

ORIGINAL RESEARCH

Open access

Spatial Clustering of Water Consumption Patterns Using Ward's Method: Implications for Infrastructure Zoning and Sensor Optimization in Smart Cities

George Wilson^{1*}, Chloe Bennett², Jack Turner¹

Abstract

In large cities with evolving infrastructure, optimizing water supply systems is critical. However, tackling these issues becomes considerably more difficult without a thorough understanding of the network layout and water-use patterns. This paper introduces a methodology based on geospatial data analysis to address two primary objectives. The first is the definition of service zones for infrastructure facilities to improve overall system manageability. The second is the development of an approach to determine the optimal locations for devices that collect and transmit hydraulic network parameters, ensuring their placement aligns with both water supply sources and the areas they serve. The research utilizes data from the water supply network of a city with a population of more than half a million, applying hierarchical clustering with Ward's method to examine spatial distribution. Four territorial clusters were distinguished, each exhibiting distinct characteristics related to consumer density and water consumption volumes. The resulting cluster boundaries were compared with the system's current service scheme, demonstrating consistency with the actual infrastructure. Clustering quality was additionally assessed using the silhouette coefficient, which confirmed the high precision and dependability of the selected method. The study illustrates the value of visualizing cluster boundaries in evaluating the even placement of pressure sensors throughout the urban water supply network. The findings indicate that combining geographic data with water consumption details not only supports efficient infrastructure planning and resource distribution but also lays the foundation for the digital transformation of the hydraulic network—an essential element of sustainable progress in contemporary smart cities.

Keywords Spatial clustering, Optimisation, Geographic segmentation, Ward's method, Water resource management, Silhouette quotient

*Correspondence:

George Wilson

g.wilson@leeds.ac.uk

¹ Department of Smart Grid Analytics, Faculty of Engineering, University of Leeds, Leeds, United Kingdom

² Department of Energy Optimization Research, Faculty of Engineering, University of Sheffield, Sheffield, United Kingdom

Introduction

In the current environment of growing demands for reliability and efficiency in resource distribution systems, the adoption of cutting-edge technologies—including the Internet of Things (IoT), machine learning, and artificial intelligence—has become increasingly relevant. These technologies facilitate the collection, processing, and real-time analysis of data, creating new opportunities to improve the performance of sophisticated technical systems, such as municipal water supply networks. At present, water supply systems consist of intricate networks of equipment and infrastructure that continuously extract, transport, and deliver potable water over extensive areas. In major urban centers, these systems are multi-level, incorporating various stages of water elevation to sustain appropriate pressure across different parts of the

network. Research shows that such systems are affected by a wide range of external and internal factors that can substantially influence their performance and dependability [1-3]. Among the primary challenges are optimizing water resource allocation, reducing water losses, and preventing operational disruptions in technological processes [4, 5].

Consequently, the effectiveness of water supply systems relies on multiple aspects, such as establishing optimal operational regimes, responding swiftly to incidents, and maintaining pipelines and equipment in good condition. Ensuring the stable functioning of these systems demands ongoing condition monitoring, in-depth evaluation of performance indicators, and prompt decision-making by management. Nevertheless, current methods for managing water supply systems often face limitations that hinder a complete understanding of the network's structure and operation. Notably, there is a lack of reliable techniques for precisely mapping the distribution of major consumers across a geographic area, which poses considerable difficulties for conducting energy audits. This complicates the assessment of object placement relative to water intake points and the evaluation of their effects on the overall network. These constraints hinder not only monitoring efforts but also the identification of strategies to enhance system efficiency.

The majority of contemporary water supply systems generate large volumes of data, offering opportunities to develop mechanisms for rapidly identifying priority consumer zones based on key parameters. This research concentrates on two such parameters: geospatial coordinates (longitude and latitude) and water consumption volumes. These variables allow for the organization and territorial analysis of water supply system components. It should be emphasized that additional factors—such as population density and the nature of industrial facilities—represent areas for future exploration. Incorporating these elements could greatly enhance analytical depth and yield a more comprehensive insight into water usage patterns through the application of the methodology developed in this study.

This paper investigates the water supply network of a city with over half a million residents, taking Gomel (Republic of Belarus) as a case study. The primary objective of the research is to propose a methodology for clustering the spatial distribution of objects within the urban water supply system, to accomplish the following goals:

1. Creating a methodology for dynamic spatial analysis that defines boundaries of territorial clusters using water consumption data alongside the geographic coordinates of objects. This approach will accommodate ongoing changes in evolving infrastructure by utilizing a steady stream of data to delineate service zones for water supply networks and improve their overall manageability.
2. Formulating a strategy for the optimal positioning of sensors (such as those for pressure, flow rate, and similar metrics) across the water supply network, ensuring their locations correspond to the identified territorial clusters and allowing for the visualization of their even distribution on a map. This will facilitate efficient gathering of data on the hydraulic characteristics of the system—from water supply origins to monitoring points within the urban environment. The acquired data will serve as the basis for constructing hydraulic models to optimize system performance and examine the impact zones of water intake facilities across different segments of the water supply network.

The anticipated results of the study include establishing a methodology that not only enhances the manageability and monitoring of the current water supply system but also provides a foundation for developing hydraulic models (digital twins) within the framework of smart city initiatives. The research outcomes constitute an important preliminary advancement toward modeling energy-efficient operating regimes for water supply networks and optimizing pressure regimes at pumping stations. Furthermore, the results can support planning for infrastructure upgrades and the formulation of resource management approaches in other cities featuring comparable water supply configurations.

Related Works

Clustering algorithms

An examination of international practices reveals that hierarchical clustering and the k-means algorithm rank among the most commonly employed clustering techniques. For example, the work by Poncelet *et al.* [6, 7] illustrates the application of these methods to identify representative historical days within models for optimizing energy consumption and generation. The authors introduced novel criteria and metrics for assessing model representativeness and devised an optimization technique that more precisely captures annual fluctuations in demand and energy output. Similarly, Arefifar *et al.* [8] employed comparable techniques to form virtual microgrids in intelligent distribution systems. This enhanced energy system oversight by incorporating probabilistic aspects of generation and consumption, thereby optimizing power allocation among users [9].

Integrating these clustering techniques with geographic information yields a strong combined effect, allowing for the successful handling of diverse challenges. In one instance, Zhang *et al.* [10] applied spatial clustering methods to evaluate energy-efficiency variations across different regions of China. Their analysis revealed a pattern of consumer grouping by energy usage intensity. To explore underlying influences on energy efficiency—such as industrial focus and levels of economic growth—a spatial quantile autoregression model was utilized [11]. This method not only pinpointed critical determinants but also laid the groundwork for crafting specific policies to boost energy efficiency.

Clustering techniques are also increasingly used to address reliability issues in water supply systems. A notable case is the research by Abokifa and Sela [12], which centers on spatial data analysis integrated with predictive modeling of pipeline breakdowns. The authors' proposed framework segmented the pipeline network by geographic location and pipe attributes, thereby highlighting zones with elevated failure risk. Employing machine learning to forecast failure likelihoods within each cluster improved the accuracy of incident predictions and the scheduling of preventive actions [13].

As evidenced by the literature survey, territorial clustering has demonstrated considerable success in managing various resource-related systems. It not only promotes optimal resource distribution and greater system control but also serves as a stepping stone for incorporation with sophisticated artificial intelligence and machine learning tools.

Water resource management in smart cities

Worldwide developments in water supply systems are grounded in the smart city paradigm, emphasizing the deployment of cutting-edge technologies that support ongoing collection and analysis of extensive datasets through artificial intelligence (AI) and machine learning (ML) to tackle specialized problems [14–16]. Until recently, the substantial cost of connecting numerous dispersed water supply facilities across large urban areas has restricted the adoption of innovative solutions. Nevertheless, recent advances in specialized low-bandwidth communication protocols for telemetry equipment have opened new possibilities. The rise of Industry 4.0 initiatives and smart city programs has made acquiring data on hydraulic network conditions far more straightforward, significantly advancing numerous processes, including municipal water distribution. Internet of Things (IoT) solutions have not only simplified data acquisition and evaluation but have also opened extensive avenues for discovering novel methods to enhance the efficiency of natural and energy resources [17–19].

Issues concerning the management of pipeline systems remain highly pertinent, with AI models capable of successfully tackling tasks such as predicting pressure at monitoring points, optimizing operational processes, anticipating failures, and identifying leaks. In these applications, high-quality information resources are essential, providing precise, up-to-date data required for training and validating the models. A representative case is the work by Aljameel *et al.* [20], which investigated the use of machine learning techniques in pipeline systems. The study assessed the performance of recurrent neural networks (RNNs) and the k-nearest neighbors (KNN) algorithm for detecting leaks. Through a comparative evaluation of different ML approaches, it was shown that RNNs outperformed others in these tasks, thanks to their capacity for autonomous learning and handling extensive datasets, leading to superior prediction accuracy in dynamic data settings.

Another relevant investigation [21] demonstrated the use of machine learning for simulating pipeline failure conditions under elevated pressure. The study employed predictive models built on artificial neural networks (ANN), gradient boosting (XGBoost, XGB), and categorical boosting (CatBoost, CAT). These models were developed using data from real-world pipeline inspections and outcomes from finite element analyses conducted for burst forecasting. Among the tested algorithms, CatBoost delivered the highest prediction accuracy with the lowest error rates. This underscores the value of integrating AI techniques with traditional engineering methods, positioning them as robust tools for evaluating pipeline integrity, especially in scenarios involving fluctuating operational conditions that demand adjustments to technical settings.

Further exploring machine learning applications in water supply systems, Sourabh *et al.* [22] opened new avenues for research in their publication in the ISH Journal of Hydraulic Engineering. The paper focuses on controlling water leaks and outlines the implementation of reverse-engineering strategies for detection. Its core concept involved using classification and regression problems with artificial neural networks and support vector machines (SVMs) to detect pressure or flow anomalies attributable to leaks. The analysis was performed under two conditions: one relying solely on pressure measurements and the other on water flow measurements within the system. The researchers constructed and trained multilayer perceptron (MLP) models along with multi-level SVMs for both classification and regression. Findings indicated that ANN models outperformed SVMs in leak identification across both scenarios. The work emphasizes that refining the volume of input data used for training can further boost model effectiveness, making this a key direction for ongoing research.

In the same vein, the study by Momeni and Piratla [23] introduced a streamlined hydraulic model for water supply networks that incorporates pipeline condition assessment derived from real flow and pressure monitoring data, combined with machine learning techniques. Applied to a standard water distribution network benchmark, this method proved successful. The research demonstrates how contemporary machine learning tools can improve the interpretation of hydraulic behaviors and enable ongoing, real-time surveillance of potential emergencies [24, 25].

To maximize the performance of ML models, securing dependable hydraulic network data that faithfully reflects the system's dynamics is indispensable. This requires a solid grasp of ideal locations for deploying data-gathering instruments to track pressure, flow rates, and other vital parameters across urban settings, and of how water-consuming entities cluster territorially, enabling targeted sensor placement that reduces their number while maintaining broad spatial coverage for information systems. Strategically positioning monitoring equipment in large, intricate water supply networks is crucial for obtaining accurate data [26-28]. Accordingly, establishing a territorial clustering methodology that groups objects by geographic position and water usage volume greatly streamlines data interpretation. Beyond aiding optimal sensor deployment, clustering yields deeper insights into regional water consumption behaviors, pinpointing areas of high demand and priority monitoring zones. Amid ongoing infrastructure expansion, this strategy proves indispensable for bolstering system oversight and control, enabling visualization of data density patterns on maps. Consequently, this supports finer calibration of pumping station operations and enhanced pressure optimization within the water supply network.

Materials and Methods

The research was organized into several sequential stages. The initial phase entailed an introductory examination of the city's infrastructure and water supply operations, with emphasis on the spatial positioning of water abstraction points, production capacities, and prevailing pressure conditions. The subsequent phase focused on gathering and preprocessing data for analysis, encompassing records of water delivery by abstraction source and georeferenced consumption figures for principal urban consumers. Intermediate steps included data digitization, refinement, and consolidation into a cohesive information platform. In the concluding phase, water-consuming entities were aggregated into clusters based on spatial attributes using hierarchical clustering, yielding segments differentiated by location and consumption characteristics. This process facilitated the recognition of territorial groupings exhibiting distinct geographic

patterns in consumer concentration and highlighted principal clusters subject to varying demands on the water infrastructure.

Sources of statistical information and methods of data processing

The water supply system in Gomel features multiple primary water intakes, each serving both residential and industrial users. The distinct features and operational specifics of these intakes influence the system's overall performance and necessitate a detailed initial assessment of the infrastructure. For this evaluation, statistical information was derived from hourly water-delivery records to the city's network, sourced from the control room logs of second-lift pumping stations throughout 2023. These records were digitized to enable visualization of load intensity on maps, thereby permitting an appraisal of system performance relative to the geographic placement of water intakes. Complementing this, pump station operations were studied by logging discharge-pipe pressure at 1-minute intervals. This provided a more precise examination of the interplay between water demand and intake pressure profiles over time and across spatial dimensions. Additionally, long-term daily water supply statistics for the city spanning 2017 to 2023 were incorporated to identify enduring trends.

The examination of water supply load distribution encompassed over 250 of the city's largest industrial consumers, accounting for their precise geographic positions. These consumers were chosen because of their substantial share of overall water usage, accounting for 66% of the total balance. As most of these facilities lacked automated water metering systems, a hybrid data collection strategy was adopted, integrating information from customer service divisions with readings from digital metering equipment. Customer service records supplied legal addresses of the consumers along with consumption figures for selected time intervals. To convert these addresses into accurate geographic coordinates, the Yandex. Maps geoservice was employed.

A range of software tools and libraries supported data acquisition, processing, and visualization in this research. Jupyter Notebook served as the primary working environment, enabling incremental code execution and immediate display of intermediate outcomes. Pandas and NumPy libraries handled data manipulation tasks, including importing, cleaning, aggregating, and normalizing datasets for subsequent examination. Temporal and spatial patterns of water consumption and pressure at pumping stations were visualized using the Matplotlib and Seaborn libraries, enabling the plotting of daily profiles and minute-level pressure variations, along with their spatial mapping. Spatial clustering of water supply objects was performed using algorithms from the scikit-learn package. The ConvexHull function from SciPy was applied to generate convex polygons delineating the geographic boundaries of the identified cluster zones.

Preliminary overview of the water supply system

The water supply network under study comprises five principal water abstraction facilities: Sozh, Iput, Korenevsky, Southwest, and Central. These sources collectively deliver water to both residential and industrial users throughout the city. All intakes function within a unified hydraulic regime, ensuring consistent technological parameters for end users. Across distinct areas of the city, each intake exhibits unique production capacity and delivered pressure. **Figure 1** illustrates the day-to-day dynamics of water delivery volumes from each intake across the studied system. To construct this time series plot, a moving-average smoothing technique was applied to minimize the effects of short-term noise and to highlight the prevailing trends in water supply patterns, for clearer visual interpretation.

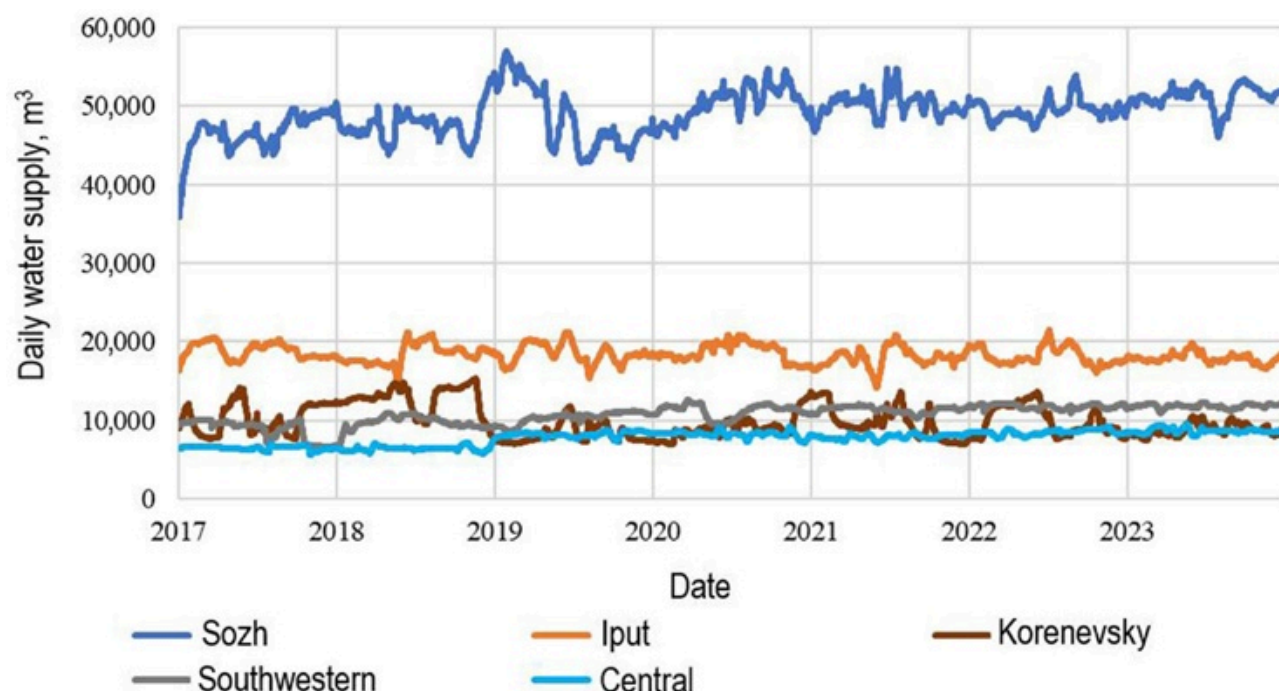


Figure 1. Smoothed data of daily water supply by water intakes for 2023 using the moving average method.

On average, the Sozh intake registers the highest daily water delivery volume (49,192 m³/day). Its coefficient of variation is 7.4%, indicating comparatively stable performance. The Iput, Southwest, and Central intakes also make considerable contributions, with mean daily supplies of 18,358, 10,713, and 7,764 m³/day, respectively. Corresponding coefficients of variation are 8.8%, 13.4%, and 13.3%, likewise indicating relatively consistent output across the city's water supply. In this regard, these four intakes serve as the primary foundations for meeting daily water needs in various urban districts. The remaining intake, Korenevsky, delivers an average of 9,820 m³/day but displays a markedly higher coefficient of variation (25.0%). This characteristic positions it as a key facility for balancing peak demand periods, particularly in adapting to fluctuating consumption in the city's southern districts.

The geographic locations of the water intakes were mapped on the OpenStreetMap platform, enabling clear visualization of each facility's location relative to the city's layout (Figure 2). An interactive 3D map was subsequently generated in Excel with the 3D Maps feature to depict the relative intensity of water supply from each source. This mapping framework facilitated georeferencing of water consumers, enabling the creation of a heat map that visually represents the spatial distribution of load across the water supply system. Through these instruments, it became feasible not only to evaluate the spatial arrangement of water abstraction points but also to seamlessly integrate their operational data with other datasets for advanced geographic zoning and clustering of consumption patterns. As a result of this visual assessment, the proximity of each intake to major consumption centers and the effective utilization rates of each intake were determined. For instance, although the Sozh intake—positioned in the northeastern sector—demonstrates substantially greater actual output than other sources, its distance from the primary consumption zones results in an average utilization rate of just over 50% of its nominal design capacity. Conversely, intakes located within or near industrial zones achieve considerably higher utilization rates, ranging from 74% to 86%.

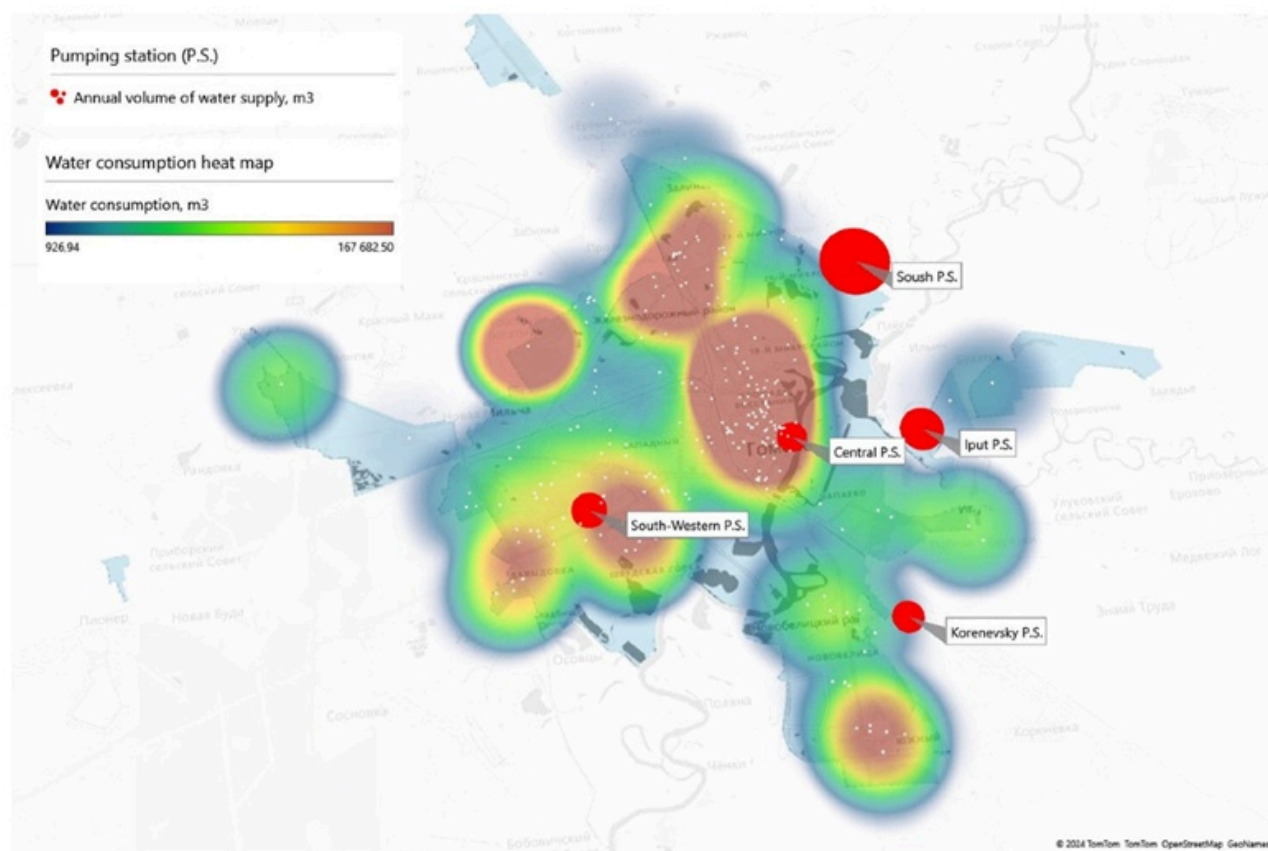


Figure 2. Location of water intakes in Gomel city with labeling of water supply intensity and a heat map of water consumption.

The distance between water supply sources and major consumption zones largely shaped the prevailing operating regimes of the intakes, resulting in noticeable variations in the pressure levels maintained throughout the urban network. **Figure 3** presents the daily pressure profiles recorded at the second-lift pumping stations, clearly demonstrating how geographic positioning affects pressure maintenance. The blue line corresponds to the ‘Sozh’ intake, which consistently maintains the highest pressure range (approximately 0.42–0.44 MPa) throughout the monitoring period. By comparison, the Korenevsky intake (orange line) operates at lower pressures of 0.26–0.30 MPa, while the Southwest intake (yellow line) maintains pressures of 0.33–0.36 MPa.

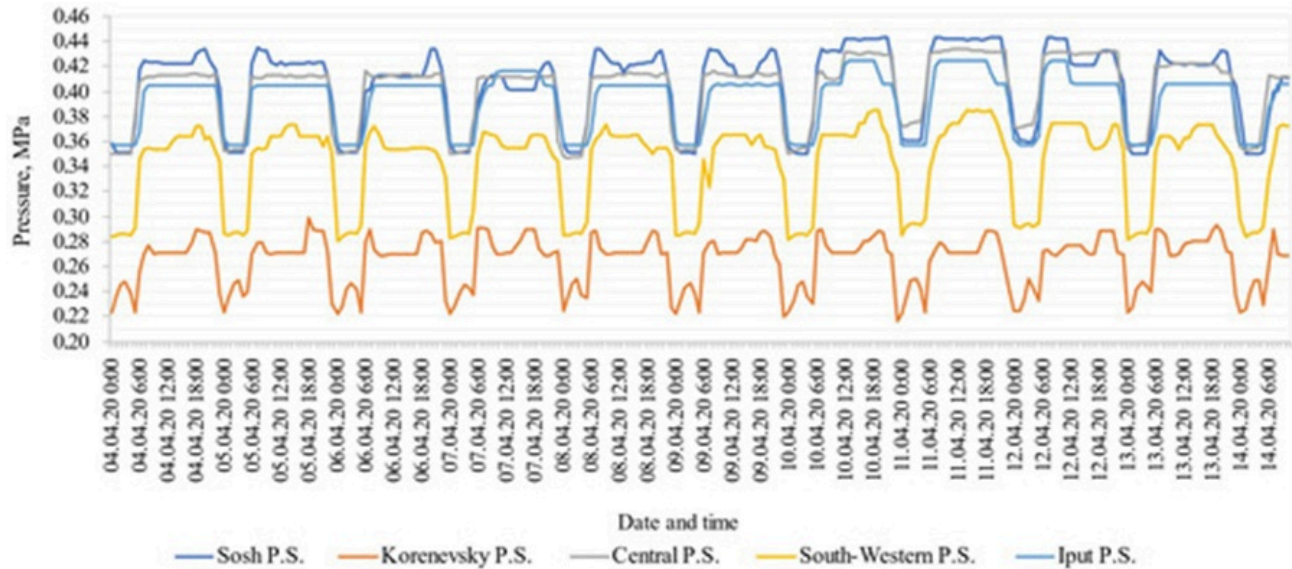


Figure 3. Pumping station (P.S.) pressure schedules of water intakes for the period from 04.04.2023 to 14.04.2023.

Methods of spatial clustering of water consumers

The initial review of water intake locations and their performance parameters revealed a strong correlation between geographic placement and the resulting load on the water supply system. This observation underscored the need for a refined clustering strategy to pinpoint zones of elevated water demand. To support improved analysis and more effective management of the water supply, spatial clustering techniques were implemented. Two distinct clustering strategies were employed in the study. The first strategy considered only the geographic coordinates of water-consuming facilities, enabling a focused examination of spatial patterns in water usage and the detection of regions with particularly intense demand. This approach yielded valuable insights into the territorial arrangement of consumption, highlighting critical areas exerting significant pressure on the city's infrastructure. The second strategy integrated both facility locations and corresponding water consumption volumes, thereby not only identifying the neighborhoods with the greatest usage but also mapping the spatial spread of high-consumption sites across the urban area.

Both strategies applied hierarchical (agglomerative) clustering. Specifically, Ward's method [29] was selected for data grouping in this research. This technique merged sites according to their spatial proximity and similarity in water consumption levels, producing clusters that closely aligned with the established understanding of the city's water supply configuration. Following Ward's algorithm, at each iteration, the pair of clusters whose merger resulted in the smallest increase in the total within-cluster sum of squared deviations was combined. This process ensured the formation of maximally homogeneous groups of water-consuming facilities [30]:

$$\Delta = \sum_i (x_i - \bar{x})^2 - \left(\sum_{x_i \in A} (x_i - \bar{a})^2 + \sum_{x_i \in B} (x_i - \bar{b})^2 \right), \quad (1)$$

x_i — coordinates of the i -th object (latitude and longitude);

$\bar{a}$, $\bar{b}$, $\bar{x}$ — centers of clusters A, B, and the merged cluster, respectively.

Clustering performance was evaluated using the silhouette coefficient, which quantifies the degree to which each object belongs to its assigned cluster relative to neighboring clusters [31]:

$$s(i) = \frac{b(i) - a(i)}{\max(b(i), a(i))}, \quad (2)$$

Where (i) — the mean distance from the i -th facility to all other facilities within the same cluster (intra-cluster distance);

(i) — the smallest mean distance from the i -th facility to any facility in a different cluster (nearest inter-cluster distance).

The silhouette coefficient is computed as the normalized difference between inter-cluster and intra-cluster distances, divided by the larger of the two. Values approaching 1 signify a strong clustering structure and clear separation among groups. Notably, a clear understanding of the optimal partitioning of water-consuming objects across the map enables better alignment of water abstraction management with actual demand patterns [32, 33]. At the same time, the effectiveness of the clustering process and the feasibility of dynamically adjusting cluster boundaries depend heavily on the level of automation in the water supply system, particularly with respect to real-time access to consumption data. Incorporating spatial analysis into water management practices, therefore, calls for adopting advanced technologies for data acquisition and processing [34].

Results and Discussion

Integration of spatial analysis into water use management

Through this study, a formalized algorithm has been developed for the territorial clustering of water supply consumers. The algorithm consists of several sequential stages designed to organize data, associate them with geographic coordinates, and assign objects to distinct clusters:

Step 1: Collection and aggregation of water consumption data. This stage gathers information from multiple sources, such as automated metering installations and records maintained by customer service units, capturing monthly or quarterly consumption volumes for individual facilities throughout the city. All collected data are standardized into a consistent format suitable for analysis and aggregated to an appropriate granularity level. The chosen sample size depends on available computational resources and the feasibility of acquiring representative statistical information. When resources are constrained, it is advisable to prioritize the largest consumers distributed across various parts of the city to maintain representativeness of the dataset.

Step 2: Geocoding of addresses. In many instances, subscriber management systems do not provide direct export of latitude and longitude coordinates, which are essential for spatial clustering, creating obstacles to real-time geographic data analysis. Consumer locations are typically available only as legal addresses, necessitating an intermediate geocoding procedure. This involves transforming textual addresses into precise geographic coordinates (latitude and longitude) through external geocoding platforms such as Google Maps or Yandex.Maps. The outcome is a unified database containing both the geographic coordinates of each consumer and their associated water consumption volumes, enabling accurate placement of facilities on a map and preparing the dataset for subsequent analytical and clustering procedures.

Step 3: Aggregation of data by address. To minimize computational demands, water consumption records are aggregated at the building or street level. This consolidation combines data from multiple consumers sharing the same address while retaining essential usage metrics and substantially reducing the dataset size. When the dataset encompasses both industrial and residential users, aggregation is particularly valuable, converting apartment-level measurements into building- or neighborhood-level summaries, yielding a more general view of consumption patterns. These aggregated values support visual interpretation through the creation of water consumption heat maps, offering an intuitive depiction of load distribution across the urban area and the relative distances from water abstraction points. In the system examined

here, the heat map (**Figure 2**) clearly highlights zones of intense demand (red), alongside regions of moderate (green) and low (blue) consumption.

Step 4: Spatial clustering. Using geocoded coordinates and aggregated consumption figures, hierarchical clustering is performed to group facilities based on their spatial proximity and water-use intensity. The resulting classification labels each data point according to its assigned cluster. These labeled points then serve as the foundation for generating convex hulls that enclose the outermost points of each cluster, thereby defining territorial boundaries characterized by varying levels of water demand.

The application of the described algorithm is illustrated below using the water supply system under investigation.

Hierarchical clustering and finding the optimal number of clusters

Initial visual inspection of the dendrogram (**Figure 4**), constructed from geographic coordinates (latitude and longitude), revealed clear groupings among data points, suggesting natural divisions. To determine the number of clusters automatically, a cut-off technique was applied by drawing a horizontal line at a level corresponding to a prominent gap in the dendrogram structure. This method identified four distinct clusters, and the validity of this division was further assessed using the silhouette coefficient. Within the dendrogram, the four clusters are distinguished by color: orange for cluster 1, green for cluster 2, red for cluster 3, and purple for cluster 4. These color assignments were carried forward into the spatial visualization of the data, facilitating a clearer understanding of each cluster's geographic positioning relative to the city's coordinate system. This combined dendrogram-and-map representation enhances the interpretability of clustering outcomes and strengthens the linkage between the hierarchical structure and the real-world spatial distribution.

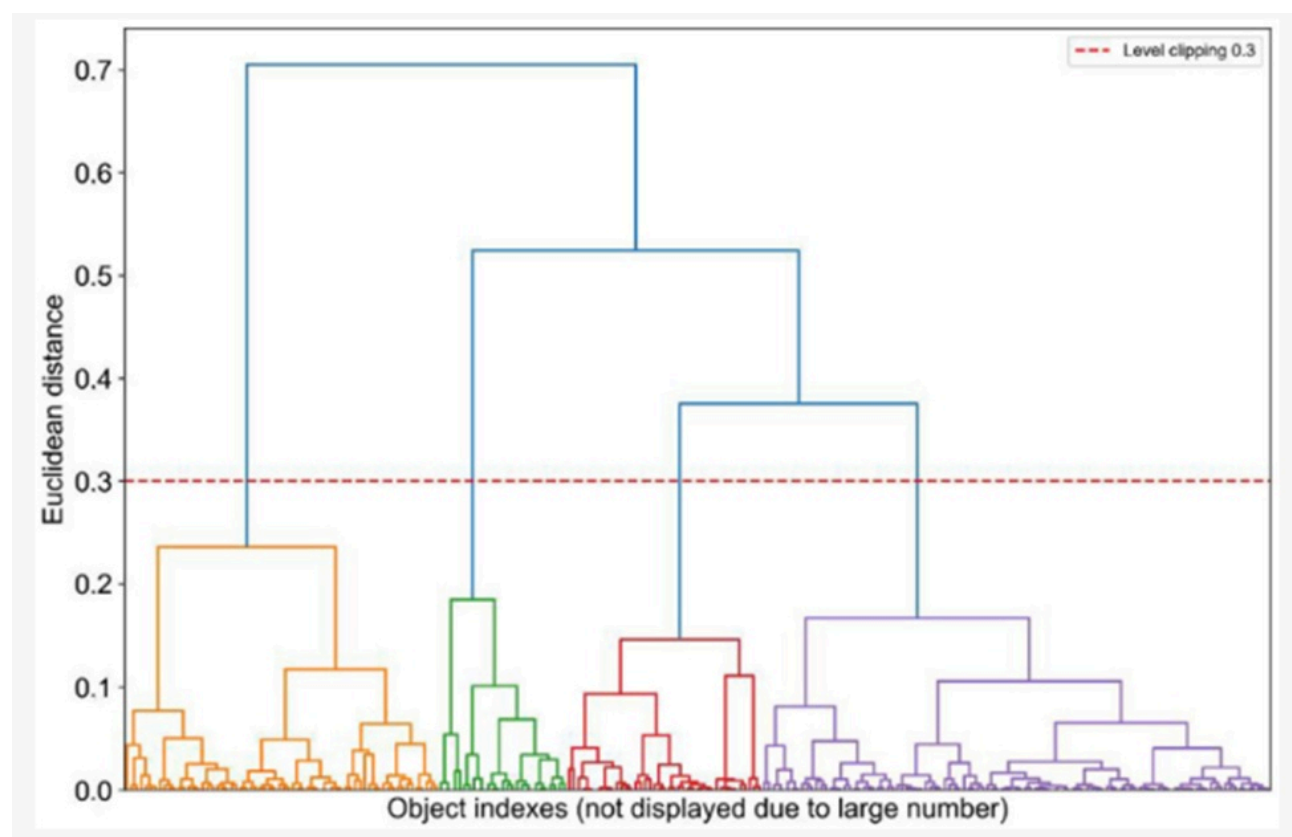


Figure 4. Dendrogram by geographical location of water consumption facilities.

The analysis verified that the maximum silhouette coefficient value (0.499) is obtained when the dataset is partitioned into four clusters. These elevated coefficient values support the selected partitioning scheme, reflecting strong cohesion within each cluster and clear separation between them. When the water consumption volume parameter was incorporated alongside geographic coordinates, the cluster structure expanded from 4 to 6 groups, accompanied by a reduction in the silhouette coefficient to 0.432. This decline suggests increased complexity in the data distribution when an additional variable is included in the geocustering procedure. **Table 1** presents the computed silhouette coefficient values for clustering based on location alone and on location combined with water consumption levels.

Table 1. Results of calculating the silhouette coefficient for various clustering methods.

Number of clusters	Silhouette coefficient (based on geographic location only)	Silhouette coefficient (based on geographic location and water consumption volume)
2	0.497	0.354
3	0.499	0.418
4	0.485	0.425
5	0.496	0.432
6	0.414	0.400
7	0.433	0.398

Figure 5 illustrates how the silhouette coefficient varies with the number of clusters when segmenting the data based on both object locations and water consumption volumes.

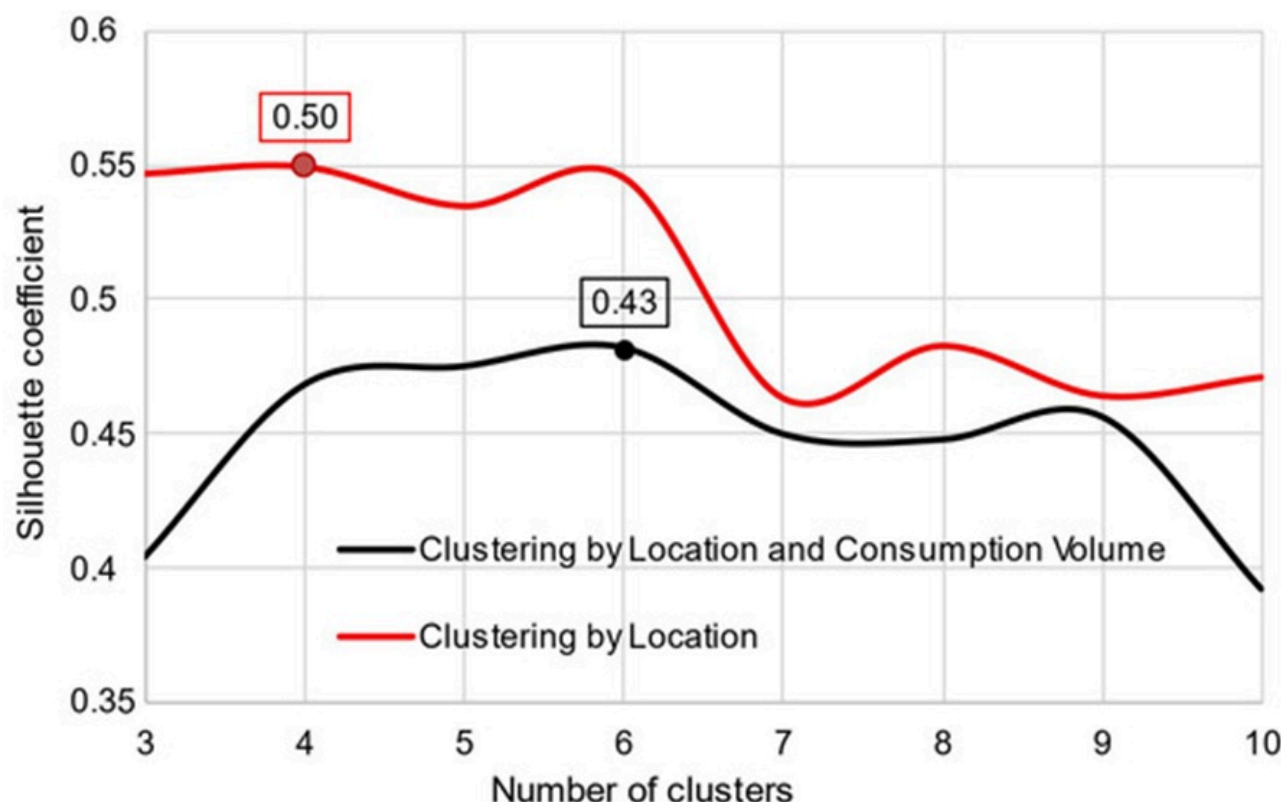


Figure 5. Change in silhouette coefficient with increasing number of clusters.

Segmentation of water consumption facilities

Using the derived clusters, water-consuming objects were mapped onto a coordinate system, enabling clear visualization of consumer groupings based on their spatial distribution. Cluster boundaries were delineated by applying the convex hull algorithm from the ConvexHull class in the SciPy library, which connects the outermost points of each group. This method effectively outlined the territorial extents of the clusters (Figure 6). Importantly, the division into four clusters proved justified not only in terms of spatial optimization on the map but also in alignment with the pipeline network's actual service zones. For instance, Cluster No. 1, positioned close to the Southwest water intake, occupies the southwestern sector of Gomel and encompasses residential and industrial zones within the Sovetsky district, which feature well-developed social infrastructure. This cluster is notable for a higher proportion of healthcare facilities and educational institutions, reflecting intensive mixed-use (residential and industrial) activity in the area. It accounts for 27.4% of the total number of water consumers analyzed.

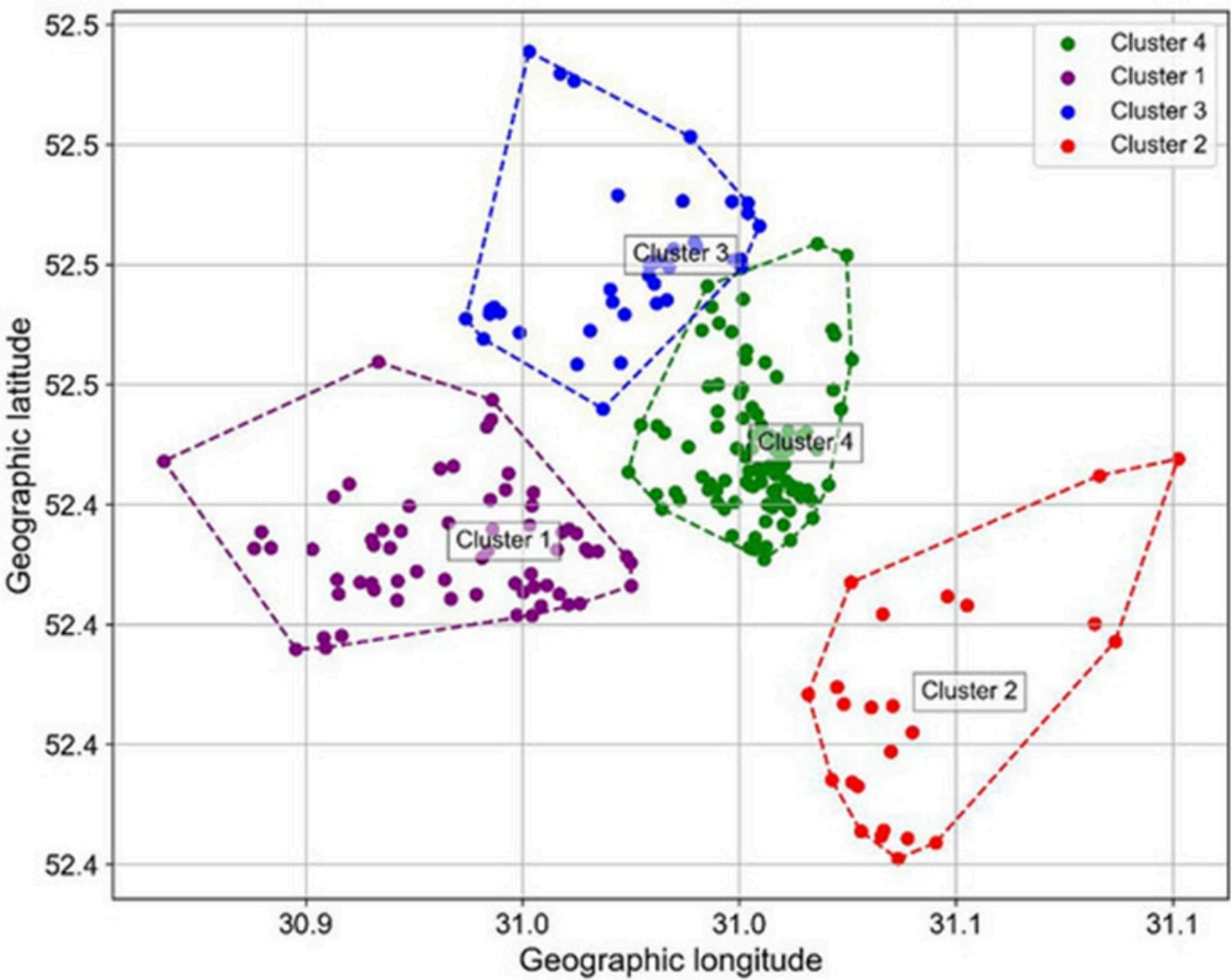


Figure 6. Clustering of water consumption objects according to geographical data.

Cluster No. 2, situated in the southern part of the city and supplied mainly by the 'Korenevsky' and 'Iput' intakes, corresponds to the Novobelitsky district. This zone exhibits low facility density (11.2% of the total), extensive residential neighborhoods, suburban territories, and only a limited number of industrial or social institutions connected to the municipal water network. Cluster 3, located in the northeastern section within the Zheleznodorozhny district and primarily served by the Sozh intake, displays comparable traits, representing 17.0% of the total facilities considered.

Cluster 4, centered in the urban core and fed by the Central water intake, stands out for its highest density and greatest variety of infrastructure elements. It houses major institutions, including central universities and the city's primary retail complexes. Accounting for 44.4% of all properties examined, Cluster 4 demonstrates the most concentrated water demand.

It is noteworthy that incorporating water consumption volumes into the clustering procedure generated two additional clusters beyond those identified using geographic coordinates alone. One of these new clusters grouped the largest individual consumers, including a food production facility with approximately 305,000 m³/year and molding enterprises consuming around 270,050 m³/year. The other comprised mid-sized high-volume users ranging from about 99,563 to 142,000 m³/year, such as hospitals and correctional institutions. These outcomes indicate that the proposed territorial analysis methods possess a degree of universality and adaptability, depending on the parameters considered. Consequently, the resulting zoning of water consumption areas can be applied to other cities of comparable population size and water supply configuration, such as Irkutsk or Vladikavkaz. In particular, Vladikavkaz relies on underground sources from the Ordzhonikidzevskoye field, utilizing dedicated water intake facilities. By comparison, Gomel's water supply system operates with five water intakes, exhibiting structural similarities to that of Vladikavkaz.

Application of clustering results for optimal installation of pressure sensors

Establishing effective optimization mechanisms for water supply operations requires continuous, reliable data on the hydraulic conditions within the urban network. To achieve this, it is essential to precisely identify suitable locations throughout the city for installing recording devices and to determine the territorial zones each device will monitor. The clustering of Gomel into distinct zones enabled the prioritization of areas for pressure sensor deployment, significantly streamlining the planning and expansion of the monitoring infrastructure. Currently, more than 20 pressure sensors are installed across Gomel, positioned at critical control points within the water supply system and at the outlets of pumping stations. The figure illustrates the spatial distribution of existing pressure sensors relative to the preliminary clustering of registered water consumption facilities. The map (**Figure 7**) displays the outer boundaries of the four clusters derived from the water consumption analysis, with all sensor locations—both at water station (WS) and pumping station outlets, and at city control points (CP)—clearly indicated.

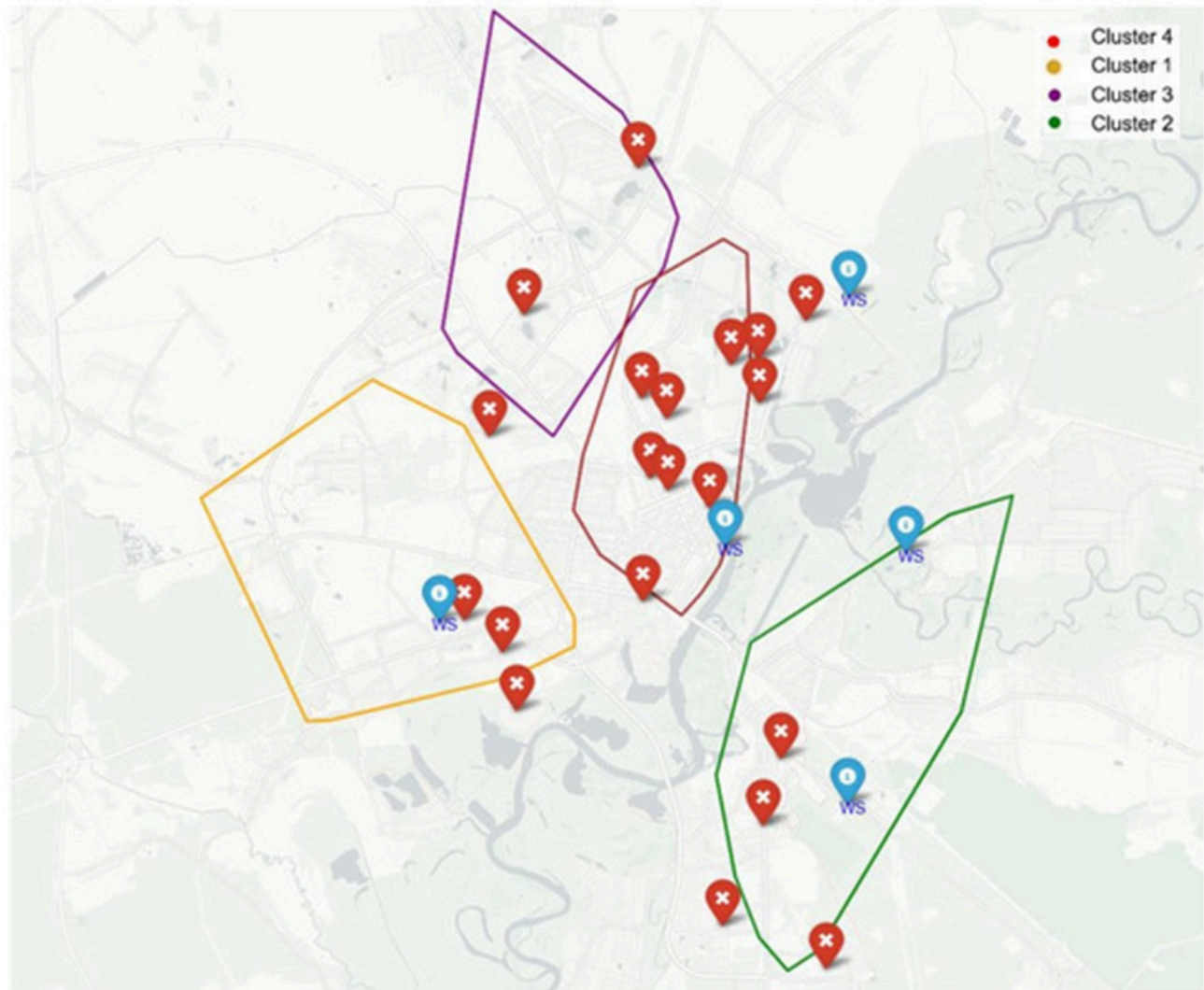


Figure 7. Construction of water consumption cluster boundaries and pressure sensor location points: WS—water station.

The placement map reveals that certain sensors lie outside the defined cluster boundaries. This occurs because the constructed cluster limits were based primarily on commercial and industrial water consumption facilities, excluding residential housing stock due to constraints in the available dataset. Despite this limitation, the current sensor layout remains valid overall and provides adequate coverage of key segments of the municipal water supply network. The distribution of sensors varies noticeably across clusters. For instance, Cluster 1 in the southwestern district requires additional sensor installations to achieve uniform pressure monitoring. In contrast, the central district (Cluster 4) is already relatively well-equipped with sensors, reducing the need for additional devices in this zone compared to other areas.

The overarching objective of this study is to establish a hydraulic modeling framework for the water supply system of Gomei. To support this goal, an IoT-based pressure monitoring platform is under development that leverages Internet of Things technology. The platform uses the MQTT protocol to facilitate efficient data exchange, enabling seamless collection and transmission of pressure sensor readings. This infrastructure also creates opportunities for advanced research, including the construction of more precise predictive models using neural network techniques. **Figure 8** presents the modeled pressure profile at one of the city network's control points alongside the corresponding measured (actual) pressure values. The model input data consist of pressure readings from the pumping stations of the studied water intakes, while the output corresponds to predicted pressures at designated system control points.

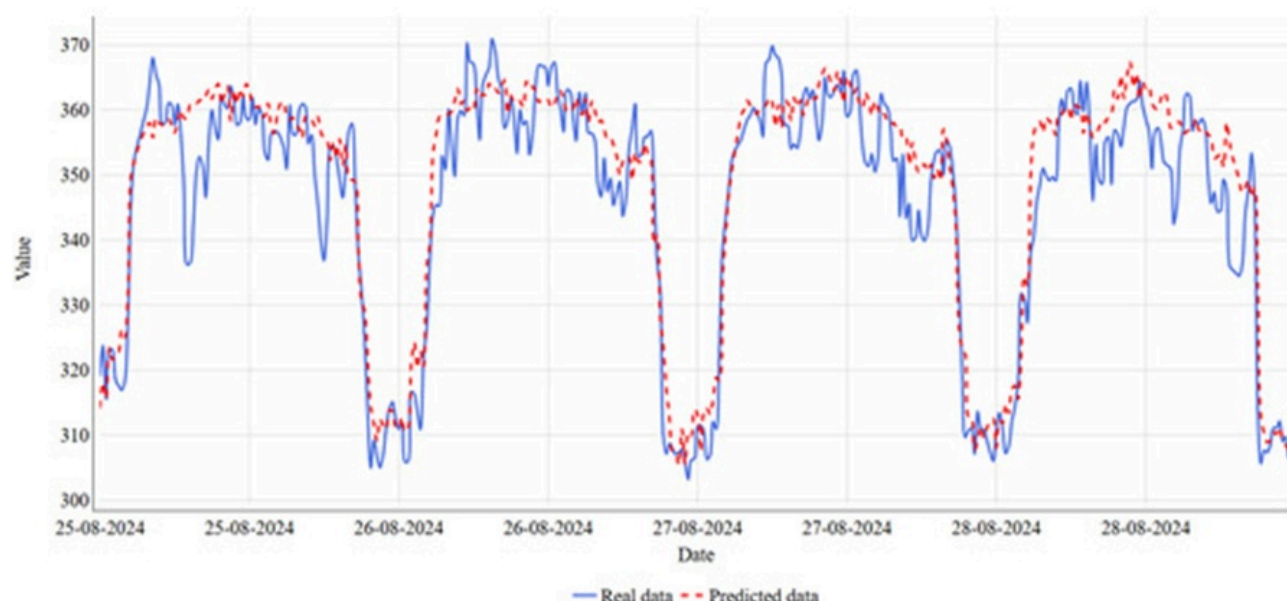


Figure 8. Pressure modeling at control points using a multilayer perceptron model.

The preliminary comparison demonstrates the strong predictive performance of the multilayer perceptron model, as evidenced by the following evaluation metrics computed for one of the reference points: the mean absolute error (MAE) was 5.07 kPa, the root mean square error (RMSE) was 6.57 kPa, and the mean absolute percentage error (MAPE) was 1.48%. The mathematical foundations of model development and the specific characteristics of hydraulic modeling constitute subjects for ongoing investigation and will be explored in greater depth in subsequent scientific publications.

Conclusion

The application of Ward's method for spatial clustering of water consumption facilities identified four primary clusters based on the geographic distribution of objects, using the city of Gomel as a case study. The resulting clustering demonstrated strong spatial optimization: 44.4% of the water consumption facilities are concentrated in the central district (cluster 4). In comparison, the southwestern sector (cluster 1) encompasses 27.4% of facilities, reflecting a balanced mix of residential and industrial consumers. The southern zone (cluster 2) exhibits the lowest density at 11.2%, and the northeastern area (cluster 3) accounts for 17% of the facilities, underscoring the varying degrees of water demand and infrastructure distribution across different parts of the city.

The use of agglomerative clustering in this research facilitated the grouping of more than 200 industrial facilities based on their geographic locations and water consumption volumes. Incorporating the water consumption volume parameter into the clustering process further refined the data structure, leading to the emergence of two additional clusters. Certain high-volume users stood out as particularly resource-intensive—for instance, a confectionery factory consuming approximately 305,000 m³/year and metal casting enterprises using around 270,050 m³/year—highlighting substantial infrastructure burdens within those clusters. Including consumption volumes enabled a more granular segmentation of facilities and pinpointed zones of elevated resource demand, which is especially valuable for optimizing both water and energy management strategies. Nevertheless, due to constraints in the available data, this study did not incorporate variables such as population density or specific types of industrial operations, factors that could substantially influence the outcomes. Addressing these additional characteristics forms a key direction for future work, to refine the methodology and develop more precise water consumption zoning applicable to a broader range of urban environments.

The findings of this research demonstrate that agglomerative clustering, when supported by precise geospatial information, is a robust tool for distinguishing consumers based on their distinctive attributes. This capability creates new

opportunities to enhance the optimization and operational efficiency of supply systems. Within the framework of smart city initiatives, such approaches assume critical importance by enabling the seamless integration of resource-use data with other components of urban infrastructure, thereby fostering the development of more sustainable, reliable, and energy-efficient systems.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 08 Mar 2022

Revised: 27 Apr 2022

Accepted: 18 Jun 2022

Published online: 20 September 2022

Rights and permissions

Open Access The author(s) retain copyright. This article is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. It may be shared and adapted for non-commercial purposes with appropriate attribution, an indication of changes, and distribution of adaptations under the same license. Third-party material may be subject to separate terms identified in its credit line. View the license at <https://creativecommons.org/licenses/by-nc-sa/4.0/>.

References

- Twomey KM, Webber ME. Evaluating the energy intensity of the US public water system. *Energy Sustain*. 2011;54686:1735–48.
- Zapata O. More water please, it's getting hot! The effect of climate on residential water demand. *Water Econ Policy*. 2015;1:1550007.
- Bezerra BG, Lee TS. Crop evapotranspiration and water use efficiency. In: *Irrigation Systems and Practices in Challenging Environments*. Rijeka: IntechOpen; 2012. p. 57–76.
- Chupin RV, Pham MN, Chupin VR. Optimal reconstruction of water supply network under conditions of reduced water consumption. *E3S Web Conf. EDP Sci*. 2019;102:03005.
- Fracasso PT, Barnes FS, Costa AHR. Energy cost optimization in water distribution systems using Markov decision processes. In: *Proceedings of the 2013 International Green Computing Conference*. IEEE; 2013. p. 1-6.

Poncelet K, Höschle H, Delarue E, Virag A, D'haeseleer W. Selecting representative days for capturing the implications of integrating intermittent renewables in generation expansion planning. *IEEE Trans Power Syst*. 2016;32:1936-48.

Poncelet K, Delarue E, Six D, Duerinck J, D'haeseleer W. Impact of the level of temporal and operational detail in energy-system planning models. *Appl Energy*. 2016;162:631-43.

Arefifar SA, Mohamed YAR, El-Fouly THM. Supply-adequacy-based optimal construction of microgrids in smart distribution systems. *IEEE Trans Smart Grid*. 2012;3:1491-502.

Arefifar SA, Mohamed YAR, El-Fouly THM. Optimum microgrid design for enhancing reliability and supply-security. *IEEE Trans Smart Grid*. 2013;4:1567-75.

Zhang J, Lu Q, Guan L, Wang X. Analysis of factors influencing energy efficiency based on spatial quantile autoregression: Evidence from panel data in China. *Energies*. 2021;14:504.

Zheng J, Mi Z, Coffman DM, Milcheva S, Shan Y, Guan D, et al. Regional development and carbon emissions in China. *Energy Econ*. 2019;81:25-36.

Abokifa AA, Sela L. Integrating spatial clustering with predictive modeling of pipe failures in water distribution systems. *Urban Water J*. 2023;20:465-76.

Abokifa AA, Sela L. Identification of spatial patterns in water distribution pipe failure data using spatial autocorrelation analysis. *J Water Resour Plan Manag*. 2019;145:04019057.

Ezechi CG, Okoroafor ER. Integration of artificial intelligence with economical analysis on the development of natural gas in Nigeria: focusing on mitigating gas pipeline leakages. In: *Proceedings of the SPE Nigeria Annual International Conference and Exhibition; Lagos, Nigeria;2023*.

Liu W, Chen Z, Hu Y. Failure pressure prediction of defective pipeline using finite element method and machine learning models. In: *Proceedings of the SPE Annual Technical Conference and Exhibition; Houston, TX, USA;2025*.

Fan X, Wang X, Zhang X, Yu X. Machine learning based water pipe failure prediction: the effects of engineering, geology, climate and socio-economic factors. *Reliab Eng Syst Saf*. 2022;219:108185.

Gunde S, Chikaraddi AK, Baligar VP. IoT based flow control system using Raspberry PI. In: *Proceedings of the 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS); Chennai, India. IEEE: Piscataway, NJ, USA;2017*. p. 1386-90.

Meena KB, Dhanalakshmi S. IoT water management and distribution system for smart city using artificial intelligence. *J Ubiquitous Comput Commun Technol*. 2023;5:339-51.

Gautam G, Sharma G, Magar BT, Shrestha B, Cho S, Seo C. Usage of IoT framework in water supply management for smart city in Nepal. *Appl Sci*. 2021;11:5662.

Aljameel SS, Alomari DM, Alismail S, Khawaher F, Alkhudhair AA, Aljubran F, et al. An anomaly detection model for oil and gas pipelines using machine learning. *Computation*. 2022;10:138.

Yang FP, Zhang L, Luo JH, Zhao XW, Zhang GL. Study on the yield behavior of oil and gas pipe in typical stress states. *Eng Mech*. 2013;30:293-7.

Sourabh N, Timbadiya PV, Patel PL. Leak detection in water distribution network using machine learning techniques. *ISH J Hydraul Eng*. 2023;29:177-95.

Momeni A, Piratla KR. Prediction of water pipeline condition parameters using artificial neural networks. *Pipelines*. 2022;2022:21-9.

Bata M, Carriveau R, Ting DSK. Short-term water demand forecasting using hybrid supervised and unsupervised machine learning model. *Smart Water*. 2020;5:2.

Saroja H, Dharshini S. Deep learning approach for prediction and classification of potable water. *Anal Sci*. 2023;39:1179-89.

Chang NB, Prapinpongsonone N, Ernest A. Optimal sensor deployment in a large-scale complex drinking water network: comparisons between a rule-based decision support system and optimization models. *Comput Chem Eng*. 2012;43:19-9.

Khorshidi MS, Nikoo MR, Sadegh M. Optimal and objective placement of sensors in water distribution systems using information theory. *Water Res*. 2018;143:218-28.

Zhang W, Yang X, Li J. Sensor placement for leak localization in water distribution networks based on graph convolutional network. *IEEE Sens J*. 2022;22:21093-100.

Eszergár-Kiss D, Caesar B. Definition of user groups applying Ward's method. *Transp Res Procedia*. 2017;22:25-34.

Ward JH Jr. Hierarchical grouping to optimize an objective function. *J Am Stat Assoc*. 1963;58:236-44.

Shahapure KR, NicholasC. Cluster quality analysis using silhouette score. In *Proceedings of the 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, Sydney, Australia, IEEE: Piscataway, NJ, USA; 2020. p. 747-8.

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.

Hruntovich NV, Kapanski AA, Baczynski D. Optimization of a variable frequency drive pump working on a water tower. In: *Proceedings of the 2019 International Scientific and Technical Conference Smart Energy Systems*; 18–20 Sep 2019; Kazan, Russia. *E3S Web Conf*. 2019;124:05060.

Sarbu I. A study of energy optimisation of urban water distribution systems using potential elements. *Water*. 2016;8:593.