AQUA
Water Infrastructure, Ecosystems and Society

IWA PUBLISHING

 AQUA — Water Infrastructure, Ecosystems and Society Vol 73 No 5, 870 doi: 10.2166/aqua.2024.294

Identifying daily water consumption patterns based on K-means Clustering, Agglomerative Hierarchical Clustering, and Spectral Clustering algorithms

Hongyuan Guo, Xingpo Liu* and Qichen Zhang

College of Ocean Science and Engineering, Shanghai Maritime University, Shanghai 201306, China
*Corresponding author. E-mail: stormmodel@163.com; xpliu@shmtu.edu.cn

ABSTRACT

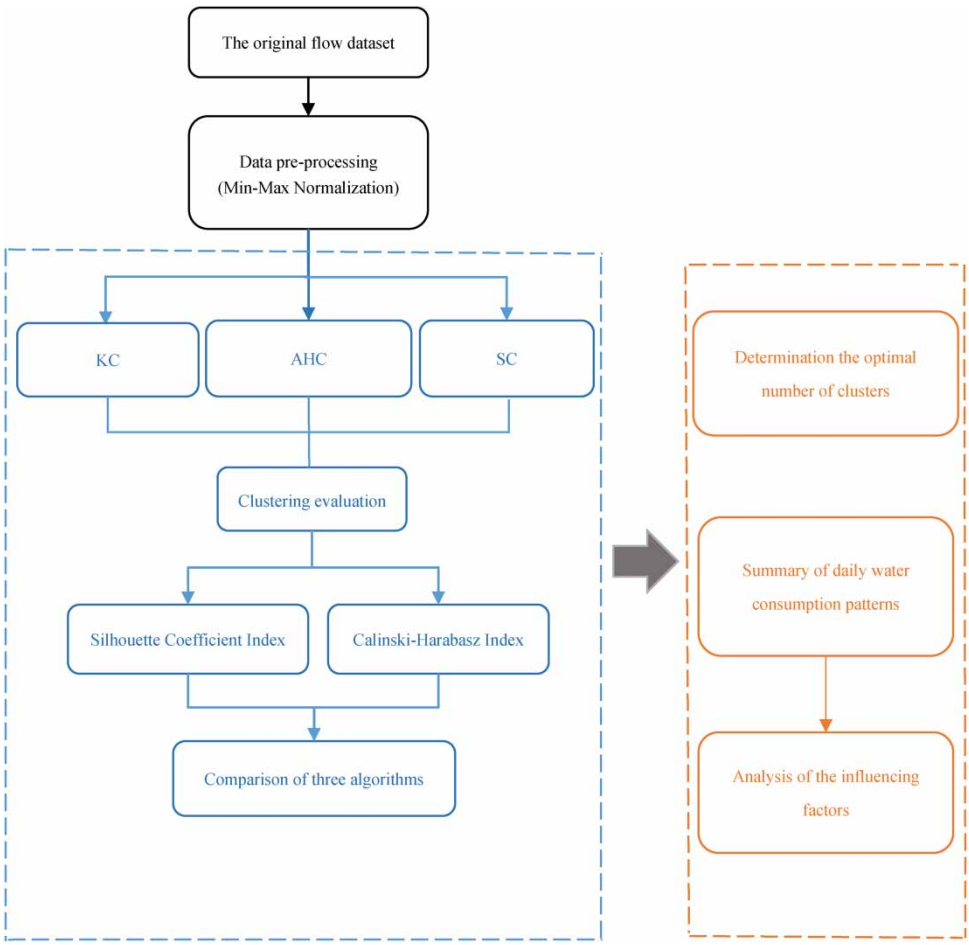
Understanding daily water consumption patterns is crucial for efficient management and distribution of water resources, as well as for promoting energy conservation and achieving carbon peaking and neutrality targets. It compares performance of three clustering algorithms, K-means Clustering (KC), Agglomerative Hierarchical Clustering (AHC), and Spectral Clustering (SC), using Silhouette Coefficient Index (SCI) and Calinski–Harabasz Index (CHI) as evaluation metrics. We conducted a case study using original hourly flow series of a water distribution division. It aims to identify typical daily water consumption patterns and explore factors that influence them. Findings are as follows: (1) among the three algorithms, KC demonstrates the best, with SCI of 0.6315, 0.5922, and 0.6272, and CHI of 305.9207, 274.1120, and 302.4738 for KC, AHC, and SC, respectively. (2) KC successfully identifies three distinct typical daily water consumption patterns. (3) Results indicate a significant impact of seasons on daily water consumption patterns. (4) Conversely, weekdays and holidays have minimal effect on daily water consumption patterns. It highlights the importance of comprehending daily water consumption patterns and underscores the effectiveness of KC in identifying such patterns. Furthermore, it emphasizes the significant influence of seasons while revealing limited impact of weekdays and holidays on daily water consumption patterns.

Key words: Agglomerative Hierarchical Clustering, cluster analysis, daily water consumption patterns, K-means Clustering, Spectral Clustering

HIGHLIGHTS

- K-means Clustering performs the best among the three clustering algorithms.
- Three typical daily water consumption patterns were identified in the case study.
- Season was found to be a significant influencing factor for the three patterns.

GRAPHICAL ABSTRACT



ABBREVIATIONS

KC K-means Clustering
AHC Agglomerative Hierarchical Clustering
SC Spectral Clustering
CHI Calinski–Harabasz Index
SCI Silhouette Coefficient Index

1. INTRODUCTION

In recent years, the severity of water scarcity caused by climate change has heightened the importance of the solar distillation system (United Nations Human Settlements Programme 2011; Ashok Kumar & Samsher 2021a; Ashok Kumar 2023). Coupled with the continuous growth of the population, this has resulted in a more urgent demand for water supply pumping stations. Understanding and studying water consumption patterns play a crucial role in controlling, scheduling, and optimizing the operation and management of water supply pumping stations, as it provides valuable insights into the demand patterns of water consumers. Additionally, accurate modeling of water consumption patterns aids in the design and optimization of pumping station control systems (Avni *et al.* 2015; Hussien *et al.* 2016). The study of water consumption patterns provides essential information for water resource planning and management (Dong *et al.* 2013) and allows us to predict and anticipate peak periods of water demand. This enables efficient allocation of water resources and ensures a reliable water supply to meet consumer needs. Furthermore, by identifying and analyzing the factors influencing water consumption patterns, we can develop strategies to encourage water conservation and promote sustainable water usage practices.

Understanding the temporal variations in water demand helps develop effective control and scheduling strategies for pumping stations, ensuring optimal operation and energy efficiency. This, in turn, leads to cost savings, reduced energy consumption, and improved overall system performance. In recent years, numerous studies have focused on conducting technology–environment–economy–energy matrix observations under specific conditions to seek more comprehensive solutions. These studies aim to delve deeper into understanding the environmental impacts, economic feasibility, and energy efficiency of different technologies, providing more accurate information for decision-making and planning processes (Ashok Kumar & Samsher 2020, 2021a, 2021b, 2022a, 2022b, 2022c).

As water scarcity becomes increasingly widespread, it is crucial to develop effective water management strategies that are based on an enhanced understanding of urban water demand and the factors that influence household water usage patterns. This understanding can be achieved through data analysis and Machine Learning (ML), particularly by analyzing live traffic sequences (Cominola *et al.* 2019).

Many researchers have shown an increasing interest in ML to address various problems in water resources and hydrology (Maier & Dandy 2000). The use of ML, a branch of Artificial Intelligence (AI), has supported numerous water studies by providing tools for high-level exploratory and statistical analysis of large-scale water consumption data (Garcia *et al.* 2017; Duerr *et al.* 2018; Rahim *et al.* 2021). From a data label perspective, there are three modes of learning: unsupervised, supervised, and semi-supervised (Ang *et al.* 2015; Li *et al.* 2017). Clustering algorithms are popular methods for data analysis, particularly in unsupervised learning tasks where items or objects are grouped based on inherent similarities among them (Saxena *et al.* 2017; Rahim *et al.* 2021). These algorithms can effectively group flow sequences through calculations, providing an effective method for identifying water consumption patterns. Chen demonstrated the successful application of similarity-based cluster analysis to sequence datasets using different distance measures (Chen 2007).

Clustering algorithms can be categorized into two main types: hierarchical clustering algorithms and non-hierarchical clustering algorithms (Gülagiz & Sahin 2017). K-means Clustering (KC) is a widely recognized non-hierarchical algorithm that has been extensively applied across different domains, with several researchers demonstrating its effectiveness in cluster analysis (Gnanadesikan 2011; Rahim *et al.* 2020; Mirzal 2022). On the other hand, Agglomerative Hierarchical Clustering (AHC) belongs to the hierarchical clustering category and has found significant usage in addressing urban water issues (Yu *et al.* 2013; Diao *et al.* 2014). While KC produces clusters that form convex sets, Spectral Clustering (SC) has been shown to effectively handle more complex problems, such as intertwined spirals (Ng *et al.* 2001; Nascimento & De Carvalho 2011; Ding *et al.* 2014; Saxena *et al.* 2017). However, in the context of water consumption research, SC has received less attention compared to KC and AHC. Previous studies have explored the application of these three algorithms in various areas, such as urban flood detection, early warning, and clustering performance in time series water depth difference and flood prediction (Li *et al.* 2020). While these algorithms have demonstrated good performance in various areas of data analysis, we discovered that SC, as an excellent algorithm, is rarely utilized in water consumption research. This observation greatly inspired our study.

Currently, there are numerous studies on residential water consumption patterns (Memon & Butler 2006; Wong *et al.* 2010; Browne *et al.* 2014; Yang *et al.* 2015; Garcia *et al.* 2017; Vieira *et al.* 2018). However, these studies vary in terms of data sources, materials, spatial scales, temporal scales, and clustering algorithms employed. Some scholars have employed self-mapping clustering algorithms, such as Self-Organizing Maps (SOMs), supplemented by KC and AHC, to analyze the water consumption behavior of individual users within a specific study area over a short period. The combination of KC and AHC has been shown to greatly enhance the performance of clustering algorithms. (Ioannou *et al.* 2021). Furthermore, the application of spectral algorithms in a real-life water distribution network (WDN) in South Italy demonstrated improved performance compared to graph partitioning in terms of minimizing the number of edge cuts, making it more efficient in both hydraulic and economic aspects. (Khoa Bui *et al.* 2020).

Some of these studies rely on questionnaires, and the data sources lack completeness, accuracy, and representativeness. (Beal *et al.* 2018; Cominola *et al.* 2019). Additionally, most of these studies primarily focus on water-consuming fixtures commonly found in households, such as dishwashers, toilets, and kitchen faucets, to segment users and explore different water consumption patterns (Russell & Fielding 2010; Aghabozorgi *et al.* 2015; Nguyen *et al.* 2015). Research based on real water consumption data yields more accurate results. These studies are conducted on a monthly or yearly time scale and have achieved favorable outcomes (Gato *et al.* 2007; Wang *et al.* 2009; Dey *et al.* 2012).

However, existing studies often have limitations in terms of their time scale and utilization of hourly water consumption data from a substantial number of users. Additionally, these studies incorporate multiple factors that influence water

consumption, resulting in complex and accurate models. However, the requirement for a significant amount of data to build these models can be inconvenient to handle. In contrast, this study focuses on real-time flow data, which offers advantages in terms of ease of clustering and manipulation. By employing KC, AHC, and SC, this research not only identifies typical daily water consumption patterns but also analyzes the factors that contribute to these patterns. The utilization of these clustering techniques enables a comparative analysis, providing novel insights into the understanding of water consumption behavior. The findings of this study not only contribute to the field of water consumption modeling methods but also offer valuable guidance for future research in this area. By emphasizing the significance of utilizing real-time flow data and employing different clustering techniques, this research contributes to the development of more efficient and effective approaches for analyzing water consumption patterns.

The remainder of this study is structured as follows: In Section 2, a comprehensive introduction to the clustering algorithms employed for identifying daily water consumption patterns is provided, including KC, AHC, and SC. Subsequently, the study data and the preprocessing methods applied to the data are presented. Section 3 presents a comparison of the three clustering algorithms using evaluation metrics. Finally, in Section 4, the impact of weekdays or holidays on clustering results, as well as the influence of seasons on clustering results, is discussed.

2. METHODOLOGY AND DATA

This study was organized into four steps: (1) preprocessing of daily water consumption data; (2) evaluation of clustering algorithms; (3) summary of daily water consumption patterns; (4) analysis of influencing factors. The flowchart of this study can be found in Figure 1.

2.1. Clustering algorithms

2.1.1. K-means Clustering

KC is a typical unsupervised clustering algorithm (Ahmed *et al.* 2020), which aims to divide the input sample dataset $D = \{x_1, x_2, \ldots, x_m\}$ into k clusters such that the items of the same cluster are as similar to each other as possible, while the items of different clusters are as different as possible. The algorithm runs as follows (Sinaga & Yang 2020):

(1) First, input the value of k, and then select the k initial centers of mass $\{\mu_1, \mu_2, \ldots, \mu_k\}$ from the dataset D.
(2) Assign each point in the sample set to a cluster. Then the distance between each point and the center of mass is calculated using the Euclidean distance measure. Finally, each point is assigned to the cluster that corresponds to the center of mass with the closest distance. The calculation formula of Euclidean distance is as follows:

$$D(x, \mu) = \sqrt{\sum_{i=1}^{n} (x_i - \mu_i)^2} \tag{1}$$

where x represents a sample point within the cluster, μ represents the center of mass of the cluster, n represents the number of features in each sample point, i represents individual features that constitute a point x.
(3) After assigning all the objects to their respective clusters, the cluster centers are recalculated by considering the existing objects within each cluster.
(4) Repeat step (2) and step (3) until the center of mass no longer changes, then output k cluster divisions $C = \{C_1, C_2, \ldots, C_k\}$.

Here, the sum-of-squared error (SSE) was used to determine the optimal k values, which can be defined as:

$$SSE = \sum_{i=1}^{K} \sum_{x_j \in C_i} |x_j - \mu_i|^2 \tag{2}$$

where μ_i represents the center of C_i.

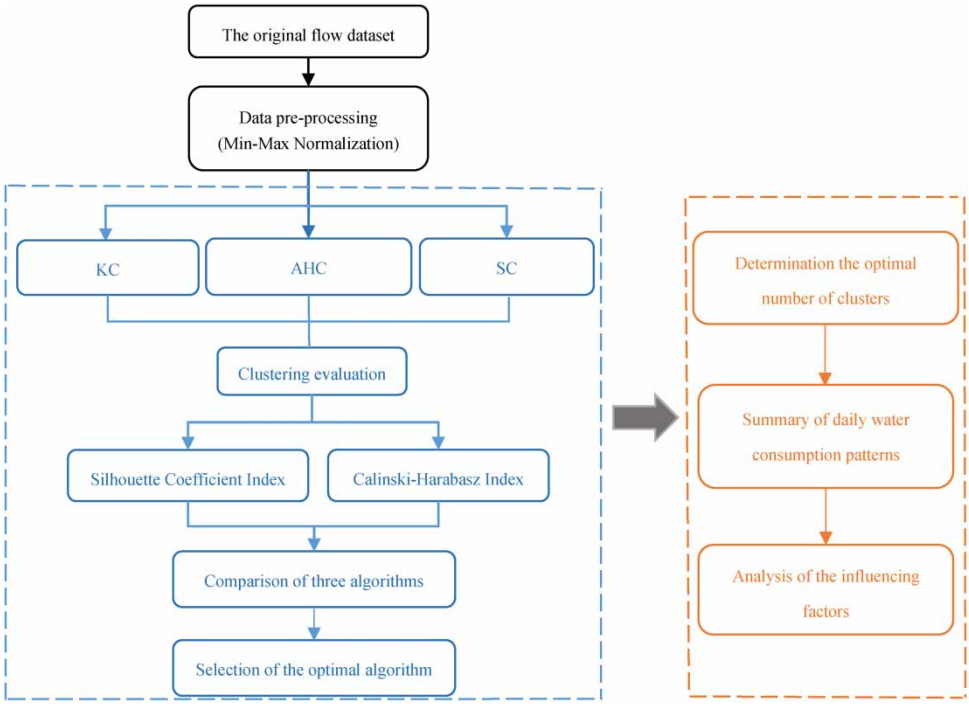


Figure 1 | Flowchart of this study.

2.1.2. Agglomerative Hierarchical Clustering

AHC is a type of hierarchical clustering, often referred to as a bottom-up method. The classical AHC algorithm relies on two essential components: a similarity or dissimilarity measure between objects and an aggregation criterion or linkage rule that governs the merging of clusters of objects (Kojadinovic 2004).

The detailed process of AHC is as follows (Day & Edelsbrunner 1984):

(1) Create n initial populations, each consisting of a single individual.
(2) Calculate the distance between each two clusters and merge them to form a new cluster.
(3) Minimizes the total within-cluster sum-of-squares between two clusters. In other words, given two clusters A and B, the value of $\Delta(A, B)$ represents the minimum sum of squares.

$$\begin{aligned}\Delta(A,B) &= \sum_{i\in A\cup B}||\vec{x_i}-\vec{m}^2_{A\cup B}|| - \sum_{i\in A}||\vec{x_i}-\vec{m}_A||^2 - \sum_{i\in B}||\vec{x_i}-\vec{m}_B||^2 \\ &= \frac{n_A n_B}{n_A+n_B}||\vec{m}_A-\vec{m}_B||\end{aligned} \tag{3}$$

where x_i represents a data point of a cluster A or cluster B, $\vec{m}_j$ represents the center of cluster j, and $\vec{n}_j$ represents the number of points in it. $\Delta(A, B)$ represents the merging cost of combining A and B.

(4) Repeat step (2) and step (3), until all clusters are combined into one.

2.1.3. Spectral Clustering

SC is a clustering method. It is based on algebraic graph theory and was proposed by Donath and Hoffman in 1973 (Donath & Hoffman 1973). In recent years, it has gained extensive attention from academia. This is due to its solid theoretical foundation and its ability to deliver good clustering performance (Jia *et al.* 2013). SC is rooted in spectral graph theory, treating the data clustering problem as a graph partitioning problem. It constructs an undirected weighted graph, where each data point in the dataset represents a vertex, and the similarity value between any two points represents the weight of the edge connecting the corresponding vertices. The SC algorithm runs as follows (Ng *et al.* 2001):

(1) Create a similarity matrix W among the observations. This is achieved by employing the fully connected method with the Radial Basis Function (RBF) for the calculations.

$$W = \begin{bmatrix} w_{11} & w_{12} & w_{13} \\ \cdots & \cdots & \cdots \\ w_{n1} & w_{n2} & w_{nn} \end{bmatrix} \tag{4}$$

$$w_{ij} = w_{ji} = e^{-\frac{1}{2}[(v_i - v_j)^T \Sigma^{-1}(v_i - v_j)]} \tag{5}$$

where $v_i^T \Sigma^{-1} v_j$ represents a scalar and the above equation represents a symmetric metric operator, W represents a symmetric matrix.

(2) For a given vertex d_i define the degree as:

$$d_i = \sum_{j=1}^{n} w_{ij} \tag{6}$$

where $i = 1, 2, 3, \ldots, n$, then create a degree matrix D as follows:

$$D = \begin{bmatrix} d_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & d_n \end{bmatrix} \tag{7}$$

(3) Calculate graph Laplacian matrix $L = D - W$, where D represents the degree matrix and W represents the similarity matrix. Clearly, since both D and W represent symmetric matrices, L represents also a symmetric matrix.

(4) The Eigen decomposition of L to obtain the smallest k eigenvalues corresponds to the eigenvectors arranged in columns to form a matrix $Q = [q_1, q_2, q_3, \ldots q_k]$, where q represents a column vector.

(5) Clustering all rows of matrix Q are clustered to obtain $[C_1, C_2, C_3, \ldots, C_k]$, output the grouping of the original data.

2.2. Evaluation indicators of clustering method performance

Clustering algorithms are typically evaluated using external and internal metrics. In this study, internal metrics, namely the Silhouette Coefficient Index (SCI) and Calinski–Harabasz Index (CHI), are employed. These metrics are unsupervised and do not depend on a benchmark dataset or an external reference model. They assess the quality of clustering results by considering the distances between sample points in the dataset and their respective cluster centers.

2.2.1. Silhouette Coefficient Index

The SCI is used to evaluate the effectiveness of clustering. It was initially proposed in 1986 (Rousseeuw 1987), where a higher SCI score relates to a model with better-defined clusters (Maulik & Bandyopadhyay 2002). The calculation formula of SCI value is as follows (Wang & Xu 2019):

$$\text{SCI} = \frac{m - n}{\max(m, n)} \tag{8}$$

where the SCI is for each individual observation; m represents the mean of dissimilarity between an observation and all other observations in the same clusters; n represents the minimum value of the mean of dissimilarity between the same observation and all observations in the next nearest cluster. Averaging the contour coefficients of all observations represents the total contour coefficient of the clustering result. Silhouette ranges from -1 to $+1$, with values closer to 1 indicating better clustering.

2.2.2. Calinski–Harabasz Index

The CHI measures the ratio of between-cluster variance to within-cluster variance. A higher CHI value indicates that the clusters are more compact internally and have greater separation between them, which signifies better clustering outcomes. The

formula for calculating the CHI value is as follows (Wang & Xu 2019):

$$\mathrm{CHI} = \frac{tr(B_k)}{tr(W_k)} \times \frac{N-k}{k-1} \tag{9}$$

where N represents number of all data; k represents number of clustering categories; $tr(B_k)$ represents between-cluster variance, $tr(W_k)$ represents within-cluster variance; B_k represents the between-cluster dispersion matrix, and W_k represents the within-cluster dispersion matrix. B_k and W_k are defined by the following equations:

$$B_k = \sum_{q=1}^{k} n_q (c_q - c_E)(c_q - c_E)^T \tag{10}$$

$$W_k = \sum_{q=1}^{k} \sum_{x \in C_q} (x - c_q)(x - c_q)^T \tag{11}$$

where C_q represents the set of all data in cluster q; c_q represents the center of cluster q; c_E represents the center of all data; n_q presents the total number of data points of cluster q.

2.3. Development environment

In this study, the clustering analysis was performed using Python programming language with the support of the PyCharm development environment. Python provided a versatile and widely-used platform for data analysis and ML, while PyCharm offered a user-friendly interface for coding and executing the clustering algorithm. These tools allowed for effective implementation and analysis, ensuring reliable and accurate results for our study.

2.4. Study data

For the case study, we utilized an hourly water consumption flow series collected from a water supply division in Shanghai, covering the period from January 1, 2020, to December 31, 2020, as illustrated in Figure 2. Thorough quality checks were conducted to ensure the reliability and validity of the data, which had a temporal resolution of 1 h, enabling a detailed analysis of consumption patterns and trends. Importantly, this study period coincided with the COVID-19 pandemic, which had a significant impact on societal behaviors and consumption patterns. The first confirmed case of COVID-19 in our study area was reported on March 15, 2020, followed by the implementation of strict lockdown measures from March 20 to April 30, 2020, aimed at limiting the spread of the virus. It is plausible that these measures potentially influenced water consumption patterns.

To gain insights into the composition of water consumption, we collected data on the proportion of residential and industrial demand through surveys and collaboration with local water supply authorities. Our findings revealed that residential demand accounted for approximately 70% of the total water consumption, while the remaining 30% was attributed to industrial demand. These proportions were taken into account during the clustering analysis to gain a better understanding of the different consumption patterns.

2.5. Data preprocessing

Standardization is a key step in data preprocessing (Armstrong & Van de Wiel 2004). In this study, we utilized the min-max normalization method to preprocess the research data, transforming the raw data into the range of [0, 1] as shown in Figure 3. The purpose of this normalization method is to eliminate the adverse effects caused by individual sample data. By scaling the data to the same range, it ensures comparability among different features and enhances the effectiveness of subsequent data processing and analysis. The min-max normalization achieves this by calculating the ratio between each sample value and the minimum and maximum values, resulting in a linear transformation of the data. Such standardization methods are commonly used in ML and data mining tasks to ensure data consistency and comparability. The Min-Max formula is as follows (Kiran & Vasumathi 2020):

$$y = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{12}$$

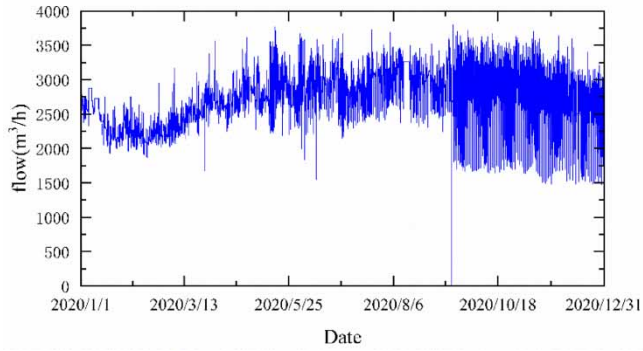


Figure 2 | The original hourly flow series.

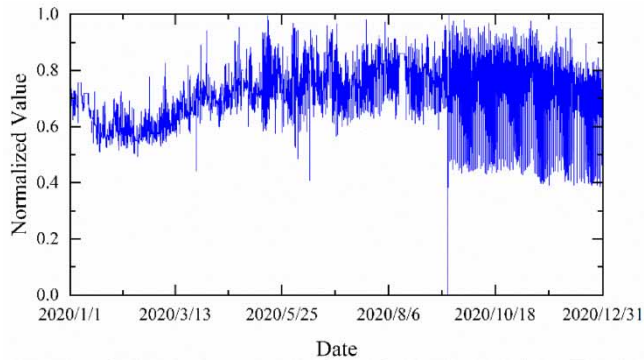


Figure 3 | Normalized water consumption flow series.

where x represents a sample data point in row dataset X, min (x) and max (x) represent the minimum and maximum values in X, y represents the normalized data point.

3. RESULTS

3.1. Results of clustering performance evaluation

Initially, the preprocessed data, which represents daily data with 24 time steps, undergoes clustering using three distinct learning algorithms. The resulting clustering outcomes for clusters 2–5 in each algorithm are illustrated in Figures 4–6, respectively. In these figures, points of the same color represent the same cluster.

According to the study data after preprocessing, the performance of the clustering algorithm is studied, and Table 1 and Figure 7 show the SCI and CHI values corresponding to the different cluster numbers of three different algorithms. It is important to note that a larger SCI and CHI indicate better clustering quality. In this study, we observed that the SCI and CHI values for all three algorithms reach their maximum when the number of clusters is set to 3. However, it is particularly noteworthy that the KC algorithm exhibits significantly higher SCI and CHI values compared to the other two algorithms. These findings suggest that the KC algorithm outperforms the other algorithms in terms of clustering evaluation. Specifically, its higher SCI and CHI values indicate its superior ability to generate well-defined and distinct clusters, which is crucial for effective data segmentation and pattern recognition.

Figure 7 shows that the optimal solution is three clusters, as indicated by the highest SCI value and CHI value. This study conclusively determines three as the ideal number of clusters, and this number was employed in the subsequent analysis.

3.2. Summarization of daily water consumption patterns

Based on the results from the previous section, it can be concluded that the optimal number of clusters for the KC algorithm in this study is three. Next, the water use data is specifically analyzed using the KC algorithm, resulting in three clusters. The

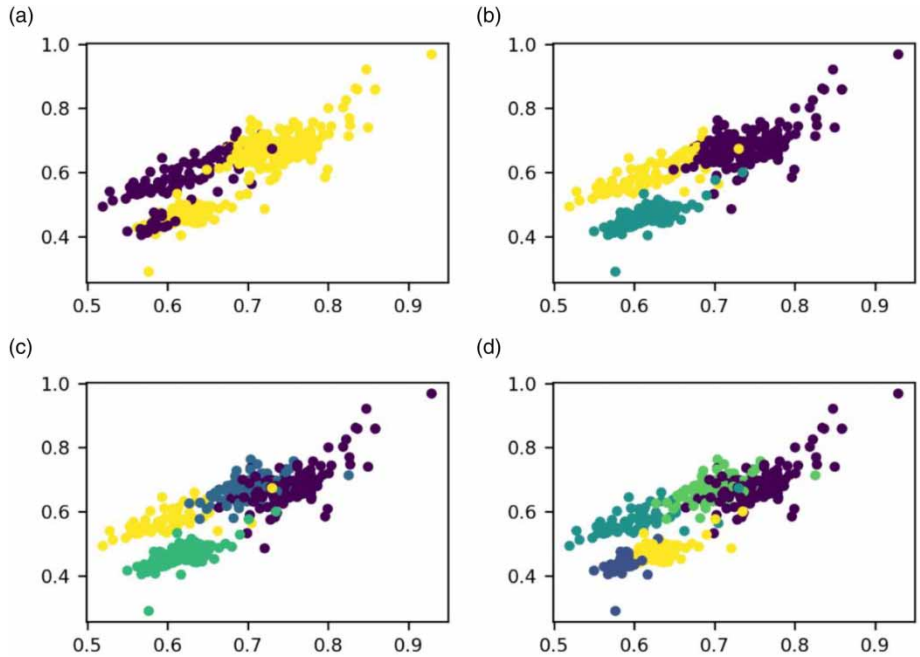


Figure 4 | The clustering results of the KC algorithm. (a) Cluster = 2, (b) Cluster = 3, (c) Cluster = 4, and (d) Cluster = 5.

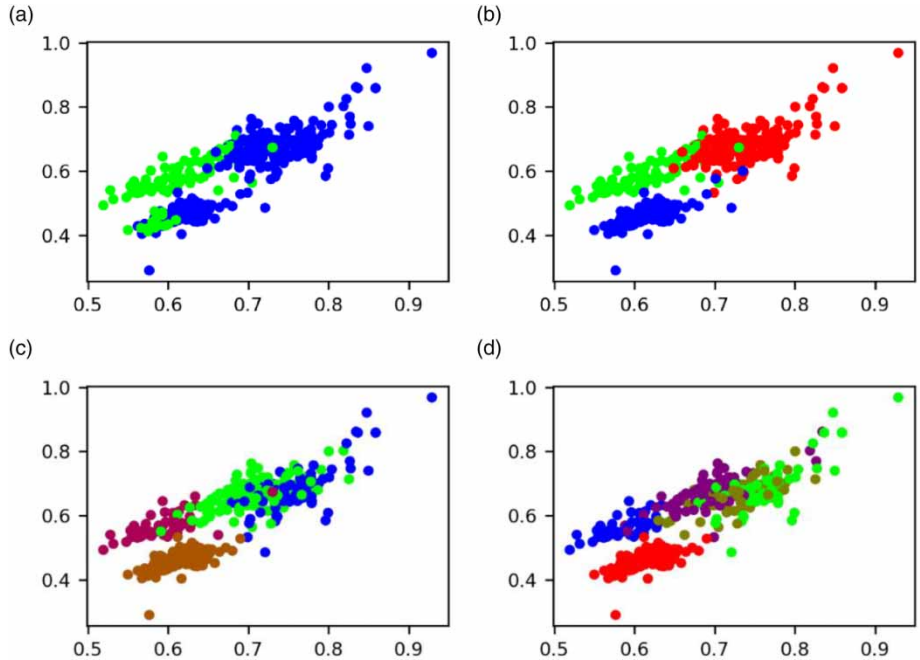


Figure 5 | The clustering results of the SC algorithm. (a) Cluster = 2 (b) Cluster = 3, (c) Cluster = 4 and (d) Cluster = 5.

dates corresponding to each cluster were obtained from the colored points in Figure 4. Figure 8 displays the 24-h water consumption curves for each cluster, representing the daily water consumption for all the dates included in each cluster.

Figure 9 illustrates the total count of days in each cluster. In 2020, Shanghai had a total of 112 holidays and 254 weekdays, excluding days involving work due to transfers. The average flow rates for each date within each cluster were calculated, resulting in the new daily water consumption curves shown in Figure 10.

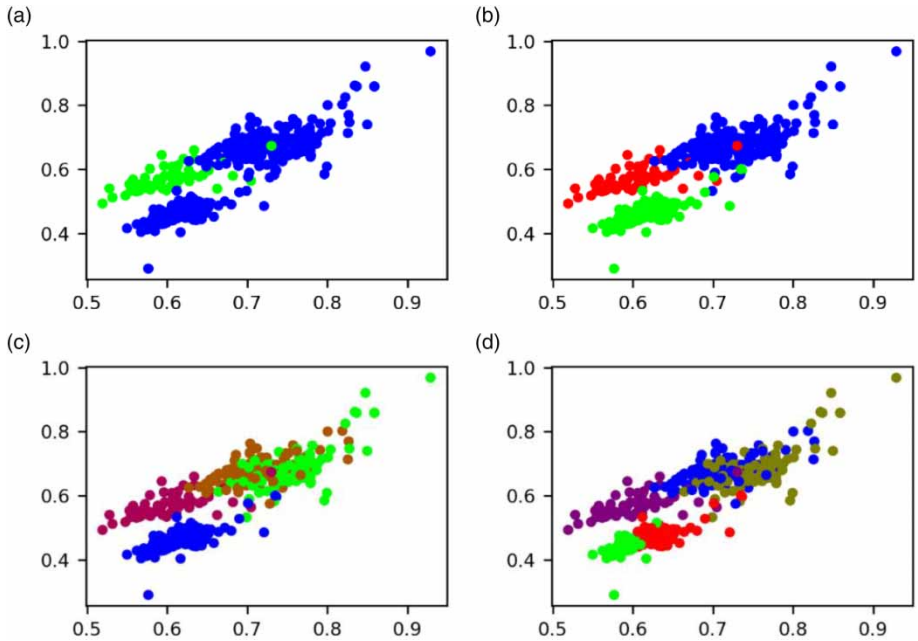


Figure 6 | The clustering results of the AHC algorithm. (a) Cluster = 2 (b) Cluster = 3, (c) Cluster = 4 and (d) Cluster = 5.

Table 1 | Evaluation indicators of the three algorithms

Clustering algorithm	Cluster numbers	SCI	CHI
K-means Clustering	2	0.5161	207.3337
	3	0.6315	305.9207
	4	0.5640	287.8707
	5	0.5125	250.1610
Agglomerative Hierarchical Clustering	2	0.5127	176.7943
	3	0.5922	274.1120
	4	0.5271	264.5629
	5	0.4774	230.7360
Spectral Clustering	2	0.5215	203.3216
	3	0.6272	302.4738
	4	0.5347	263.4528
	5	0.4341	193.4173

The green line in Figure 10 shows the daily water consumption curve for Cluster 1. It exhibits a relatively small fluctuation range but demonstrates a noticeable overall trend with an initial decrease, followed by an increase, and then another decrease. The peak occurs at 14:00 pm, while the lowest point is reached at 2:00 am. Cluster 1 consists of 29 holidays and 57 weekdays, with 32 days falling in spring and 52 days in winter.

The deep blue line in Figure 10 displays the daily water consumption curve for Cluster 2. This cluster shows a highly regular pattern with a distinctive double-hump shape. There are two peak periods at 10:00 am and 18:00 pm. The curve gradually declines and reaches the first trough at 4:00 am, followed by the second trough at 15:00 pm. Cluster 2 includes 33 holidays and 71 weekdays, with 75 days in autumn and 29 days each in winter and spring.

The light blue line in Figure 10 presents the daily water consumption variation curve for Cluster 3. The pattern in this cluster resembles that of Category 2 between 0:00 am and 8:00 am. However, from 8:00 am to 20:00 pm, the flow remains relatively flat and constant. The water usage reaches its maximum at 21:00 pm and then sharply declines from 21:00 pm to 23:00 pm. Cluster 3 encompasses 52 holidays and 124 weekdays, covering all four seasons, with 61 dates in spring, 91 dates in summer, 15 dates in autumn, and 9 dates in winter, as shown in Table 2.

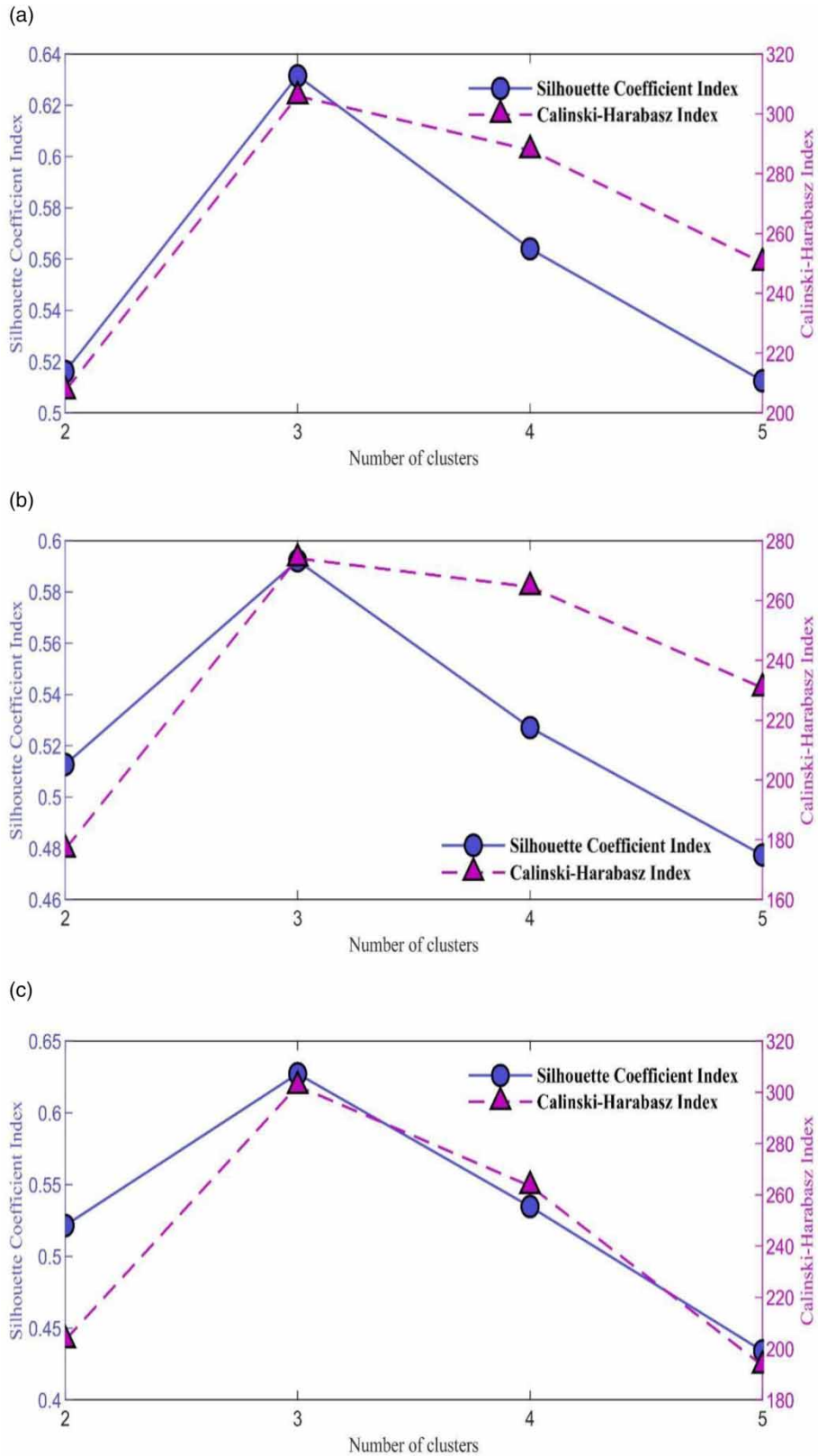


Figure 7 | Results of the cluster index evaluation. (a) The results of KC, (b) the results of agglomerative hierarchical clustering, and (c) the results of SC.

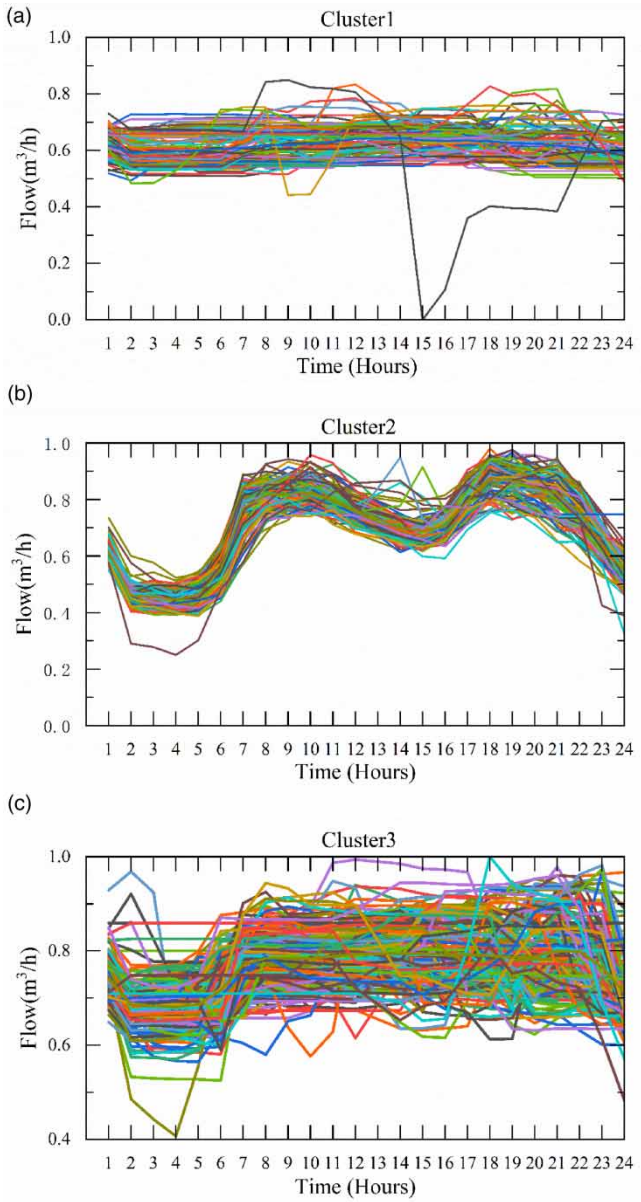


Figure 8 | Daily water consumption flow curves for three clusters.

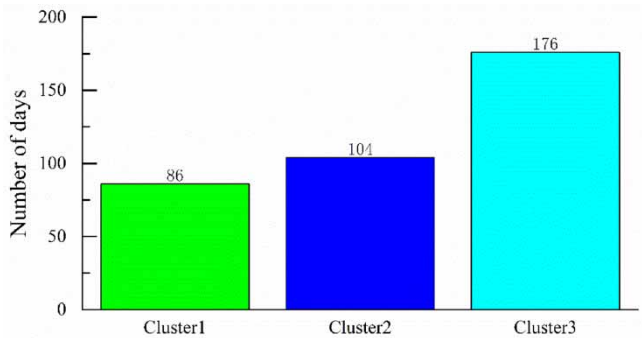


Figure 9 | Number of days for three clusters.

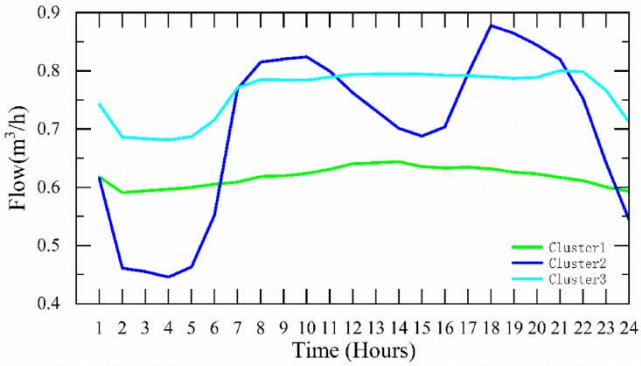


Figure 10 | Daily average water consumption curves for three clusters.

Table 2 | Distribution of six types of days for three clusters

Clusters	Holidays	Weekdays	Spring	Summer	Autumn	Winter
Cluster 1	29	57	32	1	1	52
Cluster 2	33	71	0	0	75	29
Cluster 3	52	124	61	91	15	9

4. DISCUSSIONS

4.1. Impact of weekdays or holidays on clustering results

Daily water consumption is influenced by users' daily activity habits. Speculations about differences in water consumption patterns between weekdays and weekends, as well as holidays and weekdays, are subjective and based on individual perceptions. In this study, we analyze real-time data considering additional factors associated with weekdays and holidays.

Figure 11 summarizes the distribution of dates for the three clusters. The distribution of holidays and weekdays within each cluster does not show significant distinctions, with the ratio of holidays to weekdays fluctuating around 1:2. Surprisingly, the impact of changes in daily activities during holidays on water consumption is not as pronounced as expected. In fact, there are no significant differences observed between holidays and weekdays in terms of water consumption proportions (Leitão *et al.* 2019).

Another possibility to consider is that the suburban area, where the data is sourced, may not be representative of the main urban area. The demographics of this area are predominantly agricultural, and daily activities may not significantly change during holidays. Additionally, the COVID-19 pandemic in 2020 had a major worldwide impact on water consumption

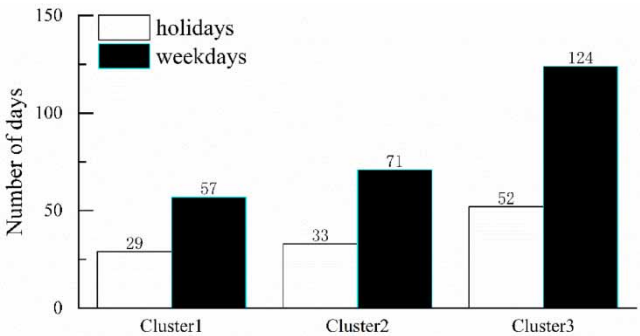


Figure 11 | Distribution of the number of weekdays and holidays for three clusters.

patterns (Dzimińska *et al.* 2021). The distinction between weekdays and holidays became less evident, especially with the rise of telecommuting and distance learning, where the 'stay at home' lifestyle became the norm. Overall, our findings suggest that there are no significant differences in water consumption between holidays and weekdays. The suburban context and the influence of the COVID-19 pandemic contribute to these observed patterns.

4.2. Impact of seasons on clustering results

Figure 12 illustrates the number of seasons corresponding to the dates included in the three clusters. In this article, we define March, April, and May as spring, June, July, and August as summer, September, October, and November as autumn, and December, January, and February as winter. It is evident that Cluster 1 comprises mainly spring and winter, summer and autumn can be ignored; Cluster 2 consists of autumn and winter; while Cluster 3 encompasses all four seasons. It is noteworthy that among the four seasons, only the summer season exclusively falls within a single category. The average daily water consumption curves for each season are depicted in Figure 13.

Cluster 1, which we have labeled as Pattern A, primarily occurs during the winter and early spring seasons, with minimal representation during summer and fall. Our analysis, as shown in Figure 10 reveals several distinctive characteristics of this pattern when compared to the other two. First, unlike the other patterns, Pattern A does not exhibit a prominent morning peak in water flow. Instead, the daily flow curve demonstrates significant rises and falls throughout the day. Notably, an afternoon peak is observed between 10:00 and 14:00, followed by a slight rebound at 16:00, forming a smaller peak. Consequently, this irregular behavior sets Pattern A apart from the other patterns. Furthermore, it is important to acknowledge the influence of the prevailing COVID-19 pandemic during the months predominantly associated with Pattern A, namely January, February, and March. This pandemic significantly impacts the morning, afternoon, and evening peaks of daily water consumption within this pattern. The altered routines and increased time spent at home during lockdown measures contribute to these notable changes in water usage patterns. However, despite the significant impact of the COVID-19 pandemic, the fluctuations in water consumption during winter and early spring – periods that hold particular significance – are relatively negligible

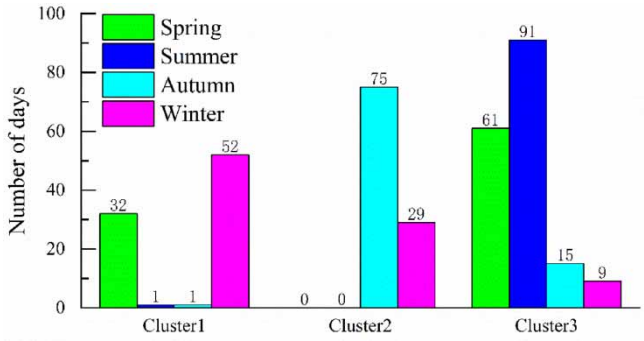


Figure 12 | Distribution of the number of days in four seasons for three clusters.

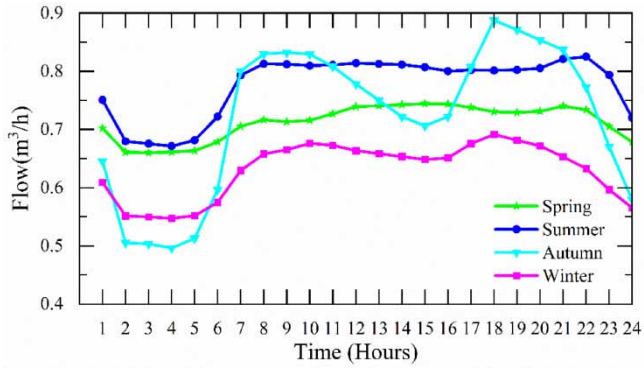


Figure 13 | Average daily water consumption curves for four seasons.

within Pattern A. Further investigation into the underlying factors contributing to these negligible fluctuations during this time frame remains an area of future research and warrants additional attention.

Cluster 2 consists of autumn and winter, with spring and summer at 0. We refer to this as Pattern B.

Cluster 3 represents a daily water usage pattern dominated by summer, which we call Pattern C. Figure 10 shows the corresponding daily average water usage variation curves, while Figure 13 displays the curves for summer. Comparing these two curves, we confirm their high similarity. From 0:00 to 7:00, there is no noticeable difference in the water consumption curves between Pattern C and Pattern B. Between 7:00 and 20:00, Pattern C shows minimal changes, maintaining consistently high levels without significant drops or valleys. This can be attributed to the hot weather in summer and the unusually long summer experienced in Shanghai in 2020, resulting in high temperatures and drought that may affect crop irrigation. Pattern C experiences a rise in water usage from 19:00 to 21:00, peaking at 21:00, likely due to residents using water for evening bathing, creating a small peak. However, unlike Pattern C, the morning peak for the summer mode occurs at 7:00, with Pattern C experiencing a 1-hour delay. Additionally, the afternoon peak for Pattern C happens between 12:00 and 14:00, peaking at 14:00, which aligns with the spring pattern. Moreover, considering the dates, more than half of the spring falls into Pattern C, further suggesting that spring may have influenced the midday peak.

Integrating daily consumption patterns into the optimization of WDNs presents significant advantages, as evidenced by the analysis of different patterns in this study. Notably, Cluster 1, representing winter and early spring as Pattern A, exhibits distinctive characteristics. This pattern lacks a morning peak and displays significant fluctuations in the daily flow curve, rendering it the most irregular pattern. Understanding and recognizing such patterns enables the scheduling algorithm to adapt and allocate resources accordingly, ensuring a stable water supply despite irregular consumption behaviors. Similarly, Cluster 3, dominated by summer as Pattern C, demonstrates consistently high levels of water usage throughout the day. By identifying and acknowledging this pattern, the algorithm can optimize supply schedules to accommodate the specific demands of summer, such as the evening bathing peak at 21:00. This targeted approach enhances operational efficiency and ensures adequate water distribution during critical periods. These observations highlight the novelty and contribution of our research in integrating daily consumption patterns into WDN optimization. By considering and analyzing these patterns, our study provides valuable insights for improving the operational efficiency of water distribution systems. This approach enables better resource allocation, resulting in optimized water supply schedules that align with the specific demands of different seasons and consumption patterns.

5. CONCLUSIONS

Table 3 presents a comprehensive summary of the research findings obtained in this study.

This study utilized three clustering algorithms (K-means, agglomerative hierarchical, and SC) to analyze daily water consumption patterns. The performance of these algorithms was evaluated using the SCI and the CHI. The conclusions of the study are as follows:

Table 3 | Summary of research findings

Research method	Innovation
Cluster method selection	The study utilized a diverse range of clustering methods, encompassing various principles and applicabilities. This facilitates the comparison of different methods in analyzing daily water usage patterns, providing a more comprehensive evaluation and insights.
Cluster analysis	The selected clustering method is applied to the preprocessed daily water usage data in this study, dividing the data samples into distinct clusters. Proper distance metrics and clustering evaluation metrics are employed to ensure the quality assessment of the clustering analysis results.
Pattern variations and feature interpretation	By comparing the results of different clustering methods, this study revealed the explanatory capacity and differences in water usage patterns among the methods. By providing detailed explanations of the patterns identified by each clustering method and their features, a more comprehensive understanding of water consumption patterns is achieved, aiding in identifying the applicability of different clustering methods in specific scenarios.
Practical application validation	This study applied the experimental method to real-world daily water usage data. By validating the effectiveness of different clustering methods in practical applications, it provides practical value and guidance for decision-makers in the field.

(1) KC outperformed AHC and SC, as indicated by higher SCI and CHI values (0.6315 and 305.9207, respectively).
(2) The data were clustered into three patterns: Pattern A, Pattern B, and Pattern C. These patterns have similar proportions of weekdays and holidays. Pattern A is dominated by winter and spring, Pattern B by autumn, and Pattern C by summer and spring. Pattern B and Pattern C exhibit similar variations from 0:00 to 7:00, while Pattern A differs during this time period. The main distinction among the patterns lies in the water consumption variation between 8:00 and 21:00.
(3) Seasons significantly influence daily water consumption patterns. In spring, the midday peak is delayed, likely due to the impact of the COVID-19 pandemic in 2020. Summer primarily affects water consumption changes between 7:00 and 20:00, influenced by high temperatures and drought conditions. The area's citrus cultivation and continuous need for irrigation water contribute to minimal variation in daily water consumption during this period, resulting in a delay in the evening peak. Autumn and winter lack a midday peak, coinciding with the citrus harvest season and the agricultural activities prevalent in the study area.
(4) The number of weekdays and holidays shows no significant effect on the three patterns, as the proportions remain proportional across all patterns.

Future scope can include investigating the underlying factors causing the delay in the midday peak during spring and examining the specific impacts of citrus cultivation on water consumption patterns in autumn and winter. Additionally, exploring the influence of external factors such as economic activities or cultural events on daily water consumption patterns can provide valuable insights.

ACKNOWLEDGEMENTS

We are very grateful to the editors and anonymous reviewers for their insightful suggestions and comments on this paper.

DATA AVAILABILITY STATEMENT

Data cannot be made publicly available; readers should contact the corresponding author for details.

CONFLICT OF INTEREST

The authors declare there is no conflict.

REFERENCES

Aghabozorgi, S., Shirkhorshidi, A. S. & Wah, T. Y. 2015 Time-series clustering–a decade review. *Information Systems* **53**, 16–38.

Ahmed, M., Seraj, R. & Islam, S. M. S. 2020 The k-means algorithm: A comprehensive survey and performance evaluation. *Electronics* **9** (8), 1295.

Ang, J. C., Mirzal, A., Haron, H. & Hamed, H. N. A. 2015 Supervised, unsupervised, and semi-supervised feature selection: A review on gene selection. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **13** (5), 971–989.

Armstrong, N. J. & Van de Wiel, M. A. 2004 Microarray data analysis: From hypotheses to conclusions using gene expression data. *Analytical Cellular Pathology* **26** (5–6), 279–290.

Ashok Kumar, S. 2023 Mathematical analysis of optimized requisites for novel combination of solar distillers. *Journal of Engineering Research* **11** (4), 515–525.

Ashok Kumar, S. & Samsher. 2020 Analytical study of evacuated annulus tube collector assisted solar desaltification system: a review. *Solar Energy* **207**, 1404–1426.

Ashok Kumar, S. & Samsher, 2021a Material conscious energy matrix and enviro-economic analysis of passive ETC solar still. *Materials Today: Proceedings* **38**, 1–5.

Ashok Kumar, S. & Samsher, 2021b An inclusive study on new conceptual designs of passive solar desalting systems. *Heliyon* **7** (2), e05793.

Ashok Kumar, S. & Samsher, 2022a Tech-en-econ-energy-exergy-matrix (T4EM) observations of evacuated solar tube collector augmented solar desaltification unit: A modified design loom. *Materials Today: Proceedings* **61**, 258–263.

Ashok Kumar, S. & Samsher, 2022b Optimum techno-eco performance requisites for vacuum annulus tube collector–assisted double-slope solar desaltification unit integrated modified parabolic concentrator. *Environmental Science and Pollution Research* **29** (23), 34379–34405.

Ashok Kumar, S. & Samsher, 2022c Techno-environ-economic-energy-exergy-matrices performance analysis of evacuated annulus tube with modified parabolic concentrator assisted single slope solar desalination system. *Journal of Cleaner Production* **332**, 129996.

Avni, N., Fishbain, B. & Shamir, U. 2015 Water consumption patterns as a basis for water demand modeling. *Water Resources Research* **51** (10), 8165–8181.

Beal, C. D., Jackson, M., Stewart, R. A., Rayment, C. & Miller, A. 2018 Identifying and understanding the drivers of high water consumption in remote Australian Aboriginal and Torres Strait Island communities. *Journal of Cleaner Production* **172**, 2425–2434.

Browne, A. L., Pullinger, M., Medd, W. & Anderson, B. 2014 Patterns of practice: A reflection on the development of quantitative/mixed methodologies capturing everyday life related to water consumption in the UK. *International Journal of Social Research Methodology* **17** (1), 27–43.

Chen, J. 2007 Useful clustering outcomes from meaningful time series clustering. *AusDM* **7**, 101–109.

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** (11), 9315–9333.

Day, W. H. E. & Edelsbrunner, H. 1984 Efficient algorithms for agglomerative hierarchical clustering methods. *Journal of Classification* **1** (1), 7–24.

Dey, S., Maity, D. & Chakraborti, D. 2012 Water consumption patterns and factors contributing to water consumption in arsenic affected population of rural West Bengal, India.

Diao, K., Farmani, R., Fu, G., Astaraie-Imani, M., Ward, S. & Butler, D. 2014 Clustering analysis of water distribution systems: Identifying critical components and community impacts. *Water Science and Technology* **70** (11), 1764–1773.

Ding, S., Jia, H., Zhang, L. & Jin, F. 2014 Research of semi-supervised spectral clustering algorithm based on pairwise constraints. *Neural Computing and Applications* **24** (1), 211–219.

Donath, W. E. & Hoffman, A. J. 1973 Lower bounds for the partitioning of graphs. *IBM Journal of Research and Development* **17** (5), 420–425.

Dong, C., Schoups, G. & van de Giesen, N. 2013 Scenario development for water resource planning and management: A review. *Technological Forecasting and Social Change* **80** (4), 749–761.

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.

Dzimińska, P., Drzewiecki, S., Ruman, M., Kosek, K., Mikołajewski, K. & Licznar, P. 2021 The use of cluster analysis to evaluate the impact of COVID-19 pandemic on daily water demand patterns. *Sustainability* **13** (11), 5772.

Garcia, D., Puig, V., Quevedo, J. & Cugueró, M. A. 2017 *Big Data Analytics and Knowledge Discovery Applied to Automatic Meter Readers Real-Time Monitoring and Operational Control of Drinking-Water Systems*. Springer, Cham, pp. 401–423.

Gato, S., Jayasuriya, N. & Roberts, P. 2007 Temperature and rainfall thresholds for base use urban water demand modelling. *Journal of Hydrology* **337** (3–4), 364–376.

Gnanadesikan, R. 2011 *Methods for Statistical Data Analysis of Multivariate Observations*. John Wiley & Sons, Hoboken.

Gülagiz, F. K. & Sahin, S. 2017 Comparison of hierarchical and non-hierarchical clustering algorithms. *International Journal of Computer Engineering and Information Technology* **9** (1), 6.

Hussien, W. E. A., Memon, F. A. & Savic, D. A. 2016 Assessing and modelling the influence of household characteristics on per capita water consumption. *Water Resources Management* **30**, 2931–2955.

Ioannou, A. E., Creaco, E. F. & Laspidou, C. S. 2021 Exploring the effectiveness of clustering algorithms for capturing water consumption behavior at household level. *Sustainability* **13** (5), 2603.

Jia, H., Ding, S., Xu, X. & Nie, R. 2013 The latest research progress on spectral clustering. *Neural Computing and Applications* **24** (7–8), 1477–1486.

Khoa Bui, X., Marlim, M. S. & Kang, D. 2020 Water network partitioning into district metered areas: A state-of-the-art review. *Water* **12** (4), 1002.

Kiran, A., Vasumathi, D., 2020 Data Mining: Min–Max Normalization Based Data Perturbation Technique for Privacy Preservation. In: *Proceedings of the Third International Conference on Computational Intelligence and Informatics* (Raju, K. S., Govardhan, A., Rani, B. P., Sridevi, R. & Murty, M. R. eds). Springer Singapores, Singapore, pp. 723–734.

Kojadinovic, I. 2004 Agglomerative hierarchical clustering of continuous variables based on mutual information. *Computational Statistics & Data Analysis* **46** (2), 269–294.

Leitão, J., Simões, N., Sá Marques, J. A., Gil, P., Ribeiro, B. & Cardoso, A. 2019 Detecting urban water consumption patterns: A time-series clustering approach. *Water Supply* **19** (8), 2323–2329.

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J. & Liu, H. 2017 Feature selection: A data perspective. *ACM Computing Surveys (CSUR)* **50** (6), 1–45.

Li, J., Hassan, D., Brewer, S. & Sitzenfrei, R. 2020 Is clustering time-series water depth useful? An exploratory study for flooding detection in urban drainage systems. *Water* **12** (9), 2433.

Maier, H. R. & Dandy, G. C. 2000 Neural networks for the prediction and forecasting of water resources variables: A review of modelling issues and applications. *Environmental Modelling & Software* **15** (1), 101–124.

Maulik, U. & Bandyopadhyay, S. 2002 Performance evaluation of some clustering algorithms and validity indices. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **24** (12), 1650–1654.

Memon, F. A. & Butler, D. 2006 Water consumption trends and demand forecasting techniques. *Water Demand Management* **2006**, 1–26.

Mirzal, A. 2022 Statistical analysis of microarray data clustering using NMF, spectral clustering, kmeans, and GMM. *IEEE/ACM Trans Comput Biol Bioinform* **19** (2), 1173–1192.

Nascimento, M. C. & De Carvalho, A. C. 2011 Spectral methods for graph clustering–a survey. *European Journal of Operational Research* **211** (2), 221–231.

Ng, A., Jordan, M. & Weiss, Y. 2001 On spectral clustering: Analysis and an algorithm. In: *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*, Vancouver, 3–8 December 2001. MIT Press, Cambridge, MA, pp. 849–856.

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.

Programme, U. N. H. S. 2011 *Cities and Climate change: Global Report on Human Settlements, 2011*. Routledge, London.

Rahim, M. S., Nguyen, K. A., Stewart, R. A., Giurco, D. & Blumenstein, M. 2020 Machine learning and data analytic techniques in digital water metering: A review. *Water* **12** (1), 294.

Rahim, M. S., Nguyen, K. A., Stewart, R. A., Ahmed, T., Giurco, D. & Blumenstein, M. 2021 A clustering solution for analyzing residential water consumption patterns. *Knowledge-Based Systems* **233**, 107522.

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** (5), W05302.

Saxena, A., Prasad, M., Gupta, A., Bharill, N., Patel, O. P., Tiwari, A., Er, M. J., Ding, W. & Lin, C.-T. 2017 A review of clustering techniques and developments. *Neurocomputing* **267**, 664–681.

Sinaga, K. P. & Yang, M. S. 2020 Unsupervised K-means clustering algorithm. *IEEE Access* **8**, 80716–80727.

Vieira, P., Jorge, C. & Covas, D. 2018 Efficiency assessment of household water use. *Urban Water Journal* **15** (5), 407–417.

Wang, X. & Xu, Y. 2019 An improved index for clustering validation based on Silhouette index and Calinski-Harabasz index. *IOP Conference Series: Materials Science and Engineering* **569** (5), 052024.

Wang, Z., Xu, J., Lu, F. & Zhang, Y. 2009 Using the method combining PCA with BP neural network to predict water demand for urban development. In *2009 Fifth International Conference on Natural Computation, vol 2*. IEEE, pp. 621–625.

Wong, J. S., Zhang, Q. & Chen, Y. D. 2010 Statistical modeling of daily urban water consumption in Hong Kong: Trend, changing patterns, and forecast. *Water Resources Research* **46** (3), W03536.

Yang, J., Li, Y., Zhang, N. F., Yang, J. F., Kuang, K., Hu, Y. H. & Qi, W. G. 2015 Analysis of urban residential water consumption based on smart meters and fuzzy clustering. In: *2015 IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing*. IEEE, pp. 1295–1301.

Yu, Y., Kojima, K., An, K. & Furumai, H. 2013 Cluster analysis for characterization of rainfalls and CSO behaviours in an urban drainage area of Tokyo. *Water Science and Technology* **68** (3), 544–551.

First received 12 November 2023; accepted in revised form 14 April 2024. Available online 27 April 2024