1693893.
[17] W.B. DeOreo, J.P. Heaney, P.W. Mayer, Flow trace analysis to assess water use, J. Am. Water Works Assoc. 88 (1) (1996) 79–90, https://doi.org/10.1002/j.1551-8833.1996.tb06487.x.
[18] S. Gato-Trinidad, K. Gan, Characterizing maximum residential water demand, Urban Water 122 (2012) 15–24, https://doi.org/10.2495/UW120021.
[19] Cardell-Oliver R. Discovering water use activities for smart metering. Proceedings of the IEEE Eighth International Conference on Intelligent Sensors, Sensor Network and Information Processing. 2–5 April 2013; Melbourne, Australia. https://doi.org/10.1109/issnip.2013.6529784.
[20] F. Mazzoni, S. Alvisi, M. Franchini, E.J.M. Blokker, Exploiting high-resolution data to investigate the characteristics of water consumption at the end-use level: a Dutch case study, Water Resour. Ind. 29 (2023) 100198, https://doi.org/10.1016/j.wri.2022.100198.
[21] K. Aksela, M. Aksela, Demand estimation with automated meter reading in a distribution network, J. Water Resour. Plann. Manag. 137 (2011) 456–467, https://doi.org/10.1061/(ASCE)WR.1943-5452.0000131.
[22] M.A. Morales, J.P. Heaney, Classification, benchmarking, and hydro-economic modeling of non-residential water users, J. Am. Water Works Assoc. 106 (2014) E550–E560, https://doi.org/10.5942/jawwa.2014.106.0150.
[23] N.A. Attallah, D.E. Rosenberg, J.S. Horsburgh, Water end-use disaggregation for six non-residential facilities in Logan, Utah, J. Water Resour. Plann. Manag. 147 (7) (2021) 05021006, https://doi.org/10.1061/(ASCE)WR.1943-5452.0001367.
[24] B. Dziegielewski, J.C. Kiefer, E.M. Opitz, G.A. Porter, G.L. Lantz, W.B. DeOreo, et al., Commercial and Institutional End Uses of Water, American Water Works Association Research Foundation, 2000.
[25] J.F. Bonnet, C. Devel, P. Faucher, J. Roturier, Analysis of electricity and water end-uses in university campuses: case-study of the University of Bordeaux in the framework of the Ecocampus European Collaboration, J. Clean. Prod. 10 (1) (2002) 13–24, https://doi.org/10.1016/S0959-6526(01)00018-X.
[26] Morales MA, Heaney JP. Estimating non-residential water use with publicly available databases. Proceedings of the World Environmental & Water Resources Congress. 16–20 May 2010; Providence, United States of America. https://doi.org/10.1061/41114(371)435.
[27] M. Farina, M. Maglionico, M. Pollastri, I. Stojkov, Water consumption in public schools for the city of Bologna, Italy. Water Sci Technol 13 (2) (2013) 257–264, https://doi.org/10.2166/ws.2013.013.
[28] E.J.M. Blokker, E.J. Pieterse-Quirijns, J.H.G. Vreeburg, J.C. Van Dijk, Simulating nonresidential water demand with a stochastic end-use model, J. Water Resour. Plann. Manag. 137 (2011) 511–520, https://doi.org/10.1061/(asce)wr.1943-5452.0000146.
[29] P.K. Baggelaar, P.J.J.G. Geudens, Prognoses Landelijke Drinkwatervraag Tot 2025, Vewin, 2008.
[30] J.S. Horsburgh, M.E. Leonardo, A.M. Abdallah, D.E. Rosenberg, Measuring water use, conservation, and differences by gender using an inexpensive, high frequency metering system, Environ. Model. Software 96 (2017) 83–94, https://doi.org/10.1016/j.envsoft.2017.06.035.
[31] E. Clifford, S. Mulligan, J. Comer, L. Hannon, Flow-signature analysis of water consumption in non-residential building water networks using high-resolution and medium-resolution smart meter data: two case studies, Water Resour. Res. 54 (2018) 88–106, https://doi.org/10.1002/2017WR020639.
[32] E.J. Pieterse-Quirijns, E.J.M. Blokker, E. Van Der Blom, J.H.G. Vreeburg, Non-residential water demand model validated with extensive measurements and surveys, Drink. Water Eng. Sci. 6 (2013) 99–114, https://doi.org/10.5194/dwes-6-99-2013.
[33] S. Kara, E. Karadirek, A. Muhammetoglu, H. Muhammetoglu, Hydraulic modeling of a water distribution network in a tourism area with highly varying characteristics, Procedia Eng. 162 (2016) 521–529, https://doi.org/10.1016/j.proeng.2016.11.096.
[34] F. Mazzoni, V. Marsili, S. Alvisi, M. Franchini, Exploring the impacts of tourism and weather on water consumption at different spatiotemporal scales: evidence from a coastal area on the Adriatic Sea (northern Italy), Environ. Res. Infrastruct. Sustain. 2 (2022) 025005, https://doi.org/10.1088/2634-4505/ac611f.
[35] G.Z. Wu, K. Sakaue, S. Murakawa, Verification of calculation method using Monte Carlo method for water supply demands of office building, Water 9 (6) (2017) 376, https://doi.org/10.3390/w9060376.
[36] Murakawa S, Nishina D, Takata H, Tanaka A. An analysis on the loads of hot water consumption in the restaurants. Proceedings of the CIB W062 31st International Symposium on Water Supply and Drainage for Buildings. 14–16 September 2005; Brussels, Belgium.
[37] S. Murakawa, Y. Koshikawa, H. Takata, A. Tanaka, Calculation for the cold and hot water demands in the guest rooms of city hotel. Proceedings of the CIB W062 34th International Symposium on Water Supply and Drainage for Buildings, 8–10 September 2008. Hong Kong, China.
[38] S. McKenna, F. Fusco, B. Eck, Water demand pattern classification from smart meter data, Procedia Eng. 70 (2014) 1121–1130, https://doi.org/10.1016/j.proeng.2014.02.124.
[39] N. Cheifetz, Z. Noumir, A. Same, A.C. Sandraz, C. Féliers, V. Heim, Modeling and clustering water demand patterns from real-world smart meter data, Drink. Water Eng. Sci. 10 (2017) 75–82, https://doi.org/10.5194/dwes-10-75-2017.
[40] S. Alvisi, M. Franchini, C. Luciani, I. Marzola, F. Mazzoni, Effects of the COVID-19 lockdown on water consumptions: northern Italy case study, J. Water Resour. Plann. Manag. 147 (11) (2021) 05021021, https://doi.org/10.1061/(ASCE)WR.1943-5452.0001481.
[41] S. Lloyd, Least squares quantization in PCM, IEEE Trans. Inf. Theor. 28 (1982) 129–137, https://doi.org/10.1109/tit.1982.1056489.
[42] A.M. Ikotun, A.E. Ezugwu, L. Abualigah, B. Abuhaija, J. Heming, K-means clustering algorithms: a comprehensive review, variants analysis, and advances in the era of big data, Inf. Sci. 622 (2023) 178–210, https://doi.org/10.1016/j.ins.2022.11.139.
[43] P.J. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, J. Comput. Appl. Math. 20 (1987) 53–65, https://doi.org/10.1016/0377-0427(87)90125-7.
[44] W.H. Press, S.A. Teukolsky, W.T. Vetterling, B.P. Flannery, Numerical Recipes: the Art of Scientific Computing, third ed., Cambridge University Press, Cambridge, 2007.
[45] L. Hubert, P. Arabie, Comparing partitions, J. Classif. 2 (1985) 193–218, https://doi.org/10.1007/BF01908075.
[46] A. Strehl, J. Ghosh, Cluster ensembles—a knowledge reuse framework for combining multiple partitions, J. Mach. Learn. Res. 3 (2002) 583–617.
[47] The MathWorks Inc, MATLAB and Statistics and Machine Learning Toolbox, Release R2024a. Natick (MA), The MathWorks Inc., 2024.
[48] J. Han, M. Kamber, J. Pei, Data Mining: Concepts and Techniques, third ed., Morgan Kaufmann, Waltham (MA), 2011.
[49] T. Walski, D.V. Chase, D. Savic, W. Grayman, S. Beckwith, E. Koelle, Advanced Water Distribution Modeling and Management, first ed., Haestad Press, 2003. Waterbury (CT).