technologies

MDPI

Article

Temporal Segmentation of Urban Water Consumption Patterns Based on Non-Parametric Density Clustering

Aliaksey A. Kapanski [1], Roman V. Klyuev [2], Vladimir S. Brigida [3,*] and Nadezeya V. Hruntovich [1]

[1] Department of Power Supply, Sukhoi State Technical University of Gomel, 246746 Gomel, Belarus
[2] Department of Automation and Control, Moscow Polytechnic University, 107023 Moscow, Russia
[3] Biomedical, Veterinary and Ecological Department, RUDN University, 117198 Moscow, Russia
* Correspondence: 1z011@inbox.ru

Abstract

The management of modern water supply systems requires a detailed analysis of consumption patterns in order to optimize pump operation schedules, reduce energy costs, and support the development of intelligent management systems. Traditional clustering algorithms are applied for these tasks; however, their limitation lies in the need to predefine the number of clusters. The aim of this study was to develop and validate a non-parametric method for clustering daily water consumption profiles based on a modified DBSCAN algorithm. The proposed approach includes the automatic optimization of neighborhood radius and the minimum number of points required to form a cluster. The input data consisted of half-hourly water supply and electricity consumption values for the water supply system of Gomel (Republic of Belarus), supplemented with the time-of-day factor. As a result of the multidimensional clustering, two stable regimes were identified: a high-demand regime (6:30–22:30), covering about 46% of the data and accounting for more than half of the total water supply and electricity consumption, and a low-demand regime (0:30–6:00), representing about 21% of the data and forming around 15% of the resources. The remaining regimes reflect transitional states in morning and evening periods. The obtained results make it possible to define the temporal boundaries of the regimes and to use them for data labeling in the development of predictive water consumption models.

Keywords: water consumption; energy consumption; intra-daily patterns; non-parametric clustering; DBSCAN; temporal labeling; energy-efficient management

Academic Editor: Miklas Scholz

Received: 30 August 2025
Revised: 28 September 2025
Accepted: 30 September 2025
Published: 3 October 2025

Citation: Kapanski, A.A.; Klyuev, R.V.; Brigida, V.S.; Hruntovich, N.V. Temporal Segmentation of Urban Water Consumption Patterns Based on Non-Parametric Density Clustering. *Technologies* **2025**, *13*, 449. https://doi.org/10.3390/technologies13100449

1. Introduction

In recent years, the capabilities for collecting information on water consumption patterns in urban water supply systems have significantly expanded. Modern technologies make it possible to form detailed databases that can be aggregated to the required level and integrated with other data sources. This creates the conditions for identifying the factors that determine water demand. Such factors may be external (climate, socio-economic conditions, demography) or internal (seasonal and daily fluctuations reflected in time series). Numerous studies have demonstrated the impact of temperature on water consumption in both agriculture and domestic use [1–3]. Socio-economic factors such as building density and population income levels also shape the basis of demand and influence the operating modes of water intake facilities [4–6]. Thus, the identification and consideration of factors affecting water consumption is a relevant task for forecasting and managing water supply systems.

The application of machine learning methods in resource management makes it possible to use such factors for planning the operating modes of water utilities. Forecasting enables the optimization of pump operation schedules and the management of electric load profiles, thereby reducing electricity costs through the use of storage reservoirs [7–9]. However, many input factors are not available in real time and can only be analyzed retrospectively, which limits their use in operational forecasting. One of the ways to generate operational features is the decomposition of time series into seasonal and calendar components. In the authors' previous research (for example, in the article «Identification of Easily Accessible Urban Water Consumption Factors for Energy-Efficient Management of Pumping Stations» [10]), the statistical significance of including such parameters as months and days of the week in models was demonstrated. In the present study, the focus is placed on investigating intra-daily variability in urban water consumption data, using one of the largest water intakes in the Republic of Belarus as a case study.

The aim of the study is to develop and justify the effectiveness of a non-parametric method for clustering daily water consumption profiles based on a modified DBSCAN (Density-based spatial clustering of applications with noise) algorithm. Unlike the classical approach, the method proposed in this article accounts for the uneven density of data across different hours of the day and allows noise points to be reassigned to neighboring clusters without specifying their number in advance. In practical application, the results make it possible to determine the temporal boundaries of high and low water demand, which can then be used for time-based labeling of data when training artificial intelligence models. The training of predictive models, however, is not considered within the scope of this article. The objectives of the work include: (1) analyzing modern methods of time series clustering and their applications in water supply systems; (2) assessing the advantages and limitations of existing approaches; (3) developing the modified DBSCAN algorithm and describing the procedure for preprocessing daily profiles; and (4) testing the method on real hourly water consumption data and interpreting the resulting demand patterns.

2. Related Work

Clustering methods for time series are widely applied in water consumption analysis tasks. The most common approaches remain classical partitioning algorithms and hierarchical methods. Algorithms such as K-means, K-medoids, and their modifications have become widespread due to their computational simplicity and interpretability [11,12]. However, their main limitation is the need to specify the number of clusters in advance. Hierarchical methods make it possible to analyze the structure of data at different levels, but they are highly sensitive to noise and require the selection of cut-off thresholds. Despite these limitations, both classes of algorithms are actively used in studies of water systems.

For example, Prakaisak and Wongchaisuwat [13] employed agglomerative clustering with preliminary extraction of statistical and spectral features, combined with the UMAP (Uniform Manifold Approximation and Projection) algorithm for dimensionality reduction. In the work of Guo et al. [14], three algorithms—K-means, hierarchical clustering, and spectral clustering–were compared for analyzing daily water consumption profiles. The authors concluded that K-means provided the best values of the silhouette index and Calinski-Harabasz score when identifying three characteristic regimes. At the same time, seasonality turned out to be the decisive factor, whereas differences between weekdays and weekends were less significant.

Recently, researchers have increasingly shifted their focus from methodological developments to applied tasks, particularly in the field of domestic water consumption. Cominola et al. [15] showed that the use of high-resolution data from smart meters (time intervals of seconds and minutes) makes it possible to classify individual consumption

profiles (shower, laundry, irrigation), thereby improving forecast accuracy and anomaly detection. Cheifetz et al. [16] applied Fourier decomposition and functional clustering to typify households. In the studies of Candelieri [17] and Ioannou et al. [18], preliminary clustering was used to improve forecasting accuracy and to obtain realistic load profiles. In turn, Arsene et al. [19] implemented an IoT platform where the K-means algorithm enabled classification of water consumption events (shower, sink, etc.) and the resulting labels were used for leak prevention.

Despite the widespread use of classical algorithms, their effectiveness decreases when working with complex and highly variable data structures. This has stimulated interest in density-based methods, particularly the DBSCAN algorithm, which can identify clusters of arbitrary shapes without the need to predefine their number. In this context, Mu et al. [20] proposed a streaming version, Stream DBSCAN, for water quality data, where K-means was used at the preliminary stage for node distribution. Song et al. [21] combined STL (Seasonal-Trend decomposition based on Loess) decomposition with DBSCAN to clean online hydromonitoring data from seasonal fluctuations and anomalies. Nasaruddin et al. [22] applied the SMOTE–PCA–HDBSCAN strategy, increasing the sensitivity of detecting rare consumption patterns by about 10%. Zhang et al. [23] integrated HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) with the generative network WSGAIN to restore missing values in monitoring data.

Thus, despite the efficiency of density-based algorithms in detecting clusters without predefined numbers, they remain sensitive to the choice of key parameters. Incorrect parameter settings can lead to errors in data merging, especially in the presence of noise or uneven density. As a result, most studies modify the base algorithms to suit the specifics of the task, confirming the trend toward creating adaptive clustering models. At the same time, existing research is mainly aimed either at identifying daily water consumption profiles or at cleaning and classifying hydrological data, whereas the problem of identifying operating regimes of resource-supplying systems, taking into account temporal structure and energy consumption levels, remains insufficiently addressed. In the present study, we propose an improved density-based clustering algorithm with automatic tuning of the neighborhood radius and the minimum number of points to form a cluster, as well as redistribution of noisy data among clusters.

3. Materials and Methods

3.1. Overview of the Study Area and Data Sources

The object of the study was the water supply system of the city of Gomel (Republic of Belarus), which includes five main water intakes. As clustering parameters, hourly water consumption data for each source for the year 2023 were used, digitized from the logbooks of pump station machine rooms. In addition, electricity consumption data with a 30 min resolution, recorded by the automated commercial electricity metering system, were utilized. To form a consistent statistical database, the hourly water supply volumes were converted into 30 min intervals. The distribution was carried out proportionally to the share of electricity consumption in each half-hour interval relative to the total hourly value. The conversion of data to a 30 min resolution was required due to the electricity billing system in the Republic of Belarus, where accounting is based on half-hourly maximum consumption values. All values were aggregated across the five water intakes. A fragment of the resulting database structure is presented in Table 1.

Table 1. Fragment of the structure of the analyzed data.

Date and Time	Hour of Day	Water Supply, m^3	Electricity, kWh
2023-01-01 00:30:00	0.50	1 468.63	893.16
2023-01-01 01:00:00	0.75	1 558.25	947.66
2023-01-01 01:30:00	1.00	1 547.99	937.35

The analysis was carried out in the form of multivariate clustering using three features: water supply, electricity consumption, and hour of the day. This approach made it possible to simultaneously account for the volumetric (hydraulic) and energy characteristics of the operating modes of water supply facilities, as well as their temporal structure. The use of the temporal feature was necessary to identify typical intra-daily patterns and to label them for further use. Electricity consumption was included to analyze the impact of these modes on consumption levels in optimization tasks.

3.2. Methodological Foundations of Density Clustering

At the first stage of the study of water supply modes, the classical DBSCAN algorithm (Density-Based Spatial Clustering of Applications with Noise) [24–26] is used. It is described by two key parameters: the radius ε and the minimum number of points m (min Pts) [27–33]. The parameter ε defines the maximum distance between two points for identifying neighborhood. This distance determines the size of the ε-neighborhood of each point in the dataset. For two points p and q in an n-dimensional space with coordinates $p = (p_1, p_2, \ldots, p_n)$ and $q = (q_1, q_2, \ldots, q_n)$ the Euclidean distance $d(p, q)$ is determined by the formula:

$$d(p,q) = \sqrt{\sum_{j=1}^{n} (p_i - q_i)^2}, \quad (1)$$

where p_i, q_i—i-th coordinate of points p and q.

Thus, the ε-neighborhood of point p, denoted as $N_\varepsilon(p)$, includes all points q for which the distance $d(p, q)$ does not exceed ε. Figure 1 illustrates the general principle of data classification in the DBSCAN algorithm with a minimum number of points in a cluster $m = 3$.

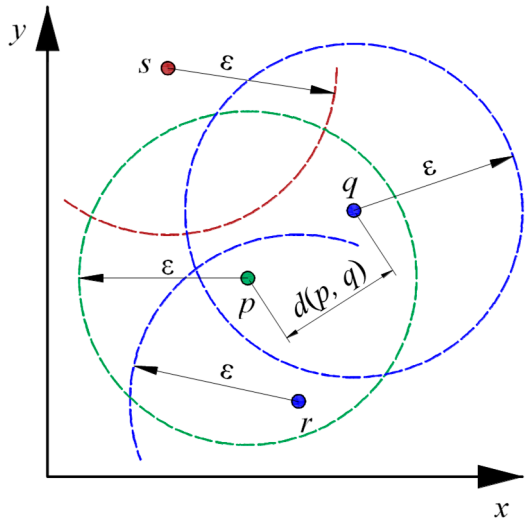

Figure 1. Identifying data points in 2D space.

Within the ε-neighborhood of point p, three types of objects are distinguished: core, border, and noise points. Core points have at least m neighbors ($|N_\varepsilon(p)| \geq m$) and initiate

cluster formation. Points q and r do not reach the threshold m ($|N_\varepsilon(q)| < m$), but since they are located within the ε-neighborhood of a core point p, they are classified as border points. Point s is isolated because its ε-neighborhood lacks a sufficient number of neighbors $q \in N_\varepsilon(p)$, and therefore it is defined as noise.

Since the definition of neighborhood is based on calculating distances between points, an important preparatory step was to bring the original features to a common scale. To ensure data consistency, z-normalization was applied:

$$z = \frac{x - \mu}{\sigma}, \tag{2}$$

where x—the initial value of the feature; μ—the average value of the feature for the sample; σ—the standard deviation of the feature.

After standardization, all features have a mean value of 0 and a standard deviation of 1. Under these conditions, ε is usually within the range $\varepsilon \in [0,\ 1]$; however, the specific value depends on the data distribution and the dimensionality of the original feature space, and, together with the parameter m, becomes a task of optimization.

3.3. Optimization of Density Clustering Parameters

To improve the efficiency of DBSCAN density-based clustering and adapt it to different datasets, an algorithm was developed to optimize the minimum number of points m and the neighborhood radius ε. The procedure includes: (1) standardizing the data using z-normalization; (2) calculating distances to the k nearest neighbors for each point; (3) computing the average distances to the k-th neighbor across the dataset; (4) analyzing the changes in these values to adaptively determine the minimum number of neighbors m (min Pts); and (5) selecting the neighborhood radius ε as the median of the distances to the m-th nearest neighbor. Figure 2 illustrates the steps of optimizing the key parameters of density-based clustering.

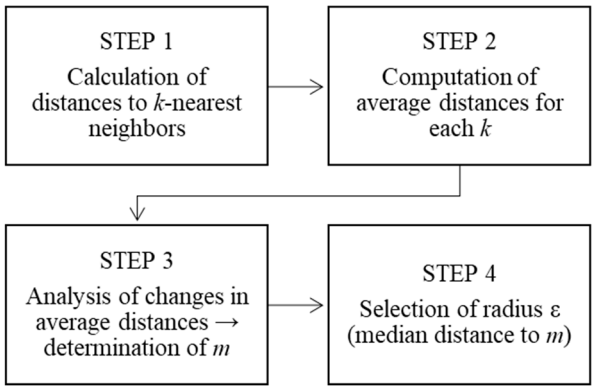

Figure 2. Stages of optimization of the main parameters of density clustering.

Let us examine the optimization steps in more detail:

STEP 1: Calculation of distances to the nearest neighbors

For each point x_i in the dataset, the distances to its k nearest neighbors are calculated. The initial value of k is set to 2 and is gradually increased up to a value limited by the sample size. Due to the computational cost of this operation on large datasets, it is reasonable to set an upper bound for k. In the conducted experiments, this limit was set to k = 100, which provided a sufficient margin for estimating data density. Next, for each point in the dataset (e.g., $x_1, x_2, \ldots, x_n$), the distances to all other points are computed. These distances are then sorted in ascending order, and the first k values are selected. If $d(x_i, x_j)$ denotes the distance between points x_i, x_j, the distance to the k-th nearest neighbor can be written

as $d_k(x_i)$. For example, if $k = 3$, then for point x_i, the distance to its third nearest neighbor is denoted as $d_3(x_i)$.

STEP 2: Calculation of average distances to the k nearest neighbors

For each value of k, the average distance to the k-th nearest neighbor across all points is calculated. This average value reflects the data density for different numbers of neighbors. Accordingly, if $d_k = (x)$ denotes the distance from point x to its k-th nearest neighbor, then the average distance for each k is defined as:

$$\bar{d_k} = \frac{1}{n}.\sum_{i=1}^{n} d_k(x_i), \tag{3}$$

where n is the total number of points in the dataset.

STEP 3: Analysis of changes in average distances

At this stage, the changes in $\bar{d_k}$ as k increases are analyzed. The critical point occurs when increasing k leads to only a negligible increase in the average distance. In such cases, the data density begins to stabilize. The search for the optimal value of k is determined by the condition:

$$\frac{\bar{d_k} - \bar{d}_{k-1}}{\bar{d}_{k-1}} < \theta, \tag{4}$$

where $\bar{d_k}$, $\bar{d}_{k-1}$—the average Euclidean distances to the k-th and (k–1)-th nearest neighbors across all points at the current and previous steps, respectively; θ—the threshold value that defines how small the change in the average distance must be in order to stop increasing k.

In the conducted experiments, the criterion θ was set to 0.01, which corresponded to a threshold of no more than 1% change between two consecutive steps in k. This approach allowed for adaptively determining the required minimum number of points m for adequate clustering of the data, taking into account their internal structure and density distribution.

STEP 4: Selection of the neighborhood radius ε

After determining the minimum number of points m based on the distance analysis, the optimal value of ε is calculated. For this purpose, the distance from each point to its m (min Pts) nearest neighbor is computed. The median of these distances is then taken as the optimal value of ε. Thus, if $d_m(x)$ denotes the distance from point x to its m-th nearest neighbor, then ε is defined as:

$$\varepsilon = \text{median}(d_m(x_1), d_m(x_2), \ldots, d_m(x_n)). \tag{5}$$

The median of these distances is used as the optimal value of ε because it represents the typical distance at which a point can still be considered sufficiently close to a group of other points. In its canonical form, optimization of density clustering parameters is aimed at finding the number of points m used in the ε-neighborhood that will ensure stable cluster formation. The objective function here is defined as:

$$m^* = min\left\{ m \middle| \frac{\bar{d_m} - \bar{d}_{m-1}}{\bar{d}_{m-1}} < \theta \right\},\ m = \in mmin, mmax \tag{6}$$

where m^*—the optimal number of points within the boundaries m_{min} of and m_{max}; $\bar{d}_m$—the average distance to the m-th neighbor.

The search for the optimal value of m^* is performed within a given maximum number of nearest neighbors $m \leq k_{max}$.

3.4. Redistribution of Noise Data to Nearby Clusters

As a result of nonparametric density-based clustering, a portion of the data points is inevitably classified as noise, since they do not meet the requirement of the minimum number of neighbors within the radius ε. Such data cannot be assigned to the formed clusters; however, their automatic exclusion is not always the optimal solution. In some cases, it is advisable to consider assigning noise points to the nearest cluster, which makes it possible to preserve potentially valuable information. According to the proposed algorithm, Figure 3 shows the correlation field of clustering by three features: water supply, electricity consumption and hours of day.

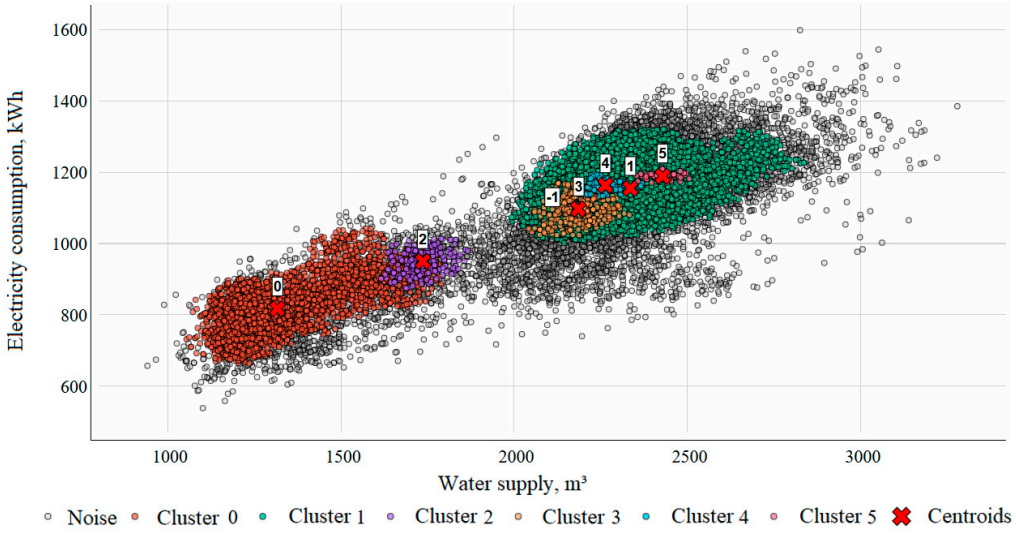

Figure 3. Clustering results based on three parameters (water consumption, electricity consumption and time of day) with noise point detection.

The resulting density groups correspond to five clusters, but around them, numerous points classified as noise (shown in gray) form. Automatically removing such data results in information loss and reduces the interpretability of the modes. In the proposed approach, noise points are not excluded but rather redistributed to the nearest clusters according to the following algorithm:

(1) Identification of noise data. At the first step, the indices of points classified as noise are determined:

$$I_\omega = \{i | L_i = \omega\}, \tag{7}$$

where L_i—the cluster label of the i-th data point; ω—the noise label.

(2) Selection of noise data. A subset of data corresponding to noise points is selected:

$$X_\omega = \{x_i | i \in I_\omega\}, \tag{8}$$

where x_i—the value of the i-th data point identified as noise.

(3) Selection of core data. Core data belonging to clusters are separated from the data identified as noise:

$$X_{cl} = \{x_i | L_i \neq \omega\}. \tag{9}$$

(4) Search for the nearest neighbors of each noise point. For each noise point x_i, the nearest point from the subset of clustered data X_{cl} is found:

$$j(i) = \underset{j \in X_{cl}}{\operatorname{argmin}}\, d(x_i, x_j), \tag{10}$$

where $d(x_i, x_j)$—the Euclidean distance between the noise point x_i and the core point x_j.

(5) Reassignment of labels. Each noise point is assigned the label of the nearest cluster:

$$L_i = L_{j(i)} \text{ for all } i \in I_\omega. \tag{11}$$

(6) Recalculation of cluster centroids. After redistributing the noise data, it becomes necessary to update the centroids taking into account the newly assigned values:

$$c_l = \frac{1}{|S_l|} \sum_{i \in S_l} x_i \text{ for all } l \in L, \tag{12}$$

where $S_l = \{i | L_i = l\}$—the set of indices of points in cluster l; c_l—is the centroid of cluster l.

3.5. Modified Non-Parametric Clustering Algorithm

To address the task of identifying stable modes from multidimensional data, a modified density-based clustering algorithm was developed. This algorithm formed the basis of the software implementation in Python-3.9.9. using the Streamlit framework [26]. Figure 4 shows the resulting block removal of the modified nonparametric density clustering algorithm that was used in the analysis of the study.

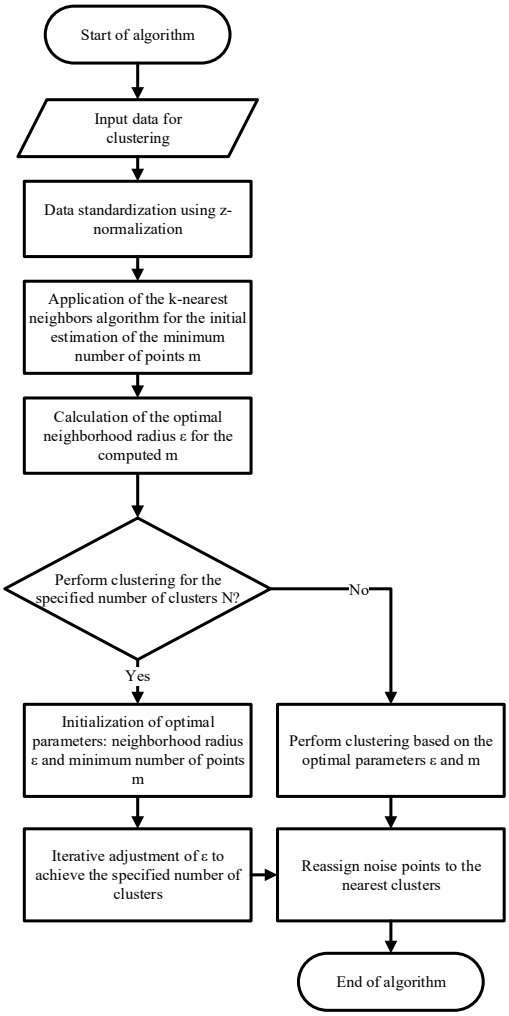

Figure 4. Modified non-parametric density-based clustering algorithm.

The algorithm includes the following steps. First, the initial statistics are loaded and structured. Next, the data is scaled using z-normalization. In the next step, the initial value

of the parameter m is determined using the k-nearest neighbors' method, after which the optimal value of the neighborhood radius is determined.

After selecting the basic parameters m and ε, the resulting number of clusters is checked to ensure that it corresponds to the specified value N. This check is necessary to exclude clusters formed by randomly combining groups that do not correspond to real-world conditions. Thus, if the number of clusters does not correspond to the desired value, the initial parameters m and ε are iteratively adjusted until a stable cluster structure is achieved. In the final step, noisy cluster points that were not assigned to the nearest groups in the first step are redistributed. The result of the algorithm is shown in Figure 5.

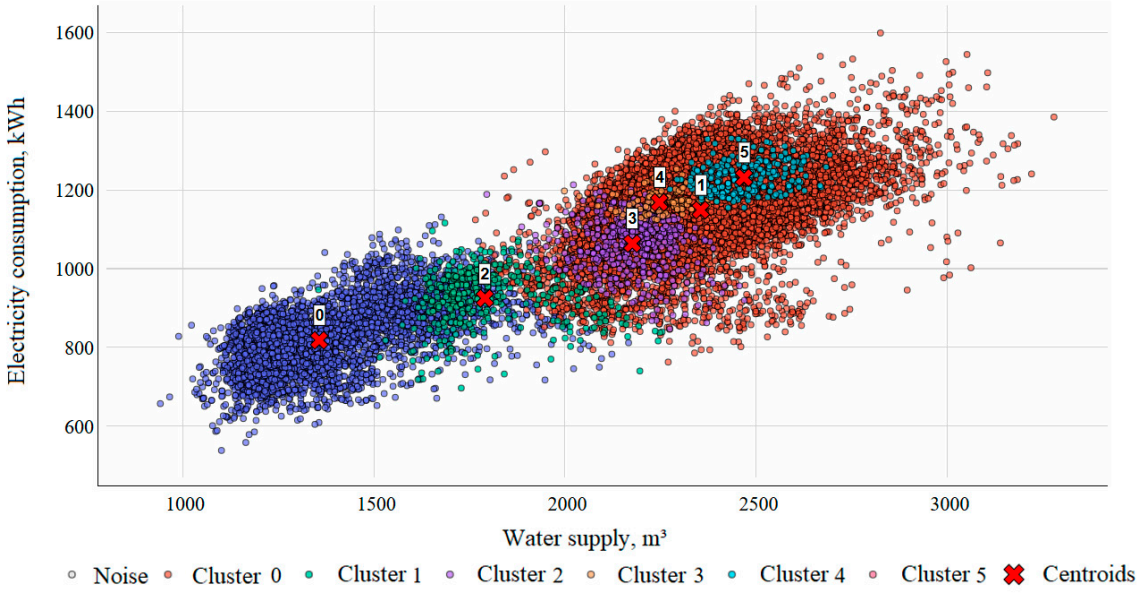

Figure 5. Results of clustering by three parameters (water consumption, electricity consumption and time of day) with redistribution of noise points to the nearest clusters.

4. Results and Discussion

This section presents the results of applying a modified density clustering algorithm to urban water supply system data. The primary objective is to label time regimes and compare the results obtained using various feature selection strategies. Two approaches are considered: (1) clustering based on time of day, water supply, and electricity consumption; (2) an extended approach that includes the factor of average water pressure at the pumping station outlet. These strategies will allow us to assess changes in cluster structure and test their stability when adding additional hydraulic components.

4.1. Clustering Based on Time of Day and Water Supply and Electricity Consumption Parameters

For the analysis of water consumption patterns, half-hourly data for 2023 were used, including the total water supply and electricity consumption from all sources in the city, supplemented with a temporal factor (hour of the day). The optimal parameters of the algorithm were determined using the k-nearest neighbors' method. Figure 6 shows the dependence of the average distance to the k-th neighbor. It can be seen that at small values of k the distance increases rapidly; however, after approximately $k \approx 25$, the curve begins to level off, and stabilization is observed at $k \approx 31$. At this point, the search algorithm terminates. As a result, the values $\varepsilon = 0.12$ and $m = 31$, were obtained, which balanced the algorithm's sensitivity to local data density and helped to avoid both excessive fragmentation and over-merging of clusters.

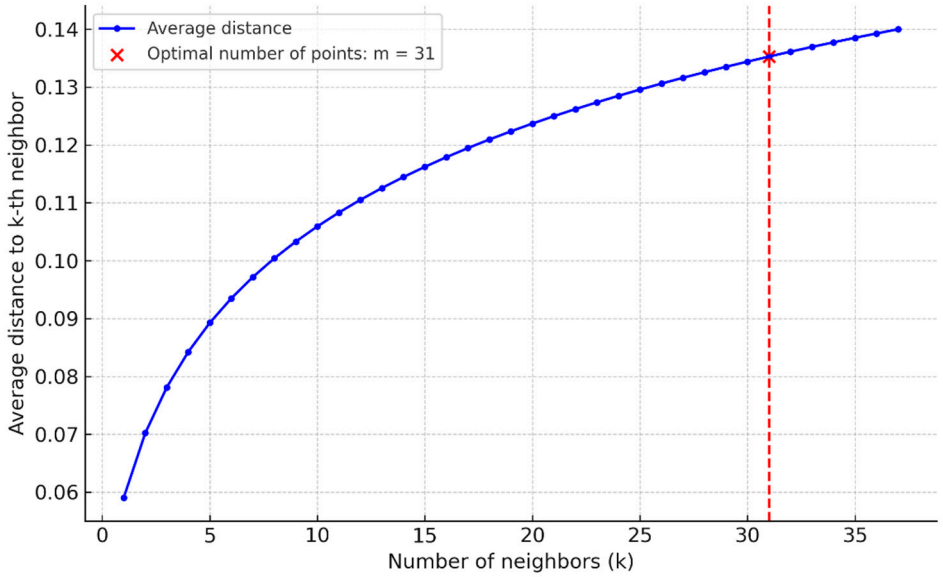

Figure 6. Variation in the average distance to the *k*-th neighbor of data points.

For clarity, the clustering results are presented in the form of a three-dimensional projection of the data, where water supply, electricity consumption, and hours of the day were used as coordinates (Figure 7). It is evident that the algorithm identified five clusters: some were obtained by merging local dense groups, whereas two clusters exhibit significant spread and cover a wide range of values. Such a structure confirms the presence of both stable modes and generalized clusters that combine several sub-modes. In Figure 7, for better clarity, the cluster centers are shown without noise points, but these points were taken into account in the subsequent analysis of the modes.

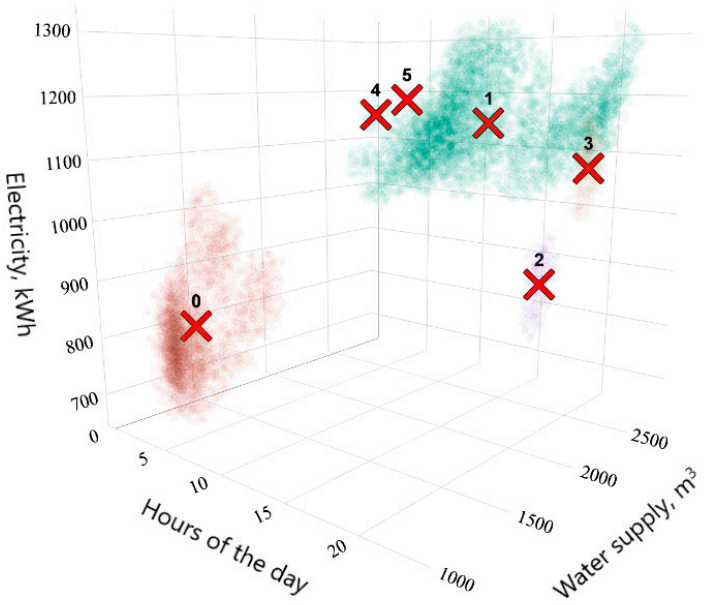

Figure 7. Three-dimensional visualization of water and energy consumption modes depending on the time of day using a modified DBSCAN method.

As a result of clustering, the data were divided into six groups. Of greatest interest for analyzing water consumption patterns are only two of them—Cluster 0 and Cluster 1. Cluster 0 corresponds to a regime with a low level of water supply and electricity consumption, whereas Cluster 1 reflects a high-demand regime characterized by increased

system loads. These two clusters form the main structure of the analyzed data and define the stable operating modes. The remaining clusters (2, 3, 4, and 5) occur much less frequently and represent transitional states arising during limited time intervals. Their role is not in forming key regimes but in describing transition processes, and therefore they have an auxiliary significance when interpreting the results. Figure 8 presents a three-dimensional distribution of the data, visualizing the frequency of occurrence of different regimes in the coordinates of water supply and electricity consumption. This illustration clearly demonstrates which system states are the most typical and which occur only occasionally.

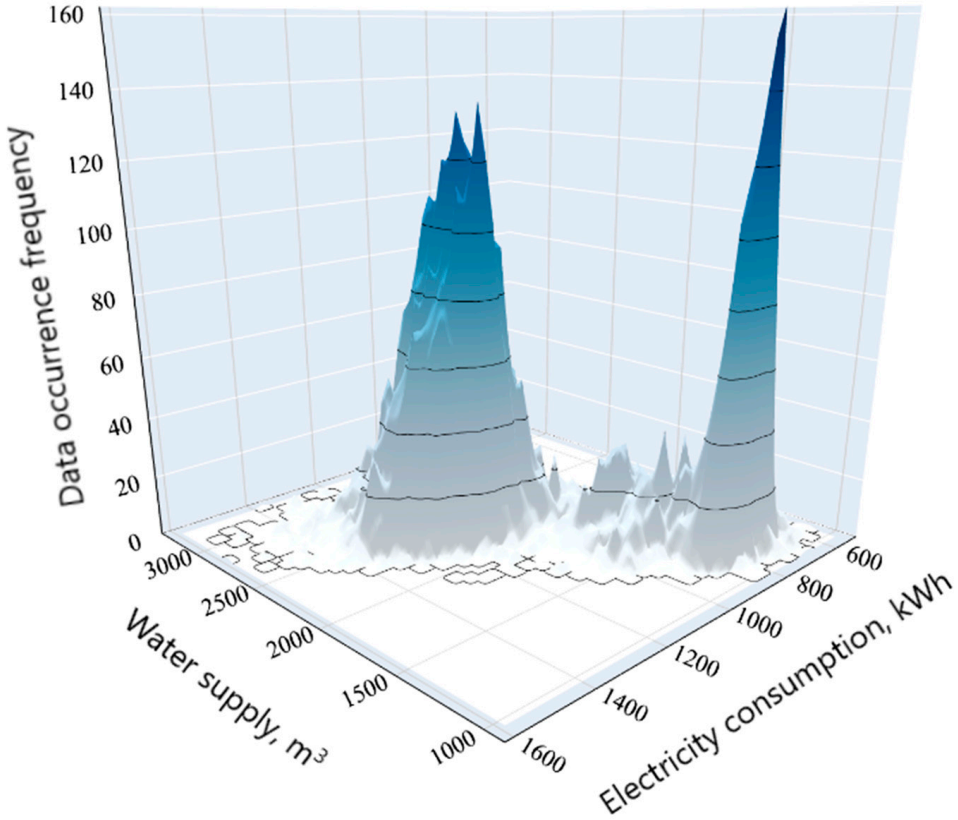

Figure 8. Three-dimensional visualization of water consumption and electricity consumption regimes.

To improve the robustness of the clustering results, extreme values were excluded from the analysis. Data within the 5-95% percentile range were considered in the calculations. Table 2 presents the main results of nonparametric clustering of the data.

Table 2. The main results of clustering modes based on the time of day and parameters of water supply and electricity consumption.

Cluster	Time Range (Within 5–95% Percentile)	Average Water Supply, m^3	Average Power Consumption, kW·h	Data Share, %
0	00:00–06:30	1357	818	29.1
1	08:00–22:00	2355	1149	63.6
2	20:30–23:30	1791	925	2.3
3	21:00–23:00	2176	1064	2.9
4	06:30–08:00	2247	1169	0.7
5	06:30–08:30	2469	1231	1.3

In the baseline scenario, the analysis of the obtained clusters allowed us to identify both stable system modes and transient states. The most significant are the modes of

morning and evening water demand. The first cluster (00:00–06:30) is characterized by the minimum value of water supply (on average 1357 m^3) and electricity consumption (818 kW·h), which accounts for 29.1% of all data. The second mode (08:00–22:00) coincides with the active phase of the city's daily activity and, on average, determines the water supply of 2355 m^3 and electricity consumption of 1149 kW·h, with a share of observations of 63.6%. The remaining groups have a significantly smaller specific weight and form characteristic transient intervals. Clusters 4 and 5 (06:30–08:30) describe the morning load peaks (on average up to 2469 m^3 of water and 1231 kW·h of electricity). Evening transient states are included in clusters 2 and 3 (8:30 p.m.–11:30 p.m.) and include individual local maxima. The proportion of transient states reflects changes in less than 6% of the data.

For clarity, the distribution of clusters by hour of day is presented in Figure 9, where on the left is a heat map of the appearance of data in the time interval, and on the right is a box plot illustrating the distribution of clusters relative to the time of day.

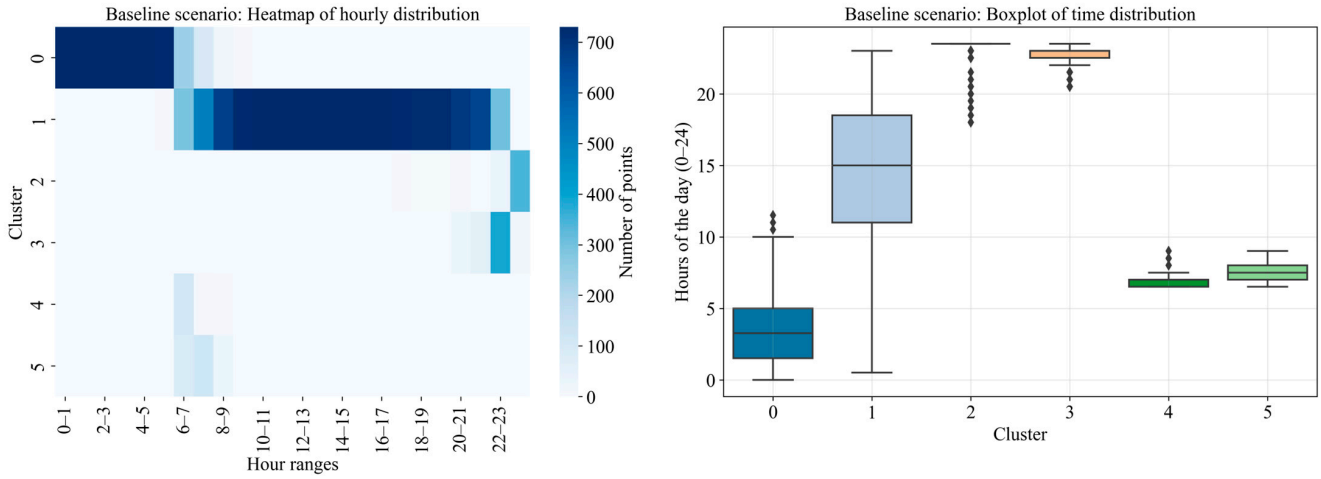

Figure 9. Distribution of clusters by hours of day in the baseline scenario (clustering by time of day and parameters of water supply and electricity consumption).

The obtained results for identifying temporary clusters are consistent with the characteristic patterns corresponding to the city's actual daily activity. However, for a more detailed assessment and identification of additional transient patterns, an additional parameter, the pressure at the pumping station outlet, will be included in the analysis.

4.2. Extended Clustering with the Addition of a Pressure Factor

To improve robustness against labeling the main operating modes, a hydraulic factor accounting for the average pressure at the pumping station outlet was added to the model. The mode analysis was performed using the same timeframe (half-hourly data for 2023), and the feature space was expanded to four dimensions: water supply, electricity consumption, hours of day, and pressure. Before clustering, all features were scaled to a single scale (z-normalization), and the density clustering parameters (ε, m) were re-evaluated using the k-nearest neighbors procedure discussed earlier. The clustering results are presented in Table 3.

A comparison of the results of the basic (Table 2) and extended clustering (Table 3) shows that the core of the daily water consumption structure has been preserved. In both scenarios, two regimes remain key: the nighttime minimum (00:00–06:30) and the daytime maximum (07:30–22:30), which together account for over 70% of the sample. However, when pressure is added, a slight shift in the boundaries of the time intervals and the emergence of additional transient regimes is observed. Thus, the nighttime cluster is

divided into two groups (0 and 2), detailing the minimum regime, while the daytime regime is formed by two clusters (1 and 3).

Table 3. Main results of extended clustering with the addition of the pressure factor.

Cluster	Time Range (Within 5–95% Percentile)	Average Water Supply, m^3	Average Power Consumption, kW·h	Average Pressure, kPa	Data Share, %
0	00:30–05:00	1254	796	324	5.7
1	12:00–22:00	2332	1149	391	11.5
2	00:00–06:30	1372	822	369	23.0
3	07:30–22:30	2361	1154	370	53.0
4	06:30–11:00	2251	1047	393	1.3
5	07:30–11:30	2355	1106	389	1.9
6	06:30–12:30	1966	915	370	1.4
7	21:30–23:30	1786	926	367	2.3

After adding hydraulic pressure, transient regimes, which were previously less pronounced, are recorded. The morning intervals are described by clusters 4–6, distinguished by the stratification of regimes across different time ranges (06:30–11:00, 07:30–11:30, and 06:30–12:30). The evening decline is classified into cluster 7 (9:30 p.m.–11:30 p.m.), which accounts for a distinct time range of reduced water supply conditions in the water supply system. The overall share of transient conditions is insignificant (approximately 6.9%), which may reflect their likely dependence on random water supply characteristics. Figure 10 shows a heat map of the data distribution over time and a cluster span diagram.

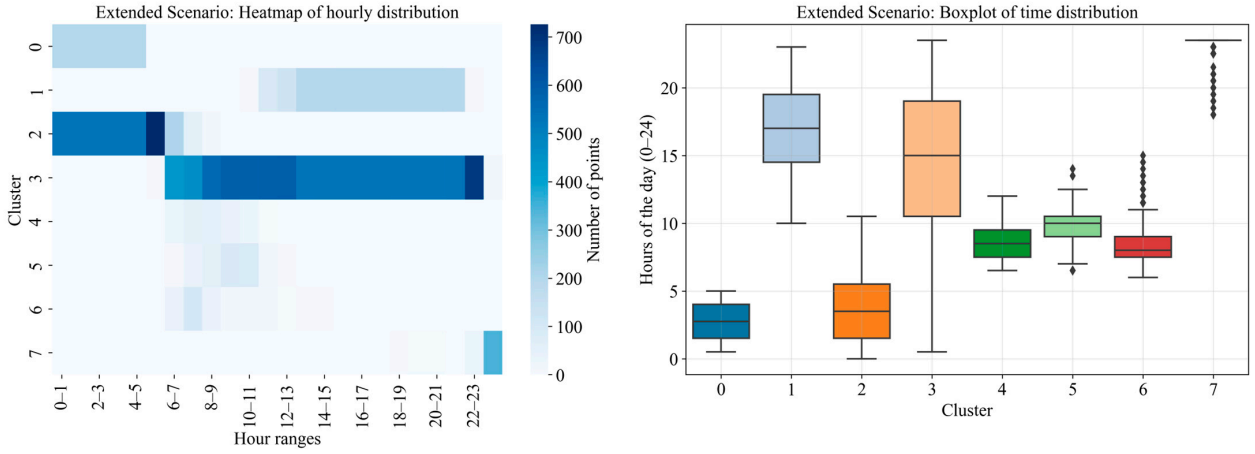

Figure 10. Distribution of clusters by hours of the day in the extended scenario (clustering by time of day, and parameters of water supply, pressure and electricity consumption).

The obtained results can be used to formalize and label time regimes in applied analysis and forecasting problems. In particular, the resulting clusters can serve as additional features in constructing water consumption prediction models by refining the structure of the time series.

4.3. Comparative Analysis of Clustering Results Using the K-Means Method and the Modified DBSCAN Method

To verify the robustness of the results, extended clustering was compared with the K-means method. Under comparable conditions, the number of clusters was pre-fixed at k = 8, allowing the obtained results to be directly correlated with previously identified modes based on water supply, pressure, power consumption, and time of day. Figure 11

presents a three-dimensional visualization of the obtained results. It can be seen that, unlike the proposed algorithm, the clarity of transient mode detection is reduced, with the resulting clusters forming smooth, centered shapes without a clear distinction between short morning and evening intervals. Furthermore, within the daily peak, an artificial division into several closely related groups is observed, complicating data interpretation.

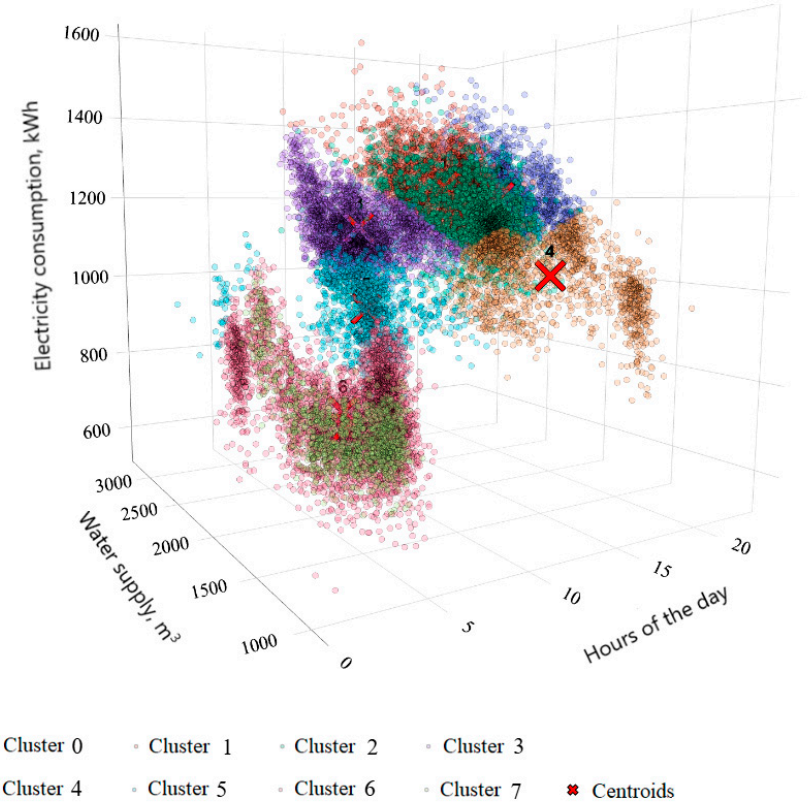

Figure 11. Three-dimensional visualization of water and energy consumption modes using the K-means method.

The resulting clusters, using the K-means method, partially captured the nighttime and daytime characteristics of consumption, but with significant limitations. Two clusters divided the nighttime patterns into the 12:00–6:00 and 12:30–5:00 ranges, which partially coincides with the proposed method. The daytime peak is also represented by several overlapping clusters (8:30–21:30, 12:00–17:30, 18:00–22:30), which partially overlap and form a smoothed structure. Morning intervals are also divided into two identical time clusters (6:00–13:00), and transitional evening patterns are blurred into the 15:30–23:30 and 18:00–22:30 ranges. This demonstrates that K-means can reproduce the main trends of the diurnal cycle but it is accompanied by excessive fragmentation and a decrease in the clarity of the patterns. Table 4 presents the results of a comparison of clustering results using the modified DBSCAN method and K-means.

The comparison showed that the K-means method is capable of capturing key trends, but is inferior to the modified DBSCAN method in terms of interpretability. The main drawback is the need to specify a predetermined number of clusters, which may not be obvious given the heterogeneity of the regimes. This results in a smoothed and less realistic data structure, in which the regimes are stratified into overlapping groups.

Table 4. Comparison of clustering results using modified DBSCAN and K-means methods.

Mode	Modified DBSCAN	K-Means	Note
Nighttime minimum	00:30–05:00 (cluster 0, 5.7%) and 00:00–06:30 (cluster 2, 23.0%)	00:00–06:00 (cluster 6, 21.6%) and 00:30–05:00 (cluster 7, 5.7%)	Both methods record the night minimum, the combined share being about a quarter of the sample.
Daytime maximum	07:30–22:30 (cluster 3, 53.0%) and 12:00–22:00 (cluster 1, 11.5%)	08:30–21:30 (cluster 2, 14.2%), 12:00–17:30 (cluster 1, 15.1%), 18:00–22:30 (cluster 0, 13.2%)	In DBSCAN, the daily maximum consistently occupies ~65% of the sample, while in K-means it is split into several intersecting intervals (a total of about 42%).
Morning transitional	06:30–11:00 (cluster 4, 1.3%), 07:30–11:30 (cluster 5, 1.9%), 06:30–12:30 (cluster 6, 1.4%)	06:00–13:00 (cluster 5, 6.4%), 06:30–13:00 (cluster 3, 15.1%)	In DBSCAN, transient modes occupy a small share (<5%), while in K-means these same intervals are extended and form up to 20% of the sample.
Evening transitional	21:30–23:30 (cluster 7, 2.3%)	15:30–23:30 (cluster 4, 8.7%), 18:00–22:30 (cluster 0, 13.2%)	DBSCAN records a short evening decline (<3%), K-means distributes it into long intervals with a share of more than 20%.

5. Conclusions

The study confirmed the effectiveness of a modified DBSCAN algorithm for analyzing intra-day water consumption patterns. The developed approach allowed us to identify clusters of varying density and account for transient states, which is difficult to achieve using traditional methods. An analysis of 2023 data revealed two main stable patterns: high demand (07:30–22:30 and 12:00–22:00), accounting for 64.5% of the data, and low demand (00:00–06:30), accounting for 28.7%. The remaining clusters, collectively accounting for approximately 6.9% of the data, are transient in nature and reflect localized data accumulations between these states, without significantly impacting the overall water consumption structure.

The practical significance of these results lies in their potential application in scheduling pumping stations and managing electrical loads. The following mode labeling is proposed for the system's operation: clusters 0 and 2 correspond to low demand (00:00–06:30), clusters 4–6 reflect the morning transitional mode (06:30–12:30), clusters 1 and 3 correspond to high demand (07:30–22:30), and cluster 7 characterizes the evening transitional mode (21:30–23:30). This classification allows for data labeling for testing predictive models of water consumption modes.

Author Contributions: Conceptualization, A.A.K. and V.S.B.; methodology, R.V.K. and N.V.H.; formal analysis, A.A.K. and N.V.H.; investigation, V.S.B.; data curation, N.V.H. and R.V.K.; writing—original draft preparation, N.V.H. and A.A.K.; writing—review and editing, R.V.K.; supervision, R.V.K.; project administration, V.S.B.; visualization, A.A.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data presented in this study are available on request from the corresponding author due to legal reasons.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Zapata, O. More Water Please, It's Getting Hot! The Effect of Climate on Residential Water Demand. *Water Econ. Policy* **2015**, *1*, 1550007. [CrossRef]
2. Brigida, V.S.; Golik, V.I.; Dzeranov, B.V. Modeling of Coalmine Methane Flows to Estimate the Spacing of Primary Roof Breaks. *Mining* **2022**, *2*, 809–821. [CrossRef]
3. Guedes, B. Crop Evapotranspiration and Water Use Efficiency. In *Irrigation Systems and Practices in Challenging Environments*; BoD—Books on Demand: Norderstedt, Germany, 2012. [CrossRef]
4. Heidari, H.; Arabi, M.; Warziniack, T.; Sharvelle, S. Effects of Urban Development Patterns on Municipal Water Shortage. *Front. Water* **2021**, *3*, 694817. [CrossRef]
5. Ghatani, S. Problems and Challenges on Urban Water Management in Darjeeling Hill Town. *Asian Res. J. Arts Soc. Sci.* **2021**, *13*, 24–33. [CrossRef]
6. Ren, W.; Bai, X.; Wang, Y.; Liang, C.; Huang, S.; Wang, Z.; Yang, L. Analysis of Water Supply-Demand Based on Socioeconomic Efficiency. *J. Sensors* **2022**, *2022*, 1–16. [CrossRef]
7. Kapanski, A.; Hruntovich, N.; Bakhur, S.; Markaryants, L.; Dolomanyak, L. Optimize the cost of paying for electricity in the water supply system by using accumulating tanks. *E3S Web Conf.* **2020**, *178*, 01065. [CrossRef]
8. Hristov, P.; Hristova, T. Explaining The DLT Applications in The Context of a Customers, Facility Managements and Utility Companies Relationship. In Proceedings of the 2019 16th Conference on Electrical Machines, Drives and Power Systems (ELMA), Varna, Bulgaria, 6–8 June 2019; pp. 1–5. [CrossRef]
9. Golik, V.I.; Razorenov, Y.U.I.; Brigida, V.S.; Burdzieva, O.G. Mechanochemical technology of metal mining from enriching tails. *Bull. Tomsk. Polytech. Univ. Geo. Assets Eng.* **2020**, *331*, 175–183. [CrossRef]
10. Kapanski, A.; Hruntovich, N.V.; Klyuev, R.V.; Brigida, V. Identification of Easily Accessible Urban Water Con-sumption Factors for Energy-Efficient Management of Pumping Stations. *Water Conserv. Sci. Eng.* **2025**, *10*, 1–16. [CrossRef]
11. Liao, T.W. Clustering of time series data—A survey. *Pattern Recognit.* **2005**, *38*, 1857–1874. [CrossRef]
12. Rani, S.; Sikka, G. Recent techniques of clustering of time series data: A survey. *Int. J. Comput. Appl.* **2012**, *52*, 1–9. [CrossRef]
13. Prakaisak, I.; Wongchaisuwat, P. Hydrological Time Series Clustering: A Case Study of Telemetry Stations in Thailand. *Water* **2022**, *14*, 2095. [CrossRef]
14. Guo, H.; Liu, X.; Zhang, Q. Identifying daily water consumption patterns based on K-means Clustering, Agglomerative Hierarchical Clustering, and Spectral Clustering algorithms. *AQUA-Water Infrastruct. Ecosyst. Soc.* **2024**, *73*, 870–887. [CrossRef]
15. Cominola, A.; Giuliani, M.; Castelletti, A.; Rosenberg, D.; Abdallah, A. Implications of data sampling resolution on water use simulation, end-use disaggregation, and demand management. *Environ. Model. Softw.* **2018**, *102*, 199–212. [CrossRef]
16. Cheifetz, N.; Noumir, Z.; Samé, A.; Sandraz, A.-C.; Féliers, C.; Heim, V. Modeling and clustering water demand patterns from real-world smart meter data. *Drink. Water Eng. Sci.* **2017**, *10*, 75–82. [CrossRef]
17. Candelieri, A. Clustering and Support Vector Regression for Water Demand Forecasting and Anomaly Detection. *Water* **2017**, *9*, 224. [CrossRef]
18. Ioannou, A.E.; Creaco, E.F.; Laspidou, C.S. Exploring the Effectiveness of Clustering Algorithms for Capturing Water Consumption Behavior at Household Level. *Sustainability* **2021**, *13*, 2603. [CrossRef]
19. Arsene, D.; Predescu, A.; Pahonțu, B.; Chiru, C.G.; Apostol, E.-S.; Truică, C.-O. Advanced Strategies for Monitoring Water Consumption Patterns in Households Based on IoT and Machine Learning. *Water* **2022**, *14*, 2187. [CrossRef]
20. Mu, C.; Hou, Y.; Zhao, J.; Wei, S.; Wu, Y. Stream-DBSCAN: A Streaming Distributed Clustering Model for Water Quality Monitoring. *Appl. Sci.* **2023**, *13*, 5408. [CrossRef]
21. Song, C.; Cui, J.; Cui, Y.; Zhang, S.; Wu, C.; Qin, X.; Wu, Q.; Chi, S.; Yang, M.; Liu, J.; et al. Integrated STL-DBSCAN algorithm for online hydrological and water quality monitoring data cleaning. *Environ. Model. Softw.* **2025**, *183*, 106262. [CrossRef]
22. Nasaruddin, N.; Masseran, N.; Idris, W.M.R.; Ul-Saufie, A.Z. A SMOTE PCA HDBSCAN approach for enhancing water quality classification in imbalanced datasets. *Sci. Rep.* **2025**, *15*, 1–12. [CrossRef]
23. Zhang, F.; Guo, J.; Yuan, F.; Qiu, Y.; Wang, P.; Cheng, F.; Gu, Y. Enhancement Methods of Hydropower Unit Moni-toring Data Quality Based on the Hierarchical Density-Based Spatial Clustering of Applications with a Noise–Wasserstein Slim Generative Adversarial Imputation Network with a Gradient Penalty. *Sensors* **2023**, *24*, 118. [CrossRef]
24. Birant, D.; Kut, A. ST-DBSCAN: An algorithm for clustering spatial–temporal data. *Data Knowl. Eng.* **2007**, *60*, 208–221. [CrossRef]
25. Bhowmik, A.; Kumar, R.; Ranganathaswamy, M.K.; Kumar, Y.K.; Samal, P.; Mahapatro, A.; Alhazaa, A.N.; Romanovski, V.; Santhosh, A.J. Predictive modeling of the mechanical behavior of 3D-printed polylactic acid/wood composite: Comparison of GEP and ANN methods. *AIP Adv.* **2025**, *15*, 045221. [CrossRef]
26. Romanovski, V.; Moskovskikh, D.; Tan, H.; Kuskov, K.; Volodko, S.; Akinwande, A.A.; Periakaruppan, R.; Kong, F.; Ma, X.; Yang, F.; et al. Gypsum Binder with Increased Water Resistance Derived from Membrane Water Desalination Waste. *Eng. Rep.* **2024**, *7*, e13028. [CrossRef]
27. Hahsler, M.; Piekenbrock, M.; Doran, D. dbscan: Fast Density-Based Clustering with R. *J. Stat. Softw.* **2019**, *91*, 1–30. [CrossRef]

28. Gholizadeh, N.; Saadatfar, H.; Hanafi, N. K-DBSCAN: An improved DBSCAN algorithm for big data. *J. Supercomput.* **2020**, *77*, 6214–6235. [CrossRef]
29. Johnston, B.; Jones, A.; Kruger, C. *Applied Unsupervised Learning with Python: Discover Hidden Patterns and Relationships in Unstructured Data with Python*; Packt Publishing Ltd.: Birmingham, UK, 2019; 500p.
30. Ester, M.; Kriegel, H.-P.; Xu, X. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD'96), Portland, OR, USA, 2–4 August 1996; pp. 226–231.
31. Hao, W.; Cominola, A.; Castelletti, A. Short-Term Memory and Regional Climate Drive City-Scale Water Demand in the Contiguous US. *Earth's Future* **2025**, *13*, e2024EF004415. [CrossRef]
32. Spinelli, D.; Giuliani, M.; Castelletti, A. Ensemble Forecasts with Blocked K-Fold Cross-Validation in Multi-Objective Water Systems Control. In Proceedings of the 2024 European Control Conference (ECC), Stockholm, Sweden, 25–28 June 2024; pp. 493–498. [CrossRef]
33. Pesantez, J.E.; Berglund, E.Z.; Kaza, N. Smart meters data for modeling and forecasting water demand at the user-level. *Environ. Model. Softw.* **2020**, *125*, 104633. [CrossRef]