

Article

Analysis of Spatial Aggregation and Activity of the Urban Population of Almaty Based on Cluster Analysis

Gulnara Bektemyssova ¹, Artem Bykov ^{1,*}, Aiman Moldagulova ², Sayan Omarov ^{1,*},
Galymzhan Shaikemelev ¹, Saltanat Nuralykyzy ¹ and Dauren Umutkulov ¹

¹ Department of Computer Engineering, International Information Technology University, Almaty 050000, Kazakhstan; g.bektemisova@iitu.edu.kz (G.B.); 33271@iitu.edu.kz (G.S.); nuralykyzy.s@gmail.com (S.N.); d.umutkulov@gmail.com (D.U.)

² Department Software Engineering, Institute of Automation and Information Technologies, Kazakh National Research Technical University named after K.I. Satbayev, Almaty 050013, Kazakhstan; a.moldagulova@satbayev.university

* Correspondence: bykov_a_a@list.ru (A.B.); osskenesary@gmail.com (S.O.)

Abstract: This study analyzes the spatial aggregation and activity of the urban population in Almaty using anonymized population density data provided by a telecommunications operator and geographic data from OpenStreetMap. The study focuses on identifying stable zones of high population activity, which facilitates the optimization of transport routes, urban infrastructure planning, and the efficient allocation of city resources. The novelty of this work lies in the integration of aggregated spatiotemporal data with advanced clustering methods, including DBSCAN, KMeans++, and agglomerative clustering. The research methodology involves dividing the city into 500 × 500 m quadrants, calculating normalized population density metrics, and identifying high-activity clusters. Based on a comparative analysis of clustering algorithms, DBSCAN exhibited the highest clustering quality according to the silhouette coefficient and the Davies–Bouldin index, allowing for the identification of key zones of urban activity. The identified clusters were utilized to assess transport load, analyze disparities in the distribution of public transport stops, and develop recommendations to improve public transport accessibility in the most congested areas. The study’s findings are applicable not only to optimizing the transport network but also to addressing a broader range of urban planning challenges, including the strategic placement of infrastructure facilities and the management of population flows. The proposed methodology is scalable and can be adapted to other cities requiring effective tools for analyzing the spatiotemporal activity of urban populations.

Keywords: spatial aggregation; urban population activity; cluster analysis; urban planning; machine learning; geographic information systems



Academic Editor: Lia Bárbara Cunha Barata Duarte

Received: 10 February 2025

Revised: 19 March 2025

Accepted: 2 April 2025

Published: 5 April 2025

Citation: Bektemyssova, G.; Bykov, A.; Moldagulova, A.; Omarov, S.; Shaikemelev, G.; Nuralykyzy, S.; Umutkulov, D. Analysis of Spatial Aggregation and Activity of the Urban Population of Almaty Based on Cluster Analysis. *Sustainability* **2025**, *17*, 3243. <https://doi.org/10.3390/su17073243>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The urban environment is becoming increasingly complex as the number of inhabitants grows, which necessitates the application of modern methodologies for analyzing population dynamics and spatial aggregation to address resource planning and infrastructure optimization challenges. The study of population dynamics plays a key role in designing a sustainable and efficient urban environment, enabling the identification of population distribution patterns and the planning of infrastructure based on real needs. Previous studies [1–3] have employed geographic information systems (GISs) for spatial data analysis in combination with machine learning methods to model population mobility

and develop strategies for urban planning optimization. Specifically, ref. [1] proposed a clustering algorithm based on Pareto theory for territorial stratification and planning; however, this approach does not account for temporal changes in population activity and focuses primarily on static assessments. Machine learning methods, including LSTM, were employed to predict mobility based on time series data, yet spatial patterns of population density were not addressed [2].

Additionally, study [4] underscores the importance of analyzing spatiotemporal patterns of urban mobility using mobile network data. However, it focuses on short-term mobility fluctuations, particularly point-to-point travel dynamics. Such approaches fail to identify stable high-activity zones that persist over extended periods and are essential for sustainable transport and urban infrastructure planning.

Similarly, ref. [5] utilizes mobile phone data to examine passenger behavior in public transportation systems, reinforcing the significance of high-resolution temporal and spatial data for transit planning. However, like other studies, it does not integrate cluster stability analysis, which is crucial for distinguishing temporary high-density areas from persistent urban activity centers. Clustering algorithms such as K-Means and Fuzzy C-Means can be used to identify patterns in mobile data usage [6], which contributes to a better understanding of urban mobility. However, these methods rely on pre-defined assumptions about the shapes and distributions of clusters. This limits their effectiveness in analyzing heterogeneous urban environments. For example, the K-Means algorithm requires a specified number of clusters and is focused on spherical clusters, making it unsuitable for identifying unevenly distributed areas of population density. The Fuzzy C-Means algorithm is also initially focused on static data and is more flexible, which may lead to less accurate results when processing data with a high temporal resolution, such as daily fluctuations in population density. Another critical limitation is the lack of integration with anonymized telecommunications data, which provides a higher temporal and spatial resolution for long-term urban analysis. Although previous studies have successfully clustered mobility patterns, they have not distinguished stable areas of population concentration over time. Our study directly addresses these gaps by using fully anonymized carrier data to estimate the network density in fine-grained urban quadrants. The work combines spatial and temporal clustering techniques to identify persistent zones of urban activity.

Many cutting-edge studies have applied big data and machine learning methods to analyze short-term urban dynamics, leveraging high-resolution datasets to predict mobility trends. To model urban mobility using time series forecasting, an LSTM-based approach [7] can be used; this provides high accuracy in predicting short-term changes in traffic patterns. These models are particularly effective for dynamic traffic management, enabling real-time adjustments to transit networks. However, they primarily focus on immediate temporal fluctuations and cannot identify persistent spatial activity zones, which are essential for long-term urban planning. LSTM-based models are designed to capture sequential dependencies in time-series data, but they do not incorporate spatial clustering mechanisms, limiting their application in detecting stable, high-density urban zones. Thus, approaches like that described in [8] are excellent for analyzing short-term temporal changes but are less effective in identifying stable spatial patterns, which are central to this study. To address these limitations, this study employed cluster analysis methods, which provide robust tools for identifying such patterns.

Three clustering algorithms were selected for the analysis of population activity: DBSCAN, KMeans++, and hierarchical agglomerative clustering [8,9]. These methods were chosen to balance computational efficiency, adaptability to different urban density distributions, and robustness to noise. However, given that urban population density is

highly heterogeneous, methods that assume a homogeneous cluster structure may not accurately reflect actual spatial aggregation.

Thus, a modified K-Means algorithm can be used; this includes noise processing, improving the quality of clusters in urban data [10]. However, even with these enhancements, K-Means-based methods struggle with arbitrarily shaped clusters and require a predefined number of clusters, limiting their flexibility in detecting natural population groupings. This is particularly problematic in high-density urban areas with significant noise, where clusters do not conform to rigid geometric boundaries.

By contrast, DBSCAN is well suited to urban mobility analysis because it identifies clusters of arbitrary shapes without requiring the number of clusters to be pre-set, making it robust to variations in urban density and spatial noise. This capability is particularly important when analyzing datasets with an uneven population distribution and significant noise, as is common in urban environments. Additionally, DBSCAN effectively handles outliers, enabling it to filter noise points that could distort clustering results in high-density areas. This approach has also been successfully applied in other domains [11], where outlier detection is essential for improving model accuracy and anomaly identification.

Hierarchical agglomerative clustering provides a multi-scale perspective on population activity, allowing for a more detailed exploration of clustering structures at different spatial resolutions. Unlike partitioning methods, hierarchical clustering does not require a predefined number of clusters, making it more adaptive to varying densities and urban layouts.

For example, in [9], this method was used to analyze the spatial data of the urban environment, thereby enabling the identification of stable data distribution patterns and a better understanding of their interrelationships. One of the advantages of this approach is its flexibility in processing data with varying densities, as well as its ability to visualize results at different levels of detail. However, a significant disadvantage of the method is its high computational complexity when working with large datasets, which makes it less suitable for tasks that require rapid processing. Hence, these established clustering methods—DBSCAN, KMeans++, and hierarchical clustering—each have distinct strengths and weaknesses when applied to urban mobility data.

Recent studies [12,13] have demonstrated the potential of using mobile data to extract dynamic patterns of urban mobility, providing useful recommendations for improving infrastructure and urban planning. In particular, ref. [13] proposed a visual analytics approach to study population movement patterns using mobile network data, providing a deeper understanding of the spatio-temporal aspects of urban mobility. Similarly, ref. [14] used telecommunication operator data to estimate the population density in Milan, Italy, showing that such data are useful for analyzing population density changes during the day. Unlike previous studies that primarily emphasize short-term mobility fluctuations, our approach is designed to identify stable spatial patterns that are crucial for long-term urban infrastructure development. A major limitation of existing research is the lack of comprehensive population density analysis based on anonymized telecommunication data, which offers a high-resolution perspective on urban dynamics. By leveraging cellular network connection data aggregated at the quadrant level, our methodology enables a more precise and scalable assessment of persistent high-activity zones, providing a robust foundation for strategic urban planning and resource allocation.

However, building modern urban infrastructure is a multifaceted task that requires the integration of various technologies, collaboration among city authorities, the private sector, and society, as well as a careful consideration of the unique characteristics of each city. At the same time, a comprehensive understanding of the spatial aggregation and activity of the urban population becomes a key element of effective urban planning [15].

Analyzing these aspects enables the identification of population distribution patterns, areas of heightened activity, and the optimization of urban resource allocation.

Despite increasing attention to population dynamics in sustainable urban development, managing urban population flows remains underexplored in planning practice. Recent studies show that deep learning models can enhance predictive analytics for urban mobility, addressing the gaps left by traditional approaches [16], yet the potential of such data remains underutilized, especially in space-constrained cities where infrastructure expansion is limited [17]. As demonstrated in a study [18], integrating dynamic control systems can significantly enhance service levels and optimize the utilization of existing infrastructure. Developing and calibrating population flow forecasting models are also crucial steps for improving urban planning and resource allocation [19].

Almaty, as Kazakhstan's largest metropolis, faces rapid population growth and intense migration flows, necessitating advanced tools to analyze and predict population activity for sustainable urban planning [20,21]. Its diverse socio-economic landscape and dynamic urban growth make Almaty an ideal subject for studying the spatial dynamics of the population and for developing practical recommendations to improve city planning and infrastructure. Consequently, the relevance of this study stems from the need to develop effective tools for analyzing and forecasting urban population activity, which will optimize urban infrastructure, enhance the quality of services provided, and promote the sustainable development of the urban environment—factors essential for the social, environmental, and economic well-being of the community.

The application of cluster analysis methods enables a more precise assessment of population density and distribution, as well as the prediction of pedestrian flows in various parts of the city at different times of the day [22]. Unlike approaches that focus solely on mobility dynamics [7], this study emphasizes the identification of stable clusters of population activity, which are critical for long-term urban infrastructure planning. Additionally, it facilitates the identification of critical patterns in transport systems and enhances our understanding of the factors influencing traffic. Employing cluster analysis for population activity data allows for the more accurate forecasting of population dynamics and optimization of urban infrastructure [23].

The scientific novelty of this study lies in its integrated application of clustering methods to spatiotemporal data and the introduction of cluster stability assessment techniques. By focusing on stable activity zones, this approach addresses urban planning gaps and offers practical infrastructure Optimisation tools. Our study focuses on adapting and optimizing modern clustering techniques for the specific characteristics of urban population density and activity data. We evaluated the performance of various clustering algorithms, including DBSCAN, KMeans++, agglomerative clustering, and HDBSCAN, to identify the most stable and accurate clusters within datasets characterized by an uneven density and the presence of noise points.

As a result, this research addresses an existing gap in urban planning practices by offering a practical methodology for assessing and forecasting population activity. This approach is essential for the effective allocation of resources, the enhancement of pedestrian infrastructure, and the overall improvement of the quality of life for city residents.

The aim of this study is to analyze the spatial aggregation and activity of Almaty's urban population using cluster analysis. This approach aims to identify key activity zones and provide practical recommendations for the development of urban environments.

The objectives of the study include:

- The collection and processing of data on population density and activity in Almaty using geographic information systems (GISs) and aggregated data provided by a telecom operator.

- The research and application of cluster analysis methods to identify patterns in population distribution and activity across different areas of the city.
- The evaluation of the quality of identified clusters using metrics such as the silhouette coefficient and the Davies–Bouldin index.
- An analysis of the temporal dynamics of population activity within the identified clusters.

2. Materials and Methods

2.1. Data Collection and Preparation

Our study utilized aggregated and anonymized data provided by one of the leading telecom operators in Kazakhstan. These data reflect subscriber activity within the coverage areas of base stations and record the number of unique connections in various locations across Almaty. The telecom operator provides only anonymized data, aggregated into a 500×500 m grid for privacy reasons, without the exact coordinates of base stations. The use of such a grid allows the maintenance of sufficient accuracy in determining the location of subscribers. The boundaries of the quadrants were obtained from OpenStreetMap (OSM) data, which reflect the urban infrastructure of Almaty and provide unified geospatial reference points.

The analyzed data include information on the total number of unique users in each quadrant; the number of users in “home locations” (when users are present in the evening and at night); and the number of users in work locations (when users are active during the day). The integration of spatial data and user activity information was performed using clustering methods. The results of the analysis helped identify areas of high population activity, which is valuable for optimizing transport routes, locating government facilities, and developing private infrastructure. Figure 1 shows a diagram of the research workflow, demonstrating the integration of telecom operator data and geospatial information for cluster analysis purposes.

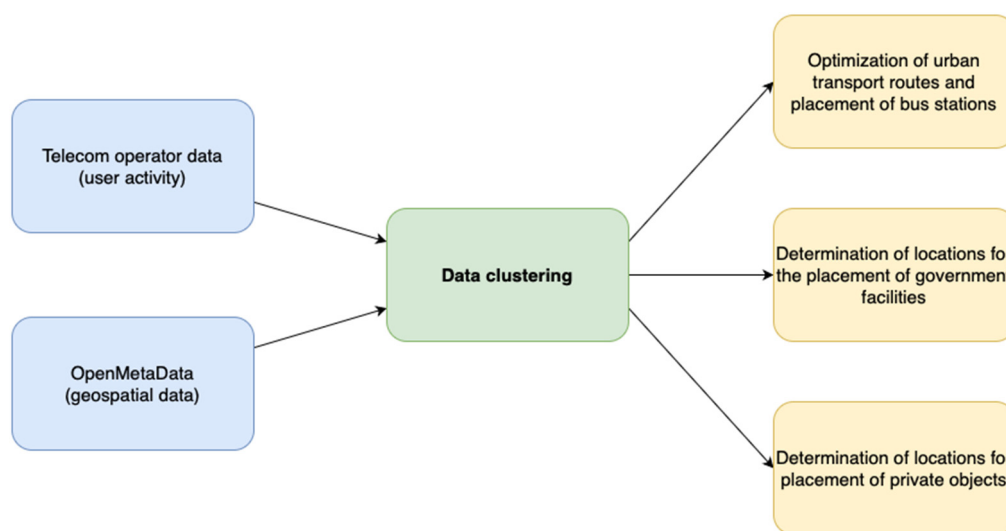


Figure 1. Scheme of integration of telecom operator data and geospatial information for the purposes of cluster analysis and strategic planning.

The stages of data preparation and processing included the following steps:

1. Geographic data on the boundaries of Almaty, its road infrastructure, buildings, and other objects (current as of 7 October 2023) were extracted from OpenStreetMap (OSM). These data served as the basis for visualizing the boundaries of Almaty (Figure 2) and for further spatial analysis.

2. The city area was divided into equal, non-overlapping 500×500 m quadrants using Python 3.11.5 algorithms and geospatial data processing libraries (Figure 3). This approach enables standardized analysis and facilitates data comparison across different parts of the city.
3. This study utilized data provided by a telecommunications operator, which aggregates and anonymizes the data to protect users' personal information. These data were employed to analyze the population density and user activity within each quadrant at various times of day, thereby capturing the true dynamics of changes in the population concentration within the urban environment. Data for the period from 1 March to 1 June 2023 were used (Table 1).

Table 1. Structure of the initial data from the telecom operator.

Field	Description
WEEK_DAY_IND	Working days indicator: 1—weekdays (from Monday to Friday inclusive), 2—weekends (other days)
TIME_HOUR	Aggregation hourly window (22:00—this means the interval from 21:00 to 21:59)
ZID_NUMBER	ZID (quadrant) number
CORNER_1(LAT_LONG)	Location of the latitude of the longitude of the TOP LEFT quadrant
CORNER_2(LAT_LONG)	Location of the latitude of the longitude of the TOP RIGHT quadrant
CORNER_3(LAT_LONG)	Location of the latitude of the longitude of the BOTTOM RIGHT quadrant
CORNER_4(LAT_LONG)	Location of the latitude of the longitude of the BOTTOM LEFT quadrant
NUM_OF_UNIQ_HOME_USERS	Average Unique Customers in Home Location Quadrant
NUM_OF_UNIQ_WORK_USERS	Average number of unique customers in the quadrant with the operating location
NUM_OF_UNIQ_USERS	Average number of unique customers per quadrant

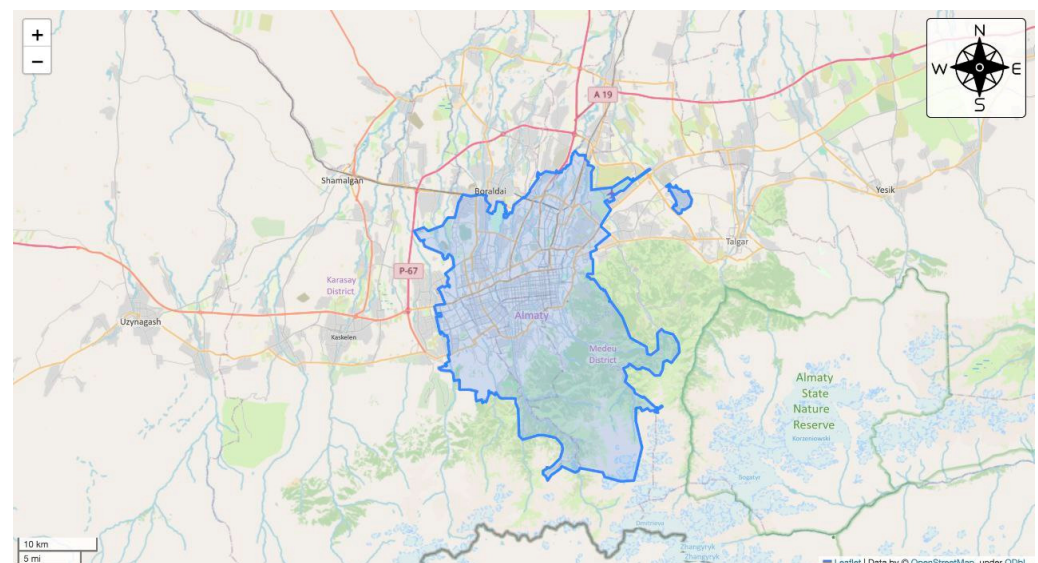


Figure 2. Almaty boundaries (blue area) based on OpenStreetMap data.

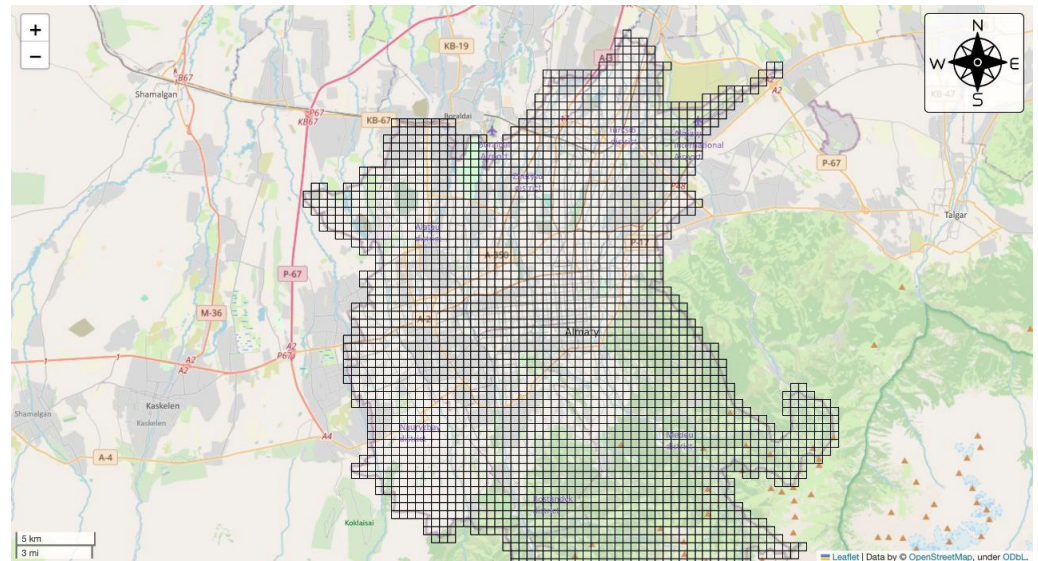


Figure 3. Division of Almaty's territory into 500×500 m quadrants.

2.2. Data Characteristics

The data used in this study possess several specific characteristics that significantly influenced the choice of clustering methods. The population data for Almaty exhibit high variability, with areas of both high and low population density. These data are anonymized and aggregated, providing hourly population load metrics for 500×500 m quadrants across the city without tracking individual movements. This structure ensures user privacy but limits the applicability of advanced spatiotemporal methods such as ST-DBSCAN, which rely on detailed temporal and spatial correlations. The dataset also includes quadrants with abnormally low or high activity, considered outliers or noise points, which were effectively handled by DBSCAN. These anomalies can distort clustering results if not properly addressed during algorithm selection. The urban environment features a complex spatial organization, and the expected clusters may have arbitrary shapes. This complexity arises from the diverse urban infrastructure, which includes residential areas, commercial zones, and transport hubs [24].

2.3. Cluster Analysis of Population Activity and Identification of Load Segments

Following the preliminary processing of the collected data, the most heavily loaded points in the city were identified, and a cluster analysis was conducted based on hourly loads to determine load segments or attraction zones within Almaty. For this purpose, clustering methods from the field of unsupervised learning were applied, and their effectiveness was compared to identify the most suitable algorithm for the dataset.

The research methodology included the following steps:

- For each sector (a 500×500 -m quadrant), centroids were calculated using the geographic coordinates of its boundaries. This approach allowed each sector to be represented as a point in space for subsequent analysis.
- Each centroid was assigned a corresponding sector load, expressed as the number of unique mobile users (NUM_OF_UNIQ_USERS indicator) per hour.
- To ensure the comparability of loads within each hour, a new indicator was introduced: the normalized load value (NUM_OF_UNIQ_USERS_NORMED), which was calculated using the following formula:

$$\text{NUM_OF_UNIQ_USERS_NORMED}_i = \frac{\text{NUM_OF_UNIQ_USERS}_i}{\max_j \text{MAX_NUM_OF_UNIQ_USERS}_j}$$

where MAX_NUM_OF_UNIQ_USERS is the maximum load value among all sectors in each hour, i is the index of the current quadrant, and j are the indices of all quadrants in each hourly interval.

- To address temporal dynamics and identify the most heavily loaded points for each hour, the 95th percentile of the normalized load indicator (upper_fence_95) was calculated. This preprocessing step ensured a focus on high-density zones while filtering out noise from low-activity areas. This approach allowed us to incorporate temporal dynamics without relying on resource-intensive spatiotemporal clustering algorithms such as ST-DBSCAN, which were less suitable due to the aggregated nature of the data.
- The centroids of sectors where the normalized load value exceeded the established threshold were selected, enabling the analysis to focus on areas with the highest population activity.

Modern approaches to urban mobility analysis often use big data methods and neural networks. For example, ref. [7] proposes an LSTM-based approach for modeling urban mobility using time series data. This method demonstrates high accuracy in predicting short-term changes in mobility, making it useful for dynamic traffic management tasks. However, such approaches do not account for stable spatial activity zones, which are crucial for long-term infrastructure planning. Thus, the approaches described [8] are suitable for the analysis of short-term temporal changes, but are less effective in identifying stable spatial patterns, which are central to this study. To address these limitations, this study employed cluster analysis methods, which provide robust tools for identifying such patterns.

Three clustering algorithms were selected for the analysis of population activity: DBSCAN, KMeans++, and hierarchical agglomerative clustering [8,9]. The KMeans++ algorithm is renowned for its efficiency in processing large datasets due to its improved centroid initialization, which ensures faster convergence compared to the classical KMeans algorithm. This method is widely used in applications that require the stable and rapid partitioning of data into clusters. For instance, ref. [24] proposes a modification of KMeans that accounts for noise points, thereby enhancing the accuracy of identifying high-activity zones in cities. However, despite this enhancement, the approach is less effective when analyzing data with arbitrarily shaped clusters, as is typically the case with high-density data containing significant noise.

In contrast, the DBSCAN algorithm is highly robust to noise and outliers, enabling it to efficiently detect arbitrarily shaped clusters without requiring the number of clusters to be specified in advance. This capability is particularly important when analyzing datasets with an uneven distribution and significant noise. Moreover, this approach has applications in other domains [11], where filtering out outliers is critical for enhancing model accuracy, especially in identifying points that lie outside clusters.

Hierarchical agglomerative clustering enables data to be represented at multiple hierarchical levels while taking into account the spatial proximity of objects. Thus, this method can be used to analyze spatial data about the urban environment [23], which made it possible to identify stable patterns of data distribution and better understand their interrelations. One of the advantages of this approach is its flexibility in processing data with varying densities, as well as its ability to visualize results at different levels

of detail. However, a significant disadvantage of the method is its high computational complexity when working with large datasets, which makes it less suitable for tasks that require rapid processing.

2.4. Evaluation of Clustering Quality and Hyperparameter Optimization

In this work, the DBSCAN, KMeans++, and agglomerative clustering algorithms were employed to analyze the spatial distribution and activity of Almaty's population. Particular emphasis was placed on the DBSCAN algorithm due to its ability to effectively identify clusters with an arbitrary shape and to work with noisy data [7].

For each of the selected algorithms, optimal hyperparameters were determined using the grid search method. Cross-validation with three folds was employed to evaluate the stability and generalization ability of the models. This approach ensured an objective comparison of the algorithms and facilitated the selection of the most suitable method for the dataset. Following the justification of the clustering algorithm choices and a description of their features, a detailed analysis was conducted to evaluate the clustering quality using the selected methods: DBSCAN, agglomerative clustering, and KMeans++. The objective was to identify which algorithm most effectively detects spatial clusters in the data.

To assess the clustering quality, we used a combination of the silhouette coefficient and the Davis–Bouldin index. The silhouette coefficient evaluates cluster separability and the degree to which points belong to their assigned clusters, while the Davis–Bouldin index measures both the compactness of clusters and the separation between them. These metrics are particularly suitable for this study because they do not require pre-labeled data and can effectively assess density-based clustering methods such as DBSCAN, which is essential for analyzing urban population aggregation. For each of the selected clustering methods, three-fold cross-validation was performed to evaluate model stability and generalizability. The optimal hyperparameters for each algorithm were determined using a grid search, which systematically tests possible parameter values and selects the best combinations. This approach is particularly important in urban environments, where the high variability in data density and the presence of noise make the reliability of the results a critical factor under dynamic conditions [25,26].

The silhouette coefficient quantifies how similar an object is to its own cluster relative to other clusters. It ranges from -1 to $+1$, where higher values signify that the object is more closely aligned with its cluster and distinctly separated from other clusters.

For a single point i , the silhouette coefficient is calculated as follows:

$$s_i = (b_i - a_i) / \max(a_i, b_i)$$

where a_i is the average distance from the i -th data point to all other points in the same cluster and b_i is the minimum average distance from the i -th data point to all points in the nearest cluster (the cluster that minimizes the average distance).

Next, the average indicator S is calculated for all points and clusters:

$$S = \text{mean}(s_i)$$

The Davies–Bouldin index is calculated as the average ratio of the within-cluster spread to the between-cluster distance. Lower values indicate better cluster separation.

$$DB = 1/N \sum_{i,j=0}^n \max((S_i + S_j) / D_{ij})$$

where S_i is the average distance between elements of cluster i and the centroid of that cluster, D_{ij} is the distance between the centroids of clusters i and j , and N is the total number of clusters.

2.5. Calculation of Transport Load

To connect the identified clusters of population activity with the city's transport infrastructure, additional metrics of transport load were calculated for each cluster. For mapping public transport stops, the corresponding coordinates of the stops were obtained from OSM. Each stop was spatially associated with the corresponding quadrant. In this case, transport stops near the edges of a quadrant can also serve neighboring quadrants. After identifying clusters, the sum of unique users for all quadrants in each cluster was calculated using DBSCAN. The sum of unique users per day is the aggregated sum of NUM_OF_UNIQ_USERS across all quadrants in the cluster. The sum of unique users during rush hour was determined by determining the peak load of each quadrant (from hourly records) and summing these maxima in the corresponding clusters. To assess the transport load, the following indicators were considered:

$$\text{daily cluster load} = \frac{\text{sum of unique users per day}}{\text{number of bus stops in the cluster}},$$

$$\text{hourly cluster load} = \frac{\text{sum of unique users during rush hour}}{\text{number of bus stops in the cluster}}$$

The values obtained show how many people on average relate to each accessible stop both in everyday conditions and during rush hours. The higher these values, the higher the load on the bus stops in this cluster.

To calculate the quality of service in each cluster, the average area per transport stop (in km²) was calculated. A less dense coverage of transport stops potentially leads to an increase in the walking distance to the stops and the accumulation of city residents at each stop.

This made it possible to link the availability of transport stops with the activity of the population in stable high-density clusters.

3. Results

3.1. Heat Maps Analysis of Population Distribution

Heat maps were created based on the collected data to illustrate the load on each quadrant during specific time intervals (Figures 4–6). The color scale indicates the level of activity, ranging from blue (low load) to red (high load), allowing for a quick visual assessment of population distribution patterns.

An analysis of the heat maps (Figures 4–6) shows that night-time activity in the city is significantly lower than daytime activity, particularly during the lunch period. This observation aligns with the typical daily cycles of urban life. However, the visual analysis of heat maps provides only a general overview of the population activity distribution and does not uncover deeper spatial patterns.

The heat map-based approach has proven effective in studying spatial patterns of urban population aggregation. For instance, a study [25] utilized heat map analysis to determine the density and distribution of Almaty's population, enabling the identification of key areas of population concentration and activity using OSM data and aggregated data from a telecom operator. A similar approach in our study provides a detailed understanding of population density and activity distribution across different parts of the city, forming a foundation for more accurate decision-making in urban planning and infrastructure development.

To gain a more detailed and quantitative understanding of population activity distribution, cluster analysis methods are required. These methods enable the identification of natural groups of quadrants with similar activity and population density characteristics, the classification of zones into high, medium, and low-activity areas—crucial for targeted infrastructure and service planning—and the analysis of spatial patterns to uncover hidden trends in population aggregation.

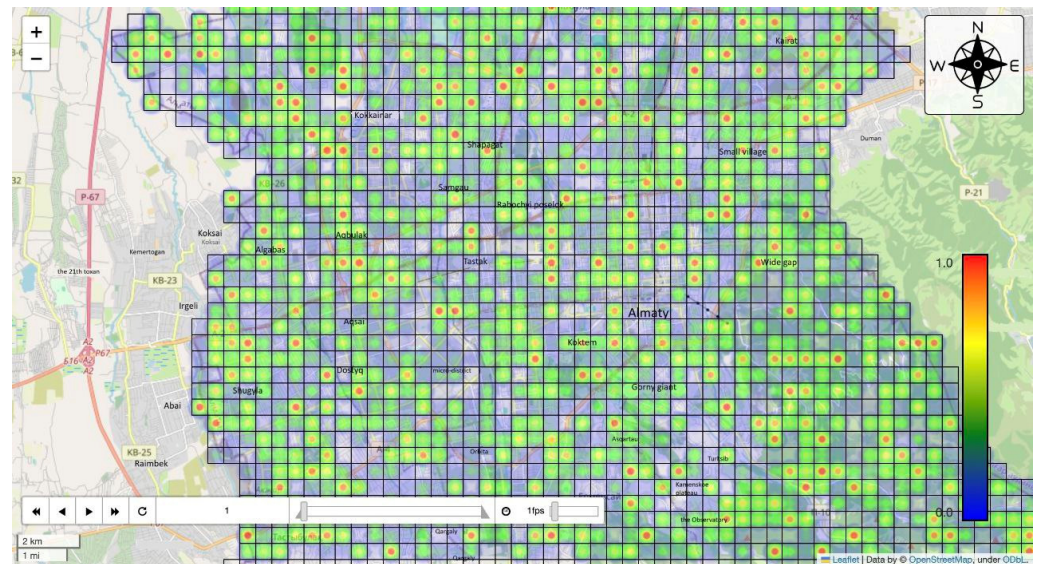


Figure 4. Heat map (fragment) of anonymous telecom data showing the number of networks connected to the link from zero (no connections) to 1 (peak flow).

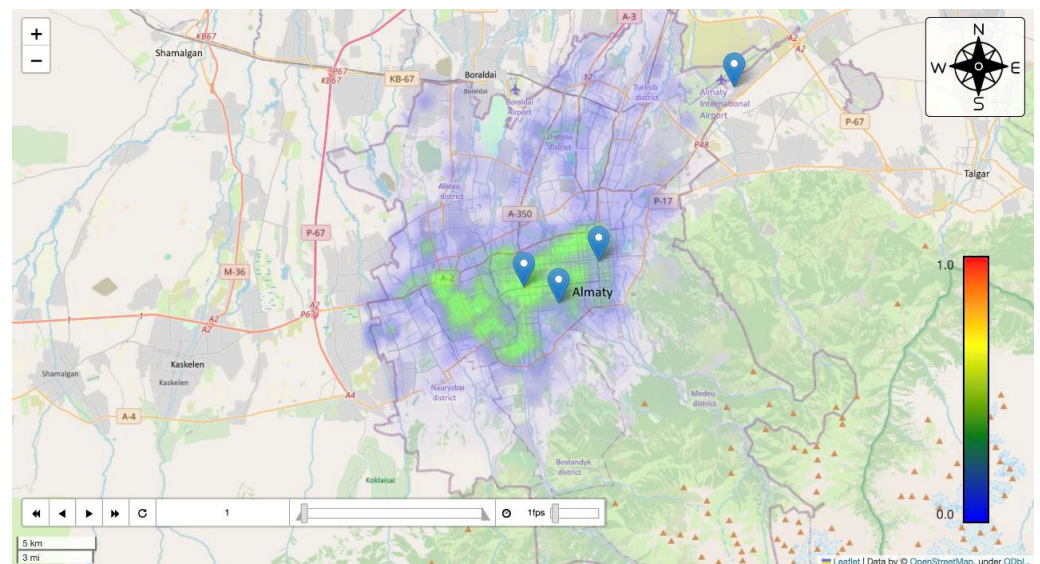


Figure 5. Heat map of pedestrian traffic intensity between 01:00 and 01:59 (0, blue—no registered users; 1, red—maximum possible user density).

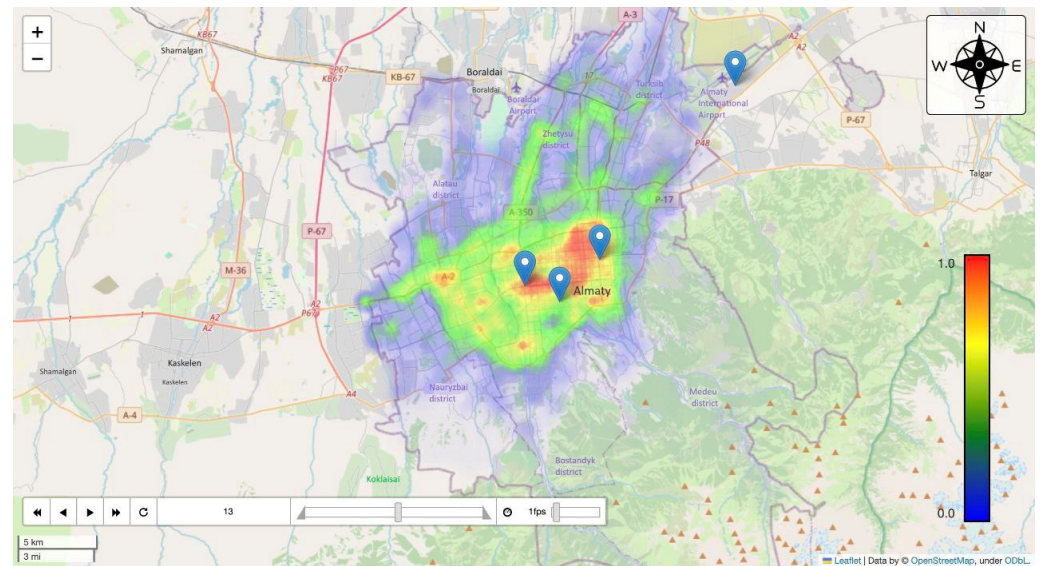


Figure 6. Heat map of pedestrian load from 13:00 to 13:59 (0, blue—no registered users; 1, red—maximum possible user density).

3.2. Comparison and Evaluation of Clustering Algorithms

The evaluation of the clustering algorithms aimed to select the most effective method for analyzing population activity in Almaty, taking into account an uneven data density, the presence of noise, and specific spatial characteristics. In our case, each data element corresponds to a 500×500 m quadrant containing three key aggregate metrics: the average number of unique users (NUM_OF_UNIQ_USERS), the number of users in “home” areas (NUM_OF_UNIQ_HOME_USERS), and the number of users in “work” areas (NUM_OF_UNIQ_WORK_USERS). These values represent the total number of cellular users and provide the basis for clustering. The algorithm comparison results (see Table 2) demonstrated that DBSCAN achieved the best performance across both metrics. For the K-means algorithm, the number of clusters was set to three based on the analysis of the silhouette coefficient, ensuring an optimal balance between cluster separability and compactness.

Table 2. Cluster comparison metrics with clustering parameters.

Algorithm	Silhouette Coefficient	The Davis-Bouldin Index	Best Parameters
DBSCAN	0.390781150	1.01724855	{‘eps’: 0.03, ‘min_samples’: 2}
Agglomerative	0.389376287	1.304225119	{‘linkage’: ‘average’, ‘n_clusters’: 3}
KMeans++	0.370775957	1.280799701	{‘init’: ‘random’, ‘n_clusters’: 3}

To select the optimal parameters `eps` and `min_samples` for the DBSCAN algorithm, a grid search method with cross-validation was employed [26]. The `eps` parameter was interpreted in a spatial context as a radius of approximately 3.34 km, which corresponds to an `eps` value of 0.03 degrees under the conditions in Almaty. This scale was chosen based on the city’s size (approximately 650 km²) and the spatial distribution of the telecom operator’s base stations. The `min_samples` parameter was varied from 1 to 3 to optimize the clustering results. We used the Euclidean metric, which is justified for relatively small areas like a single city; at this scale, differences from a geodesic metric are

minimal. The `min_samples` parameter was varied from 1 to 3, and the final values of `eps` = 0.03 and `min_samples` = 2 achieved the best results according to the silhouette and Davis–Bouldin indices.

For hierarchical agglomerative clustering, the “average” linkage method was applied, which ensured stable results at various levels of cluster detail.

While ST-DBSCAN extends DBSCAN’s capabilities by incorporating temporal correlations, its application was not feasible in this study due to data limitations. The aggregated and anonymized nature of the dataset, which provides an hourly population load per quadrant without individual movement tracking, ensured confidentiality but limited the use of methods that require the detailed spatiotemporal tracking of each telecommunications network user’s movement. Additionally, our primary objective was to identify stable high-density zones to support urban infrastructure planning, making DBSCAN an optimal choice. DBSCAN’s ability to handle noise and detect arbitrarily shaped clusters aligns well with the irregular spatial patterns observed in Almaty, providing robust and reproducible results.

Hyperparameter optimization was performed using grid search, and model stability was evaluated using three-fold cross-validation. The best performance—yielding a silhouette coefficient of 0.39 and a Davies–Bouldin index of 1.017—was achieved with the parameters `eps` = 0.03 and `min_samples` = 2.

The evaluation results presented in Table 2 clearly highlight the advantage of DBSCAN in detecting stable high-density zones amidst noisy and complex urban data. DBSCAN proved to be the most effective method for identifying clusters in datasets with uneven point distribution and noisy data. Unlike K-means, which requires the number of clusters to be predefined, DBSCAN automatically determines the number of clusters based on point density, making it particularly suitable for analyzing complex, noisy datasets. Additionally, DBSCAN excels at detecting arbitrarily shaped clusters and identifying high-density areas without being constrained by assumptions about the cluster structure. However, it is sensitive to the selection of parameters such as `eps` (the maximum distance between points to be considered neighbors) and `min_samples` (the minimum number of points required to form a dense region), which necessitates careful data preprocessing and parameter optimization. While agglomerative clustering offers flexibility for analyzing data at different levels of granularity, its high computational complexity limits its applicability for large datasets. Given the specific characteristics of our data, including irregular spatial patterns and the presence of noise, DBSCAN was selected as the most suitable clustering method for this study.

To achieve a good clustering quality, a heatmap analysis of the DBSCAN algorithm parameters was performed on the dataset (Figure 7). The results revealed that `eps` values around 0.03 and above improve the silhouette coefficient, indicating better cluster separation. Conversely, for the Davies–Bouldin index, `eps` values below 0.03 are preferable, reflecting greater cluster compactness. Additionally, optimal results are achieved by selecting the `min_samples` parameter in the range of 1 to 3 for low `eps` values and 4 to 9 for high `eps` values.

The heat maps indicate that `eps` values around 0.03 and higher lead to an increase in the silhouette coefficient, whereas lower `eps` values improve the Davies–Bouldin index. Considering the specific characteristics of our data, such as their uneven point distribution and the presence of noise, the DBSCAN algorithm was selected as the most suitable clustering method. Its ability to accurately identify high-population concentration areas with complex and irregular shapes makes it particularly effective for this analysis.

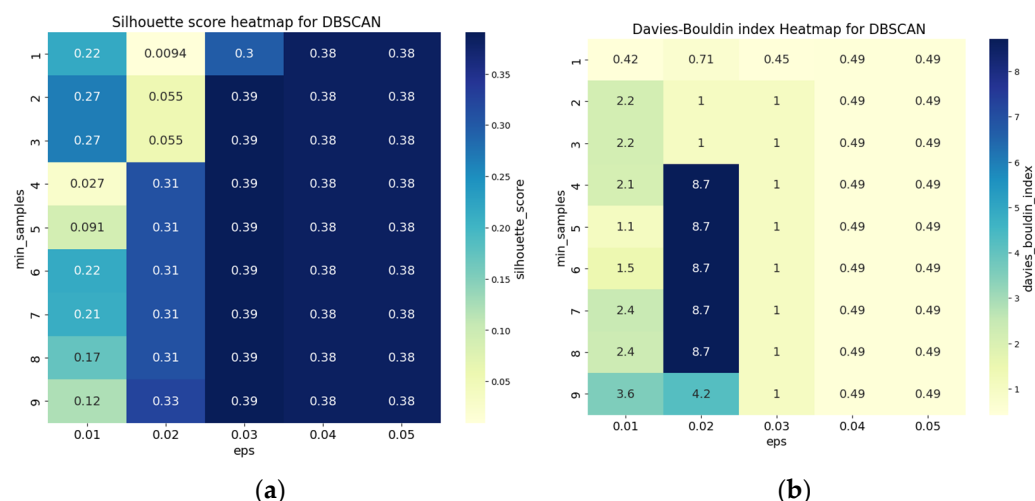


Figure 7. Heat maps of DBSCAN algorithm parameter estimation: (a) silhouette coefficient; (b) Davies–Bouldin index.

3.3. Visualization of Clustering Results and Analysis of Population Activity Dynamics

The analysis demonstrated that DBSCAN provides the highest cluster stability under changes in data and model parameters, as evidenced by high silhouette coefficient values and low Davies–Bouldin index values. This indicates that the algorithm is particularly well suited for processing complex urban datasets characterized by uneven distributions and the presence of noise points.

Based on OSM data and mobile phone density data, a clustering map that identifies twelve areas within the city exhibiting consistently high population density was constructed, regardless of the time of day (see Figure 8). This demonstrates the effectiveness of the DBSCAN algorithm in clustering urban population density data, which are characterized by high spatial heterogeneity, variable point density, and the absence of clearly defined boundaries between activity zones.

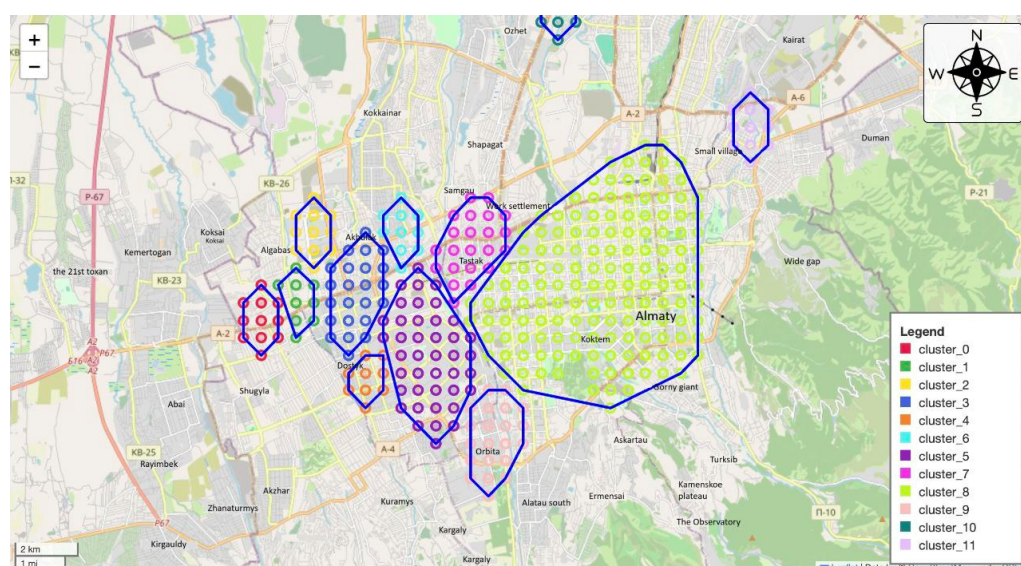


Figure 8. Identification of zones (blue lines) with a consistently high population density across the city, regardless of the time of day.

A comparison of the clustering maps constructed for different time intervals allowed us to analyze the dynamics of population activity within the static clusters throughout the day. To achieve this, the cluster boundaries were overlaid on an activity heat map

(see Figure 9). The analysis revealed that between 14:00 and 14:59, the highest load falls on clusters located in commercial and business districts, likely due to the lunch break. In the evening, between 18:00 and 18:59, the load is redistributed, reflecting the end of the workday and the movement of the population toward residential and recreational areas. The alignment of the cluster boundaries with the zones of highest population density confirms both the accuracy of their delineation and the adequacy of the chosen clustering method.

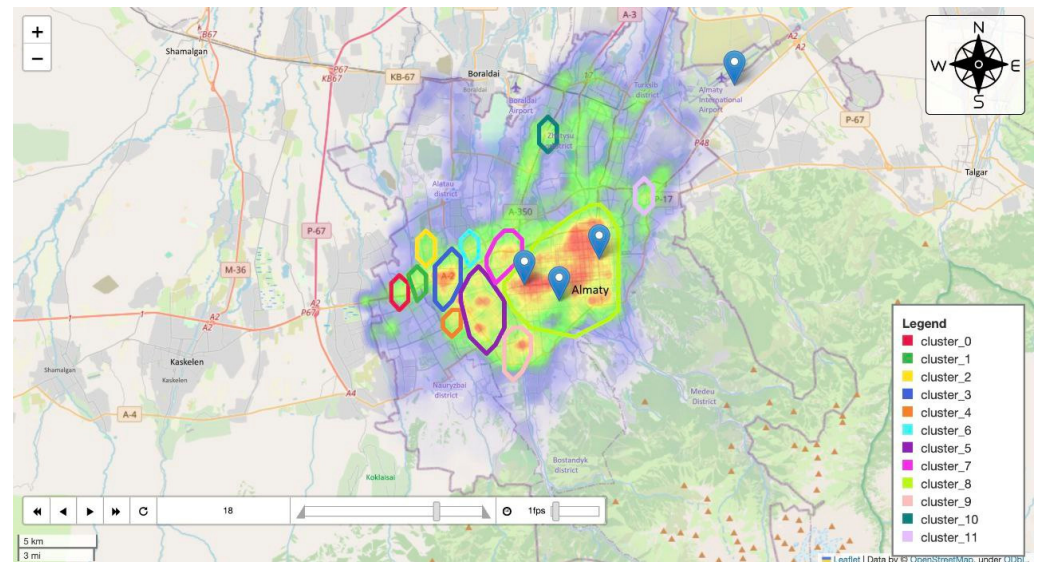


Figure 9. Visualization of identified cluster boundaries overlaid on a heatmap (18:00–18:59, peak hour).

One of the key aspects of applying cluster analysis in urban studies is the ability of algorithms to process large volumes of heterogeneous data and reveal hidden spatiotemporal patterns. Unlike traditional methods such as K-Means, which require a predetermined number of clusters, DBSCAN adaptively identifies dense groups while ignoring noise points—a feature that is especially important in highly heterogeneous urban environments [27]. The proposed methodology not only identifies key activity zones but also accounts for their temporal fluctuations, thereby providing opportunities for monitoring and forecasting changes in the load on urban infrastructure. The use of unsupervised clustering methods confirms their effectiveness in spatiotemporal analysis by revealing patterns that are inaccessible with traditional approaches. Machine learning helps increase the accuracy of analysis in conditions of high data variability [28], making this method especially relevant for optimizing urban planning and traffic flow management.

3.4. Analysis of the Dynamics of Unique Users over Time in Clusters

As part of the study, a detailed analysis of the temporal dynamics of unique user activity within various clusters identified by the DBSCAN algorithm was performed. For each cluster, graphs were created to depict changes in the total number of unique users throughout the day (Figures 9–11). The X-axis on the graphs represents the hours of the day (from 0 to 23), while the Y-axis indicates the total number of unique users at each specified hour. The graph colors correspond to the colors of the clusters shown in Figures 7 and 8.

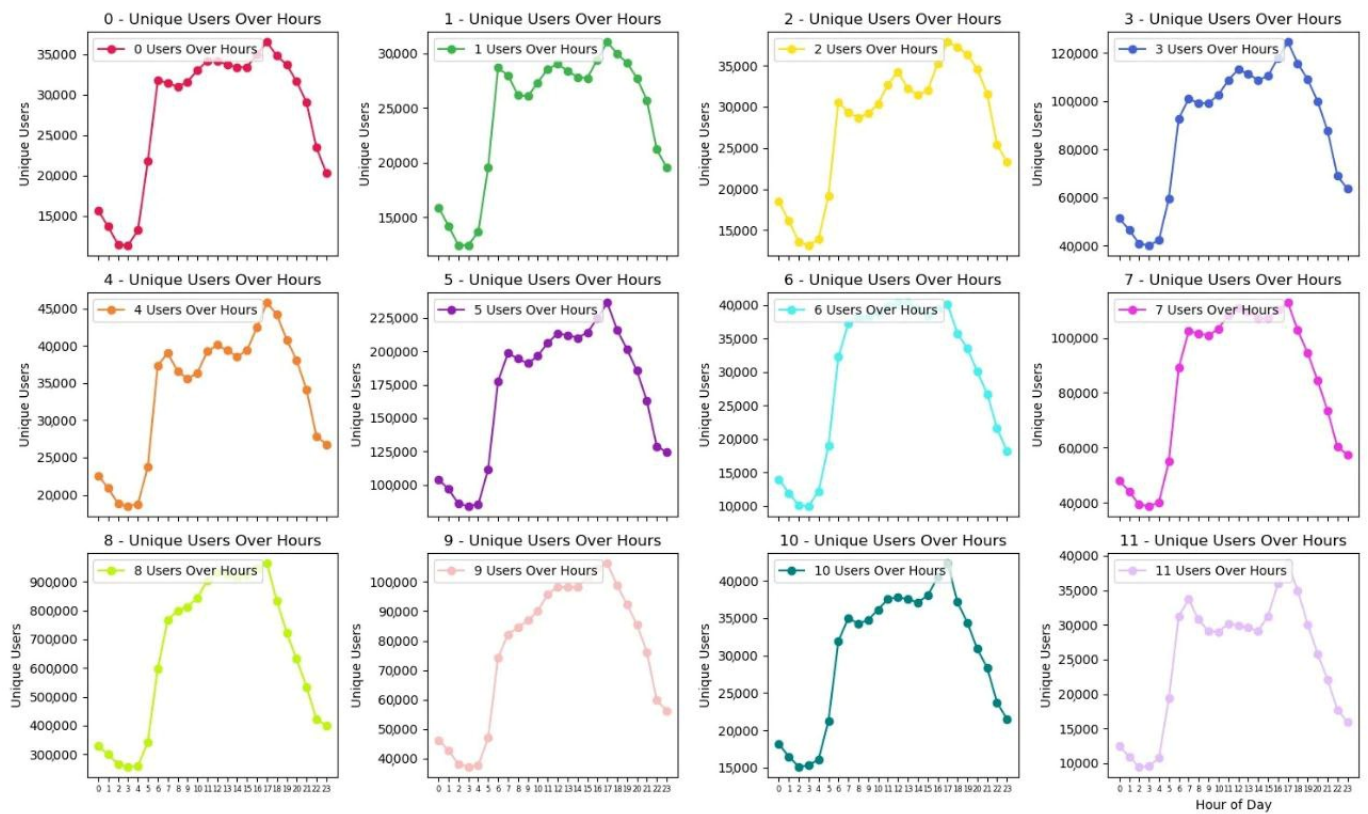


Figure 10. Dynamics of the unique number of people by hours in each cluster.

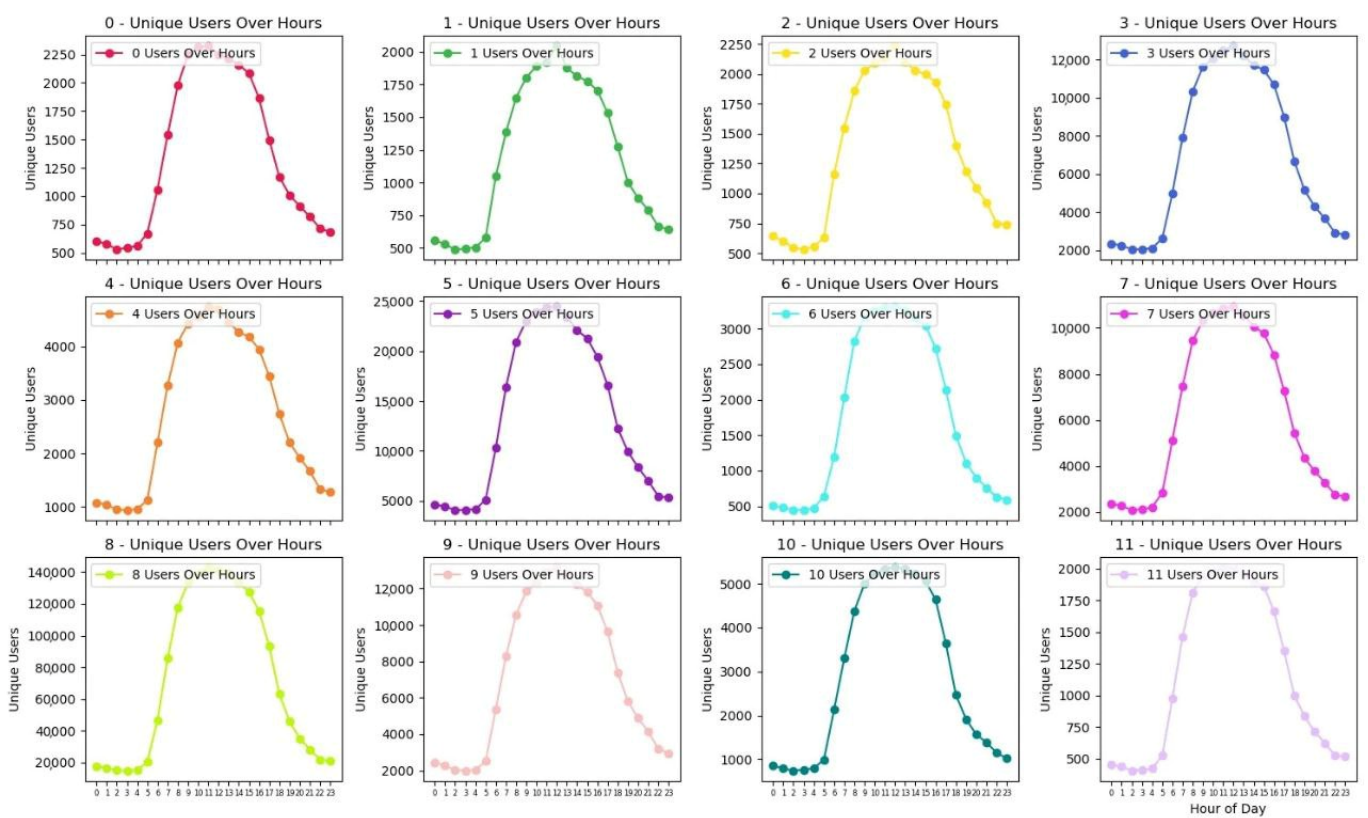


Figure 11. Dynamics of the unique number of people at the work location by hours in the context of each cluster.

The temporal dynamics analysis uncovered distinct patterns of user activity across the clusters, reflecting both common daily rhythms and unique variations tied to specific urban zones. The graphs (Figures 10–12) demonstrate an overall pattern consistent with typical urban life cycles, where morning activity peaks between 8:00 and 10:00, followed by either a plateau or a decline during the daytime, and a second peak in the evening between 17:00 and 19:00. This alignment with expected behavioral patterns serves as a validation of data quality, confirming its reliability and representativeness of real urban dynamics.

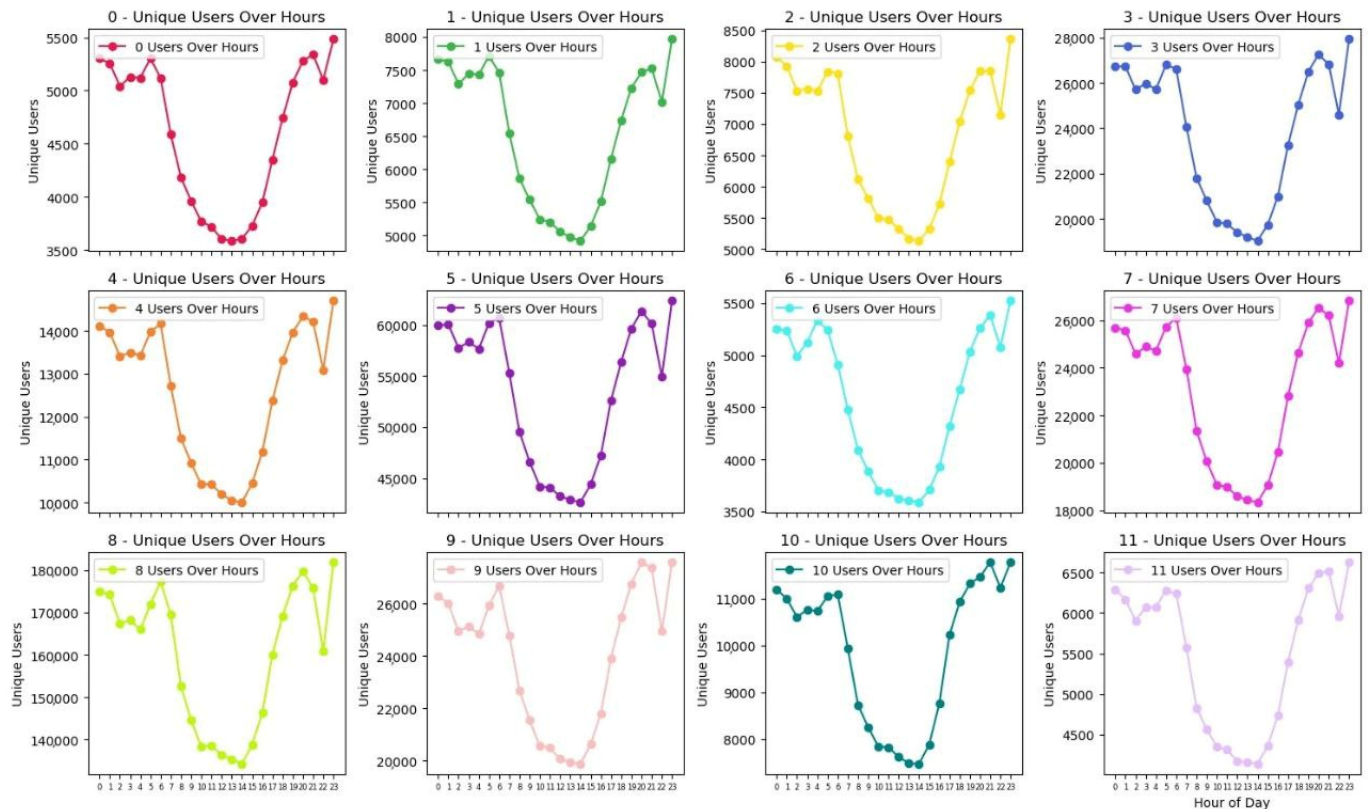


Figure 12. Dynamics of the unique number of people at home location by hours in each cluster.

While the general rhythm is shared across clusters, the magnitude and distribution of activity reveal significant differences. Clusters such as 3 and 5 maintain consistently high user counts throughout the day, indicative of central business districts or transport hubs where commercial and transit activities dominate. In contrast, clusters like 7 exhibit sharp morning and evening peaks but lower activity levels during the daytime, pointing to predominantly residential areas. Smaller clusters, such as 8, display overall lower user numbers, which may correspond to localized zones like small neighborhoods or regions with limited infrastructure. These differences not only highlight the heterogeneity of urban zones but also illustrate how population activity dynamically redistributes throughout the day—a finding further supported by the works [27,29].

The analysis also reveals patterns of movement between clusters, shedding light on urban population flows. Residential clusters experience a decrease in activity during the daytime as people move towards commercial or transit hubs, reflected in increased activity in these zones. Conversely, evening hours show a reverse flow, with residential clusters regaining activity as people return home. These inter-cluster movements provide a critical understanding of urban mobility dynamics, offering a foundation for designing efficient transportation systems.

The variations in user density across clusters, as shown on the Y-axis of the graphs, also offer insights into public transportation planning. High-density clusters with sustained

activity throughout the day, such as 3 and 5, require increased transportation capacity to accommodate population flows effectively. Conversely, smaller clusters with lower activity levels can operate with fewer resources, enabling the targeted optimization of transport services. This approach not only enhances efficiency but also ensures equitable resource allocation based on cluster-specific needs.

Temporal analysis, while robust, also opens new opportunities for future research. One promising direction is to assess the causal impact of various infrastructure elements, such as shopping malls, social facilities, or transit hubs, on the dynamics of specific clusters. Additionally, exploring how targeted interventions, such as introducing new transport routes or modifying land use policies, affect activity levels could provide actionable insights for urban planning. By extending the findings of this study, these future investigations could deepen our understanding of urban dynamics and further contribute to the advancement of urban planning strategies.

Particular attention is drawn to clusters with the highest activity, such as clusters 3, 5, and 7. These clusters exhibit significantly higher user counts compared to others, suggesting that they encompass central or highly frequented areas of the city, including business districts, major transportation hubs, or popular public spaces.

When analyzing users based on work and home locations, distinct patterns emerge that align with a typical work schedule. The number of users with a work location peaks during work hours and declines in the evening and at night. Conversely, users with a home location display the opposite trend: activity increases in the evening and at night and decreases during work hours. These findings confirm the expected behavioral patterns of the urban population and underscore the reliability of the collected data.

3.5. Analysis of Transport Load in Clusters

In addition to the analysis of the temporal dynamics of population activity, an assessment of the load on the transport infrastructure in each of the identified clusters was carried out in accordance with the procedure described in Section 2.5, and transport load indicators were calculated for each cluster (Table 3).

Table 3. Public transport load and stop coverage for each cluster.

Cluster ID	Bus Stations	Sum of Unique Users daily	Cluster Load Daily	Sum of Unique Users Peak Hour	Cluster Load Peak Hour	Area, km ²	Station Area Recall, km ²
0	9	659,825.913	73,314	34849	3872	3.05	0.3389
1	8	579,801.9565	72,475	29,976.47826	3747	2.54	0.3175
2	16	666,888.3043	41,681	37,301.82609	2331	3.05	0.1906
3	46	2,115,066.304	45,980	115,664.4348	2514	7.89	0.1715
4	17	804,989.2609	47,352	44,188.52174	2599	2.29	0.1347
5	96	4,062,139.043	42,314	216,134.7826	2251	17.07	0.1778
6	8	705,626.0435	88,203	35,739	4467	2.8	0.35
7	32	2,001,579.696	62,549	102,719.3913	3210	7.89	0.2466
8	470	15,630,093.48	33,256	834,368.3478	1775	70.81	0.1507
9	43	1,838,653.826	42,759	98,810	2298	7.15	0.1663
10	13	721,371.4348	55,490	37,170.08696	2859	3.04	0.2338
11	7	598,000.1739	85,429	34,962.08696	4995	3.05	0.4357

The analysis of transport load showed that the maximum concentration of passengers during peak hours was recorded in clusters 11, 6, 0, 1, 7, and 10, where the number of users exceeds 2800 people per hour. This indicates a significant overload of transport hubs in these areas. Meanwhile, clusters 8, 5, and 3, despite the high daily load, demonstrate a more uniform distribution of passenger traffic throughout the day, possibly due to the presence of a well-developed public transport network and a sufficient number of transport stops.

A further analysis of the stop coverage ratio revealed that in clusters 11, 6, 0, 1, 7, and 10, the average area per stop is 0.3–0.4 km², which exceeds the overall average. This suggests that the stop density in these congested areas is insufficient, potentially reducing public transport availability and increasing the distance between passenger entry and exit points. These findings underscore the need to improve transport coverage in high-load areas to evenly distribute passenger flow and reduce congestion on individual routes.

4. Discussion

The results of this study underscore the importance of clustering algorithms in urban planning, particularly for population mobility analysis and infrastructure optimization. Previous research has extensively applied cluster analysis methods and mobility data to examine the spatial structure of cities. For instance, K-Means and DBSCAN algorithms have been successfully employed to identify functional zones in Shanghai, demonstrating their effectiveness in addressing spatial planning challenges [12,27]. Similarly, mobile operator data have been used to analyze population movement, providing valuable insights into urban dynamics [5]. At the same time, machine learning methods have proven effective in identifying hidden patterns and improving predictive analytics in complex datasets [28]. This aligns with previous research demonstrating the applicability of machine learning algorithms in analyzing unknown and complex datasets [28]. However, most of these studies have focused on short-term mobility patterns or relied solely on dynamic data.

In contrast to traditional approaches, our study integrates anonymized telecommunication data with modern clustering methods—particularly DBSCAN—enabling us to identify stable areas of high population density that are critical for long-term urban planning. This method demonstrates a high degree of robustness to irregular spatial patterns and noise, aspects that are often overlooked in conventional approaches. Unlike studies that focus on short-term mobility changes (e.g., clustering weekly movements based on mobile operator data [5,6,27]), our work concentrates on identifying stable activity zones. This approach contributes to the development of sustainable urban infrastructure and facilitates more efficient long-term resource allocation. Further improvements in DBSCAN's performance, as proposed by [29], allow for the analysis of high-dimensional spatial data with optimized computational costs, making it even more effective for large-scale urban analytics.

However, DBSCAN also has some limitations. The algorithm requires two key parameters (eps and min_samples), the choice of which may depend on the initial data distribution. This may affect the resulting cluster structures, especially in datasets with widely varying densities. Also, DBSCAN's dependence on the distance metric may introduce minor inaccuracies. The use of 500 × 500 m quadrant-based data aggregation may obscure some subtle patterns. Despite this, DBSCAN's robustness to noise, ability to detect arbitrarily shaped clusters, and lack of a required predetermined number of clusters make it well suited to analyzing complex urban environments. The proposed analysis methodology is flexible and scalable, allowing it to be adapted to various urban conditions. The use of readily available telecommunication and geoinformation data makes our study applicable in situations where individual movement data are unavailable [4].

Recent research has explored different clustering approaches to classify urban areas more effectively, emphasizing the advantages of machine learning-based clustering meth-

ods in optimizing urban spatial structures [30]. These studies demonstrate that selecting an appropriate clustering approach significantly influences the accuracy of urban classification and infrastructure planning. Furthermore, studies have demonstrated that combining spatial indicators, social media activity, and geo-statistical methods allows for a more comprehensive assessment of urban dynamics and infrastructure needs [31]. In the future, incorporating these data sources could further enhance the accuracy and granularity of urban planning strategies.

One of the key applications of this approach is assessing the imbalance in public transport provision across different parts of the city. Analysis revealed that the most overloaded clusters (11, 6, 0, 1, 7, and 10) experience high loads during rush hours, leading to a concentration of passenger traffic at specific transport hubs and resulting in overload on key routes. This situation is exacerbated by the low density of stopping points, which forces passengers to converge at a limited number of locations, thereby causing overload on individual routes. In contrast, clusters 8, 5, and 3 exhibit a more uniform distribution of passenger traffic throughout the day, likely due to a well-developed transport infrastructure and a high density of stops. These results confirm that increasing the number of transport hubs contributes to a more uniform distribution of passenger load and enhances the accessibility of transport services.

Taking into account the temporal dynamics of population activity enables the optimization of public transport schedules and more efficient resource allocation. Identifying zones with a consistently high passenger flow density during specific time intervals creates opportunities for the dynamic redistribution of routes, thereby reducing congestion and enhancing mobility. In addition, these data can be used to develop environmentally sustainable infrastructure—for example, by creating bicycle routes that connect residential, business, and recreational areas—which will contribute to the sustainable development of urban mobility and improve residents' quality of life.

To address the identified imbalances in transport infrastructure provision, we propose the following measures: (1) Increase the number of public transport stops in the most congested clusters (11, 6, 0, 1, 7, and 10) to reduce the concentration of passengers on a limited number of routes and distribute the load more evenly; (2) Optimize public transport routes to shorten the distance between stops and enhance accessibility in high-density areas; and (3) Implement adaptive traffic flow management, including increasing the number of trips during rush hours and dynamically adjusting routes based on actual load data. The implementation of these strategies will significantly improve public transport efficiency, reduce congestion at individual hubs, and enhance passenger comfort. These findings align with previous studies [32–35], which emphasize the impact of both polycentric and monocentric city structures on traffic flow formation and the optimization of urban transport networks.

Although this paper focuses on the analysis of transport infrastructure, the proposed methodology has broader applications. For example, the identified high-density population zones can be used to optimize emergency response systems by ensuring that rapid response services (police, fire department, ambulance) are located in close proximity to areas with a high population concentration; they can also inform the strategic placement of commercial facilities—targeting locations with stable pedestrian traffic for new shopping centers or business districts—and support the development of environmentally sustainable urban solutions, such as parks, pedestrian zones, and recreational areas in regions with high population activity. Thus, the present study not only confirms the effectiveness of clustering methods in urban mobility analysis but also demonstrates their potential across various areas of urban planning.

This study demonstrates the potential use of modern big data and machine learning methods for analyzing the spatiotemporal dynamics of urban activity. The integration of the DBSCAN algorithm with temporal analysis provided valuable insights into the processes occurring in the urban environment. These results contribute to the advancement of research in geoinformatics and urban studies by offering a scalable methodology that can be applied to other cities and regions. Future research should further integrate machine learning algorithms—such as optimization methods (e.g., ant colony algorithms)—to enhance the planning of transport flows and routes. Another promising direction is the real-time monitoring of traffic flows between clusters, which would improve the accuracy of forecasting transport bottlenecks and congestion zones. Additionally, incorporating social network analysis, IoT device data, and other open sources of information will enable an even more detailed study of population mobility and further improve the accuracy of the results.

5. Conclusions

In this study, we analyzed the spatial aggregation and activity of Almaty's population using cluster analysis methods. The identified high-activity clusters enabled us to determine key areas that require strategic urban infrastructure planning and traffic management. Moreover, the application of the DBSCAN algorithm demonstrated its effectiveness in identifying stable clusters under conditions of high heterogeneity and noisy data, which is particularly crucial for analyzing the spatiotemporal patterns of urban activity.

The results obtained have practical significance, particularly in the context of public transport route optimization, urban planning, and sustainable urban development. The analysis revealed an imbalance in the availability of Almaty's transport infrastructure, underscoring the need to expand the public transport network and optimize routes in congested clusters to achieve a more even distribution of passenger flows. These findings provide a foundation for more effective mobility planning, including adaptive route management and dynamic traffic load control.

Beyond transport optimization, this study demonstrates the potential of clustering methods to address a broader range of urban planning issues. The integration of spatial data and machine learning algorithms enables a more accurate analysis of spatiotemporal population activity, which can be instrumental in developing emergency response strategies, optimizing the placement of commercial facilities, and designing public spaces.

By focusing on population dynamics and the stability of the identified activity zones, this study proposes a systematic approach to analyzing urban infrastructure aimed at improving the availability and quality of urban services. Future research prospects include expanding the dataset by integrating IoT data and implementing real-time analysis to enhance forecasting and decision-making in urban management.

Author Contributions: Conceptualization, G.B.; methodology, A.B.; software, S.O., G.S. and D.U.; validation, A.M. and S.N.; formal analysis, A.B.; investigation, G.B.; resources, G.B.; data curation, S.O.; writing—original draft preparation, A.B.; writing—review and editing, A.B.; visualization, G.S. and D.U.; supervision, G.B.; project administration, A.M. and S.N.; funding acquisition, G.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research has is funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. AP19674517- Development of a smart map for planning and evaluating the efficiency of urban infrastructure based on human activity analyzing models).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The geographic data used in this study were obtained from OpenStreetMap and are publicly available at <https://www.openstreetmap.org>, accessed on 7 October 2023. Aggregated and anonymized data on population activities were provided by the telecom operator for the period from 1 March to 1 June 2023 under a data sharing agreement. This data is not publicly available due to privacy and ethical restrictions.

Conflicts of Interest: Aiman Moldagulova was employed by the “Non-Profit Joint-Stock Company “K.I. Satbayev Kazakh National Research Technical University”. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

- Du, Z.; Qian, Y.; Li, C.; Hu, Y.; Qiu, J. An Innovative High-Dimensional Clustering Algorithm and its Application to Urban Stratification and Planning. In Proceedings of the 2024 7th International Conference on Advanced Algorithms and Control Engineering (ICAACE), Shanghai, China, 1–3 March 2024; pp. 107–111. [\[CrossRef\]](#)
- Cheng, J.; Li, K.; Liang, Y.; Sun, L.; Yan, J.; Wu, Y. Rethinking Urban Mobility Prediction: A Multivariate Time Series Forecasting Approach. *IEEE Trans. Intell. Transp. Syst.* **2024**, *26*, 2543–2557. [\[CrossRef\]](#)
- Shi, H.; Huang, H.; Ma, D.; Chen, L.; Zhao, M. Capturing urban recreational hotspots from GPS data: A new framework in the lens of spatial heterogeneity. *Comput. Environ. Urban Syst.* **2023**, *103*, 101972. [\[CrossRef\]](#)
- Liu, X.; Payakkamas, P.; Dijk, M.; de Kraker, J. GIS Models for Sustainable Urban Mobility Planning: Current Use, Future Needs and Potentials. *Future Transp.* **2023**, *3*, 384–402. [\[CrossRef\]](#)
- Xu, Y.; Shaw, S.L.; Zhao, Z.; Yin, L.; Fang, Z.; Li, Q. Understanding aggregate human mobility patterns using passive mobile phone location data: A home-based approach. *Transportation* **2015**, *42*, 625–646. [\[CrossRef\]](#)
- Mekeawd, S.; Khamitkar, S.; Bhalchandra, P.; Lokhande, S. Discovery of Usage Pattern from Mobile Call Data Using Clustering Approaches. In Proceedings of the 2nd International Conference on Cognitive and Intelligent Computing, ICCIC 2022, Cognitive Science and Technology, Hyderabad, India, 27–28 December 2022; Kumar, A., Ghinea, G., Merugu, S., Eds.; Springer: Singapore, 2023. [\[CrossRef\]](#)
- Liu, Y.; Dong, B. Modeling urban scale human mobility through big data analysis and machine learning. *Build. Simul.* **2024**, *17*, 3–21. [\[CrossRef\]](#)
- Wang, R.; Zheng, W.; Huang, M.; Li, G. Driving Behavior Evaluation Based on DBSCAN and Kmeans++ Clustering. In Proceedings of the 2022 5th International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE), Wuhan, China, 22–24 April 2022; pp. 188–193. [\[CrossRef\]](#)
- Cesario, E.; Vinci, A.; Zhu, X. Hierarchical Clustering of Spatial Urban Data. In *Numerical Computations: Theory and Algorithms*; Sergeyev, Y., Kvasov, D., Eds.; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2020; Volume 11973. [\[CrossRef\]](#)
- Ran, X.; Zhou, X.; Lei, M.; Tepsan, W.; Deng, W. A Novel K-Means Clustering Algorithm with a Noise Algorithm for Capturing Urban Hotspots. *Appl. Sci.* **2021**, *11*, 11202. [\[CrossRef\]](#)
- Daurenbayeva, N.; Nurlanuly, A.; Atymtayeva, L.; Mendes, M. Survey of Applications of Machine Learning for Fault Detection, Diagnosis and Prediction in Microclimate Control Systems. *Energies* **2023**, *16*, 3508. [\[CrossRef\]](#)
- Yuan, Y.; Raubal, M. Extracting Dynamic Urban Mobility Patterns from Mobile Phone Data. In *Geographic Information Science. GIScience 2012*; Xiao, N., Kwan, M.P., Goodchild, M.F., Shekhar, S., Eds.; Springer: Berlin, Heidelberg, Germany, 2012; Volume 7478, pp. 354–367. [\[CrossRef\]](#)
- Senaratne, H.; Mueller, M.; Behrisch, M.; Lalanne, F.; Bustos-Jiménez, J.; Schneidewind, J.; Keim, D.; Schreck, T. Urban Mobility Analysis With Mobile Network Data: A Visual Analytics Approach. *IEEE Trans. Intell. Transp. Syst.* **2018**, *19*, 1537–1546. [\[CrossRef\]](#)
- Douglass, R.W.; Meyer, D.A.; Ram, M.; Rideout, D.; Song, D. High resolution population estimates from telecommunications data. *EPJ Data Sci.* **2015**, *4*, 4. [\[CrossRef\]](#)
- Sevtsuk, A. Estimating Pedestrian Flows on Street Networks: Revisiting the Betweenness Index. *J. Am. Plan. Assoc.* **2021**, *87*, 512–526. [\[CrossRef\]](#)
- Wang, Y.; Currim, F.; Ram, S. Deep Learning of Spatiotemporal Patterns for Urban Mobility Prediction Using Big Data. *Inf. Syst. Res.* **2022**, *33*, 579–598. [\[CrossRef\]](#)
- Han, C.; Seshadri, P.; Ding, Y.; Posner, N.; Koo, B.W.; Agrawal, A.; Lerch, A.; Guhathakurta, S. Understanding Pedestrian Movement Using Urban Sensing Technologies: The Promise of Audio-Based Sensors. *Urban Inform.* **2024**, *3*, 22. [\[CrossRef\]](#)
- Molyneaux, N.; Scarinci, R.; Bierlaire, M. Design and Analysis of Control Strategies for Pedestrian Flows. *Transportation* **2021**, *48*, 1767–1807. [\[CrossRef\]](#)
- D’Apuzzo, M.; Santilli, D.; Evangelisti, A.; Pelagalli, V.; Montanaro, O.; Nicolosi, V. An Exploratory Step to Evaluate the Pedestrian Flow in Urban Environment. In *Computational Science and Its Applications—ICCSA 2020*; Gervasi, O., Murgante, B., Misra, S.,

- Garau, C.; Blečić, I.; Taniar, D.; Apduhan, B.O.; Rocha, A.M.A.; Tarantino, E.; Torre, C.M., et al., Eds.; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2020; Volume 12255. [CrossRef]
20. United Nations Economic Commission for Europe. Smart Sustainable Cities Profile: Almaty, Kazakhstan. UNECE 2024. Available online: <https://unece.org/housing-and-land-management/publications/smart-sustainable-cities-profile-almaty-kazakhstan> (accessed on 20 August 2024).
 21. Kazakhstan: Regions, Major Cities & Settlements—Population Statistics, Maps, Charts, Weather and Web Information. Available online: <https://citypopulation.de/en/kazakhstan/cities/> (accessed on 15 September 2014).
 22. Toshniwal, D.; Chaturvedi, N.; Parida, M.; Garg, A.; Choudhary, C.; Choudhary, Y. Application of Clustering Algorithms for Spatio-Temporal Analysis of Urban Traffic Data. *Transp. Res. Procedia* **2020**, *48*, 1046–1059. [CrossRef]
 23. López Baeza, J.; Carpio-Pinedo, J.; Sievert, J.; Landwehr, A.; Preuner, P.; Borgmann, K.; Avakumović, M.; Weissbach, A.; Bruns-Berentelg, J.; Noennig, J.R. Modeling Pedestrian Flows: Agent-Based Simulations of Pedestrian Activity for Land Use Distributions in Urban Developments. *Sustainability* **2021**, *13*, 9268. [CrossRef]
 24. Schirmer, P.M.; Axhausen, K.W. A Multiscale Clustering of the Urban Morphology for Use in Quantitative Models. In *The Mathematics of Urban Morphology; Modeling and Simulation in Science, Engineering and Technology*; D’Acci, L., Ed.; Birkhäuser: Cham, Switzerland, 2019. [CrossRef]
 25. Bektemyssova, G.; Moldagulova, A.; Shaikemelov, G.; Omarov, S.; Nuralykyzy, S. Research on Spatial Aggregation Patterns of Urban Population in Almaty City Based on Heat Map. In Proceedings of the 10th International Conference on Control, Decision and Information Technologies, CoDIT 2024, Valletta, Malta, 1–4 July 2024; pp. 2194–2198.
 26. Li, Z.; Zhao, G. Revealing the Spatio-Temporal Heterogeneity of the Association between the Built Environment and Urban Vitality in Shenzhen. *ISPRS Int. J. Geo-Inf.* **2023**, *12*, 433. [CrossRef]
 27. Luo, T.; Ebrahimpour, Z.; Wan, W.; Cervantes, O. A Study on Functional Planning in Shanghai Regions Using K-Means and DBSCAN Clustering. In Proceedings of the 2018 International Conference on Audio, Language and Image Processing (ICALIP), Shanghai, China, 16–17 July 2018; pp. 251–256. [CrossRef]
 28. Daurenbayeva, N.; Atymtayeva, L.; Nurlanuly, A.; Bykov, A.; Turusbekova, U.; Shuitenov, G. A Machine Learning Approach to Microclimate Monitoring and Fault Detection. *Appl. Math. Inf. Sci.* **2025**, *19*, 327–334. [CrossRef]
 29. Chen, Y.; Tang, S.; Bouguila, N.; Wang, C.; Du, J.; Li, H. A Fast Clustering Algorithm Based on Pruning Unnecessary Distance Computations in DBSCAN for High-Dimensional Data. *Pattern Recognit.* **2018**, *83*, 375–388. [CrossRef]
 30. Vera, C.; Lucchini, F.; Bro, N.; Mendoza, M.; Löbel, H.; Gutiérrez, F.; Dimter, J.; Cuchacovic, G.; Reyes, A.; Valdivieso, H.; et al. Learning to Cluster Urban Areas: Two Competitive Approaches and an Empirical Validation. *EPJ Data Sci.* **2022**, *11*, 62. [CrossRef]
 31. Bernetti, I.; Alampi Sottini, V.; Bambi, L.; Barbierato, E.; Borghini, T.; Capecchi, I.; Saragosa, C. Urban Niche Assessment: An Approach Integrating Social Media Analysis, Spatial Urban Indicators and Geo-Statistical Techniques. *Sustainability* **2020**, *12*, 3982. [CrossRef]
 32. Volpati, V.; Barthelémy, M. The Spatial Organization of the Population Density in Cities. *arXiv* **2018**, arXiv:1804.00855.
 33. Abozeid, A.S.M.; AboElatta, T.A. Polycentric vs Monocentric Urban Structure Contribution to National Development. *J. Eng. Appl. Sci.* **2021**, *68*, 11. [CrossRef]
 34. Alizade, M.; Kheni, R.; Price, S.; Sousa, B.C.; Cote, D.L.; Neamtu, R. A Comparative Study of Clustering Methods for Nanoindentation Mapping Data. *Integr. Mater. Manuf. Innov.* **2024**, *13*, 526–540. [CrossRef]
 35. Akhmer, Y.; Akhmer, Y.; Bektemyssova, G.; Uskenbayeva, R.K. Applications of Machine Learning Algorithms to the Problem of Detecting Unknown Data. *J. Theor. Appl. Inf. Technol.* **2019**, *97*, 1948–1958.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.