Civil Engineering Infrastructures Journal xx,xx
DOI: 10.22059/ceij.2025.385897.2189

Unveiling Spatiotemporal Patterns in Public Transportation: A Smart Card Data Analysis

Shariat Radfar [a], Hamidreza Koosha [b], Ali Gholami [c]* and Atefeh Amindoust [d]

[a] *Department of Industrial Engineering, Na.C, Islamic Azad University, Najafabad, Iran; Email: shariatradfar@sin.iaun.ac.ir*

[b] *Department of Industrial Engineering, Faculty of Engineering, Ferdowsi University of Mashhad, Iran; Email: koosha@um.ac.ir*

[c]* *Department of Civil Engineering, Faculty of Engineering, Golestan University, Gorgan, Iran, Email: a.gholami@gu.ac.ir (corresponding author)*

[d] *Department of Industrial Engineering, Na.C, Islamic Azad University, Najafabad, Iran; Email: aamindoust@iau.ac.ir*

Received: 24/11/2024

Revised: 19/04/2025

Accepted: 05/07/2025

Abstract

Travel patterns in public transportation systems are shaped by a complex interplay of spatial and temporal factors, including passenger location, time of day, and urban infrastructure. Unraveling these patterns is crucial for effective planning and development. This study delves into zonal-based public transport travel behavior in Mashhad, Iran, leveraging smart card data from bus and metro systems. K-means clustering unveils distinct temporal patterns (morning, noon, evening) in passenger trips across 253 traffic zones.

Meanwhile, Mean Shift clustering explores the spatial dimension, analyzing population density and built-up areas within each zone. The analysis yields unique clusters for both temporal and spatial aspects, highlighting the intricate relationship between travel patterns, demographics, and land use. Notably, the study confirms correlations between: morning trips and residential areas, midday trips and commercial/educational areas, and the transactions of marginal traffic zones with near outer residential areas. These insights provide valuable knowledge for policymakers and planners in optimizing land-use and transportation strategies.

Keywords:
Smart Card Data (SCD);
Temporal and Spatial Patterns;
Clustering of Traffic Zones;
K-Means Algorithm;
Mean Shift Algorithm;

1- Introduction

Understanding people's travel behaviors and activity patterns is important in addressing social justice and effectively supporting urban traffic planning and traffic flow prediction as well as the operation of public transportation systems (Lin et al., 2020; Tang et al., 2020; Zhang et al., 2021). People's travel patterns are not uniform. These patterns are usually characterized by specific features such as spatial and temporal diversity of passengers, population density, land use, and access to public transportation. By using these features, analyses such as passenger identification and clustering and the evolution of movement patterns can be performed (Cheng et al., 2020; Li et al., 2021). The pattern of movement and travel behaviors of individuals in a city are closely related to urban structures (Li et al., 2018; Kim et al., 2018). In a longer time frame, the mobility patterns of urban travelers can be changed if the land use structure changes. This may cause unexpected change and in some cases increase the risk of overloading the urban system. Therefore, an effective mobility pattern prediction method is needed to assess the long-term effect of changing the land use structure of an area (Qi et al., 2018).

Data collected from smart cards provides an opportunity for researchers to analyze large data sets and extract meaningful information from them (Kim et al., 2018). One of the important advantages of these cards is improving the quality of travel data with a large sample volume, which means better spatial and temporal coverage (Chen et al., 2019). Smart card data of the public transport system has provided an opportunity to track activities and travel patterns and the spatial and temporal distribution of public transport demand (Kim et al., 2018; Lin et al., 2020). Considering some of the characteristics and shortcomings of smart card data, such as the lack of personal information of travelers and their travel purposes, some researchers have combined smart card data with data from other sources such as travel survey data, land use characteristics, demographic characteristics, socio-economic indicators, days of the week, holidays and other factors such as weather to examine the travel patterns of public transport passengers. Combining the advantages of smart card data with other data improves the accuracy and interpretability of models (Lin et al., 2020; Marinas-Collado et al., 2022).

The development of information technology has caused an explosive growth in the volume of data and made it possible to explore different aspects of mobility in public transport such as changing travel behavior, the perceived and actual differences in travel behavior and reaction to travel motivations, traveler loyalty, community well-being, etc. by using data mining tools in smart card data and other data sources (Briand et al., 2016).

Two widely used concepts for data analysis and problem solving in machine learning and data mining problems are the concepts of classification and clustering. Classification is a supervised learning and clustering is an unsupervised learning that analyzes data objects without consulting the class labels. Clustering can be used to generate class labels for a group of data. Clustering is important for personal passenger services, improving demand models, methods for extracting information about mobility patterns, and various other applications in transportation (Briand et al., 2016; Cheng et al., 2020). Several studies have been conducted on the use of smart card data in urban mobility patterns. A station-centric operational perspective and a passenger-centric perspective are used for clustering mobility patterns in public transportation systems based on smart card data (El-Mahrsi et al., 2017).

To discover passenger behavior and temporal and spatial travel patterns, Zhao et al. (2014), Kim et al. (2017) and Zhang et al. (2021) used the K-Means clustering method in their studies, while Ma et al. (2017), and Medina (2018) used the DBSCAN method. Additionally, Lee & Hickman (2014) and Long & Thill (2015) used the decision tree method, Briand et al. (2016) and Yu et al. (2017) used the Gaussian mixture method, Cheng et al. (2020), Lin et al. (2020), and Liu et al. (2020) used the Dirichlet latent allocation method in clustering and pattern extraction in buses or subways. Mariñas-Collado et al. (2022) in Salamanca bus transport chose the hierarchical clustering algorithm for travel pattern analysis. K-Spectral Centroid (Kim et al., 2018), K-Medoids (Zhao et al., 2019), NCP decomposition (Tang et al., 2020), and Tucker3 decomposition (Frutos-Bernal et al., 2022) are among other methods in this research area. In other studies, to predict and identify regional travel patterns of passengers, Qi et al. (2018) in Beijing proposed and implemented an integrated multi-step method of inner-restricted fuzzy C-means clustering, non-negative tensor factors, and artificial neural network. Also, Wang et al. (2020) and Li et al. (2021) in Sydney used neural networks and deep learning for this purpose. In another study, Zhou et al. (2017) presented a method for inferring the performance of urban areas of Wuhan, China at the metro station level based on the persistent activity patterns obtained from DBSCAN and K-Means. Faroqi et al. (2018) measured the similarities between passenger activities through probabilistic decision trees in the Brisbane Queensland bus and metro system in Australia. El Mahrsi et al. (2014) used the unigram model to analyze the distribution of socio-economic characteristics on temporal clusters in public transport in Rennes, France, and were able to identify different occupational groups commuting between home and work at different times of the day. Ding et al. (2019) investigated the non-linear effects of built environmental characteristics around metro stations on passenger number prediction through gradient boosting decision trees. Huang et al. (2020) used the DBSCAN method to obtain bus arrival time in Suzhou, China. Zhou et al. (2021) in Wuhan Metro used the random forest algorithm to examine changes in individual activity patterns at home and work. In the study of Zhao et al. (2021), using K-Means++, passenger travel characteristics including travel time, travel demand, and travel purposes were presented in bus and metro travel in Beijing through a visual analysis system to identify passenger mobility correlations based on their interests in group and individual forms.

This research delves into public transportation ridership patterns in Mashhad city. We employ a clustering approach to identify patterns, considering both temporal features (morning, noon, evening) and spatial characteristics (population density, built-up area types) within traffic zones. This approach, which clusters traffic zones based on district-total level built-up areas, expands the understanding of spatiotemporal travel patterns beyond station-centric methods. K-means and Mean Shift clustering algorithms are used to analyze ridership patterns. We incorporate spatial data on population and built-up area types (residential, commercial, educational, etc) alongside ridership data for different time intervals. The study's findings aim to inform improvements in public transportation, such as optimized fleet allocation, better land-use distribution across districts, and reduced unnecessary trips, ultimately leading to a more efficient system.

The remainder of this paper is structured as follows. Section 2 details the methodological framework for our traffic zone clustering model, considering both temporal and spatial data. Section 3 describes the research environment, data preparation, and results for each of the clustering models employed. Section 4 analyzes the combined findings from the different models. Section 5 presents conclusions and highlights potential avenues for future research.

2- Materials and Methods

The research framework investigates public transport user behavior using data classification and clustering techniques. The analysis incorporates both temporal (time-based) and spatial (location-based) dimensions to provide insights for:

- *Metropolitan Planning:* By understanding travel patterns, city planners can make informed decisions about land-use allocation. This could involve strategically placing residential areas near public transport hubs, commercial districts near high-traffic zones, or educational institutions within easy reach for students.
- *Transportation Operational Planning:* This analysis can inform strategies for optimizing public transport operations. Insights can be used for matching the number of vehicles on a route with passenger demand during different time periods (e.g., having more buses running during rush hour) and optimizing service schedules (e.g., increasing frequency during peak hours or adjusting schedules based on spatial variations in ridership).

Data classification and clustering techniques will be employed to identify distinct patterns in travel behavior based on factors like temporal features (time of day, day of the week) and spatial characteristics (passenger location, population density, land use). By analyzing these patterns, the research aims to reveal the relationship between travel behavior, time periods, and spatial contexts. Figure (1) shows the general framework diagram of this research.

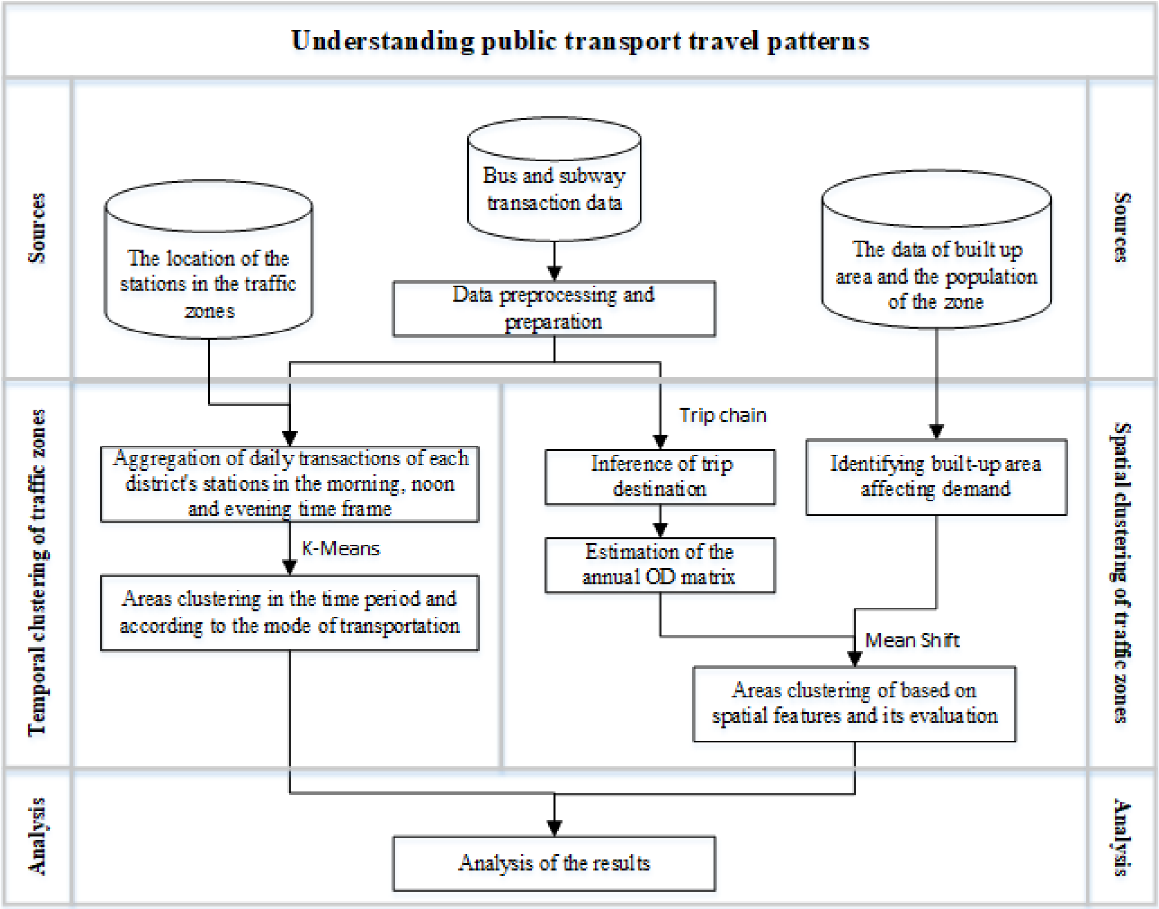

Understanding public transport travel patterns
Sources
Sources
Bus and subway transaction data
The location of the stations in the traffic zones
Data preprocessing and preparation
The data of built up area and the population of the zone
Temporal clustering of traffic zones
Spatial clustering of traffic zones
Trip chain
Aggregation of daily transactions of each district's stations in the morning, noon and evening time frame
Inference of trip destination
Identifying built-up area affecting demand
K-Means
Estimation of the annual OD matrix
Areas clustering in the time period and according to the mode of transportation
Mean Shift
Areas clustering of based on spatial features and its evaluation
Analysis
Analysis
Analysis of the results

Figure (1) The general framework diagram of this research

2-1- Research environment and data

Mashhad is the second-largest city in Iran with a population exceeding 3 million. The city is divided into 253 traffic zones. The study leverages transaction data from the "Man-Card" system, a rechargeable AFC smart card used for public transportation fare payments in Mashhad. Passengers tap their cards on payment devices when entering buses or metro stations. Transaction information was obtained through Mashhad Municipality for this research. Table (1) shows the summary of public transportation information of Mashhad in March 2019 - 2020. Figure (2) also shows the bus and subway routes.

Table (1) Summary of public transport information in Mashhad, from March 2019 - 2020

Title	Bus	Metro
Number of active lines	136	2

Number of active station	3511	33
Active route length (km)	2162	43.5
The number of urban transport fleets	2250	-
Number of active Man-Cards (per year)	2,315,017	
Number of transactions (per year)	261,967,075	

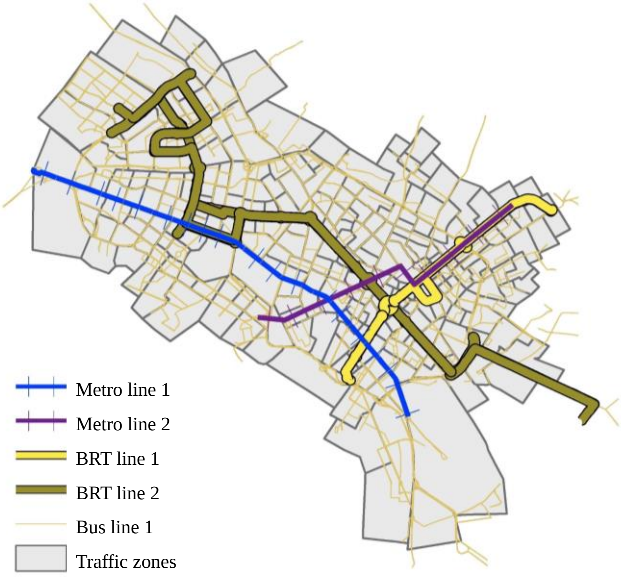


Figure (2) The bus and subway routes

This research incorporates various data sources beyond smart card transactions (Man-Card data) to enhance the prediction of public transport travel demand. These additional sources provide a more comprehensive understanding of travel patterns within the city. Table (2) summarizes the information collected from these sources. By combining smart card transaction data with these additional sources, the research aims to create a more robust model for predicting public transport travel demand across different traffic zones.

Table (2) Summary of the information collected and used to predict the demand

Title	**Important Features**	**Time Period**	**Source**
Daily OD matrices	Zonal transit production and attraction	from March 2019 - 2020	Current study
Land use data	Types of land uses, number, area, density, built up area	from March 2019 - 2020	Department of Planning and Architecture
Traffic zone data	Zone borders, available stations	from March 2019 - 2020	Mashhad Transportation Network Engineering and Management Organization
Population data	Population of different zones of the city, Number of households	from March 2016 - 2017	Iran Statistical Center

2-2- Clustering Traffic Zones based on Temporal Ridership Patterns

Traffic zones were clustered based on temporal patterns in public transport ridership. Understanding these patterns provides valuable insights for transportation decision-makers. The analysis utilizes public transport transaction data (e.g., passenger entries or exits) across various time periods for each traffic zone. The well-known K-means clustering algorithm (McQueen, 1967) was used due to its effectiveness as a distance-based method. K-means groups data points based on their similarity to cluster centroids. To identify the optimal number of clusters (k), two approaches were used:

- ***Elbow Method:*** This method visually analyzes the sum of squared distances within each cluster (intra-cluster distance) plotted against the number of clusters. The "elbow" point on the curve indicates the most suitable number of clusters (Shi et al., 2021).
- ***Silhouette Score:*** This score measures how well a data point is placed within its assigned cluster compared to neighboring clusters (Shahapure and Nicholas, 2020). Higher silhouette scores indicate better cluster separation.

The data used in this section are the transaction time and transaction station data from the bus and metro fare collection system, as well as the traffic zone information of different metro and bus stations. In order to preprocess the transaction data for clustering traffic zones based on temporal patterns and transaction counts, several steps are undertaken. The data originates from the bus and metro fare collection system, includes transaction time and transaction station. Traffic zone information for stations is also used to link transactions to specific zones.

In the cleaning of the data, irrelevant features that are not used in the clustering process are removed. This ensures the focus remains on the information crucial for clustering traffic zones effectively. For example, details like passenger identification or ticket type might be excluded if not relevant to zone analysis. Next, cash payment transactions made by bus drivers are specifically excluded. The rationale behind this exclusion is twofold. Firstly, these transactions are often replaced with the passenger's cash payment, making it difficult to pinpoint the exact boarding time and location. Secondly, the absence of this information for bus driver transactions introduces inconsistencies when compared to passenger transactions.

Once the data is cleaned, it is further processed for efficient clustering. Separate tables are created for bus and subway transactions. This separation allows for tailored analysis considering potential differences in travel patterns between the two modes of public transport. Within each table, the transactions are then sorted by transaction time in ascending order. This chronological arrangement facilitates the subsequent aggregation steps. The traffic zone for each transaction is identified using the corresponding station's information. This step links the transaction data with the specific geographic zones it pertains to, enabling the clustering process to consider the spatial distribution of public transport usage. Bus and subway flows were separated due to differences in travel patterns, operational schedules and headways, capacity and demand, spatial accessibility and coverage, and organizational structure and planning.

After sorting and zone identification, the transaction data undergoes temporal aggregation to capture travel patterns across time intervals. This aggregation is achieved in two stages:

1- Hourly Aggregation: The number of transactions for each station is aggregated by one-hour intervals for each day. This provides a detailed picture of how transaction volume fluctuates throughout the day at each station.

2- Intra-Zone Aggregation: Following the hourly aggregation for individual stations, transactions between stations located within the same traffic zone are further aggregated. Here too, one-hour intervals are considered. This step identifies how frequently transactions occur within a zone, reflecting the internal movement of passengers within that specific zone.

To account for potential variations in travel behavior across different times of the day, the final step involves period aggregation. Cumulative transaction values for each zone are calculated for three distinct periods: morning, noon, and evening. The definitions for these periods are derived from the methodology established in "Determining Public Transportation Fare Collection Policies in Mashhad," a foundational report published by the Mashhad Transportation Network Engineering and Management Organization. Here's a breakdown of the periods used:

- Morning: 00:00 - 11:00
- Noon: 11:00 - 17:00
- Evening: 17:00 - 24:00

This period aggregation allows for the creation of separate zone clusters based on their unique morning, afternoon, and evening travel patterns.

Following these preprocessing steps, the final dataset prepared for clustering traffic zones is obtained. This dataset combines the hourly transaction counts for each zone within each of the three time periods (morning, noon, evening) for both buses and subways. This comprehensive dataset captures the temporal and spatial characteristics of public transport usage within the city, providing a robust foundation for effective traffic zone clustering.

2-3- Clustering Traffic Zones based on Spatial Patterns

Traffic zones have different built-up areas (residential, commercial, educational, etc.) and population densities. These spatial features influence public transportation demand, impacting planning decisions like fleet allocation and land-use distribution. To analyze this relationship, Mean Shift clustering (Jin and Han, 2017) was employed. This method excels because it doesn't require pre-defined clusters or knowledge of their distribution shape (Zhu et al., 2022). It only needs one parameter: bandwidth, which acts as a search radius for data points around a cluster center (Netzer et al., 2020). Selecting an optimal bandwidth poses a significant challenge in clustering analysis. In this study, the mean-shift algorithm from the Scikit-learn Python library is used to cluster traffic zones. The library's built-in function, "sklearn.cluster.estimate_bandwidth", automatically computes the optimal bandwidth by analyzing the data distribution and the pairwise distances between points.

The Mean Shift algorithm iteratively refines clusters by considering each data point as a potential cluster center. It identifies neighboring data points within the defined bandwidth (search radius). Then, it shifts the cluster center towards the average location (mean) of these neighbors. This process repeats until convergence, resulting in well-defined clusters (Mendonça et al., 2024).

To assess the quality and separation of the clusters, two metrics were employed: silhouette score and Davies-Bouldin (Davies and Bouldin, 1979) index. The silhouette score measures how well data points are placed within their assigned clusters compared to neighboring clusters. The Davies-Bouldin Index evaluates the ratio of within-cluster distance to the between-cluster distance, aiming for well-separated clusters.

This methodology aims to identify distinct groups of traffic zones based on built-up areas, population, and public transport demand.

Demand Data: Mashhad's fare collection system only records entry data, lacking information on passenger destinations. To address this limitation, researchers employed the improved trip chain model (Radfar et al., 2025) using one year of available Automatic Fare Collection (AFC) data from March 2019 - 2020. The key improvements including: data comprehensiveness; card retention with only one daily transaction; using real-time arrival times instead of scheduled arrival times; and using data that does not allow destination inference for estimating the OD matrix. Trip chain model infers the destination of a trip by selecting transactions for each card on each day and examining the records of its other transactions on that day. This process is repeated for all cards and all days of the year. This model allowed for destination estimation and the generation of daily and annual public transport origin-destination matrices based on traffic zones, which were then used as demand data.

Spatial Data: Traffic demand is known to be influenced by the characteristics of an area, such as land use and population. Researchers considered the relationship between transportation and land use, acknowledging how travel patterns are closely linked to urban land use patterns. Proper land-use planning is believed to influence overall travel behavior, potentially encouraging residents to shift from private vehicles to public or non-motorized transportation, ultimately improving traffic flow and contributing to sustainable development. Conversely, travel patterns and transportation demand can also influence urban development factors like land prices and facility distribution within a zone (Sarkar and Mallikarjuna, 2013; Hu et al., 2016; Kim et al., 2018; Zhou et al., 2019; Cai et al., 2020; Saleem et al. 2025). Additionally, population growth plays a role in shaping public transportation demand, with population increases leading to higher ridership and potentially influencing commuting patterns (Zhang and Xu, 2022).

For this study, urban land use information was categorized into thirteen distinct groups based on existing land use classification criteria specific to the city of Mashhad. ArcGIS 9.3 software was used to manage the data preparation process. First, shapefile maps for traffic zones, land uses, and population were imported into the software environment. After combining the information from these layers into a single map, each traffic zone was intersected with the corresponding land use and population data. This process allowed researchers to extract relevant data such as the number of regions, area of the region, built-up area, and population within each traffic zone. The attribute table was used to manage this data extraction. Unnecessary features, such as the number and area of land uses not applicable to the study, were removed during the creation of the final spatial data table. Notably, the research focused on the built-up area within each zone, excluding the total area due to the presence of undeveloped land or zones with mixed uses. The data was examined individually for duplicate entries, missing data points, and outliers that might exist for various reasons. These issues were then addressed to clean the data. Additionally, Python was used to convert all data into decimal or integer formats, ensuring compatibility with the chosen model type.

To identify the most influential features affecting the dependent variable (demand), the Random Forest algorithm (Tin Kam Ho, 1995; Zhou et al., 2021) was utilized. This selection process offers several benefits including improved processing speed for the model and prevention of overfitting due to an excessive number of dependent variables.

The Random Forest algorithm, built upon the decision tree algorithm, consists of several hundreds of trees. These trees analyze the impact of each independent variable on the dependent variable, as well as the combined effect of a set of independent variables. By running the algorithm, the following features were identified as having a significant impact:

- Zone Population
- Traffic Zone Area
- Build up area for residential land uses
- Build up area for commercial land uses
- Build up area for educational land uses
- Build up area for residential-commercial land uses

Table (3) displays part of the final data set created. Additionally, Table (4) shows the correlation results of the variables with the dependent variable of demand.

Table (3) Part of the final data set number

Zone Population (person)	Zone Area	Residential (m^2)	Commercial (m^2)	Educational (m^2)	Commercial – Administrative (m^2)	Demand (person)
54	656,422	6,989	68,520	7,474	5,429	621,851
455	321,143	59,264	38,932	15,328	6,640	1,951,430
766	205,456	86,709	3,935	3,127	5,082	13,007
275	211,477	57,503	4,502	0	7,406	229,106
136	238,852	21,877	2,843	0	3,724	1,272,545
...	...	...	...	...	...	...

Table (4) The correlation report of independent variables with demand

Variable name		Correlation with demand
Area of traffic zone		0.0786
Population of traffic zone		0.008
Bui	Residence & Hotel	0.295

	Educational	0.205
	Health	0.096
	Commercial	0.063
	Commercial - Administrative	0.083
	Commercial - Residence	0.354
	Recreational & Sports	0.025
	Religious	0.002
	Residential	0.154
	Residential - Commercial	0.214
	Administrative	0.044
	Military	0.096
	Other land use	0.058

Due to potential variations in data scales and dispersions within the datasets, a standardization process was implemented to improve model prediction accuracy. This process transforms the data such that the mean becomes zero, the standard deviation becomes one, and outliers have a reduced impact on the model.

3- Results and Discussion

3-1- Results of clustering traffic zones based on temporal data and number of transactions

Determining the optimal number of clusters (k) is crucial for the K-Means algorithm. This study employed two methods to identify k: Elbow Method and Silhouette Score.

- ***Elbow Method:*** This method analyzes the sum of squared distances within clusters for various k values. Figure (3) and Figure (4) depict the elbow method results for bus and metro data, respectively. These figures suggest a range of 4 to 6 clusters for both datasets.
- ***Silhouette Score:*** This score measures clustering quality, ranging from -1 (worst) to 1 (best). The Silhouette score was calculated for k values between 2 and 19. Based on the Silhouette scores the optimal k for bus data is 2 due to the highest score (0.710158), and the optimal k for metro data is 7 due to the highest score (0.6594371).

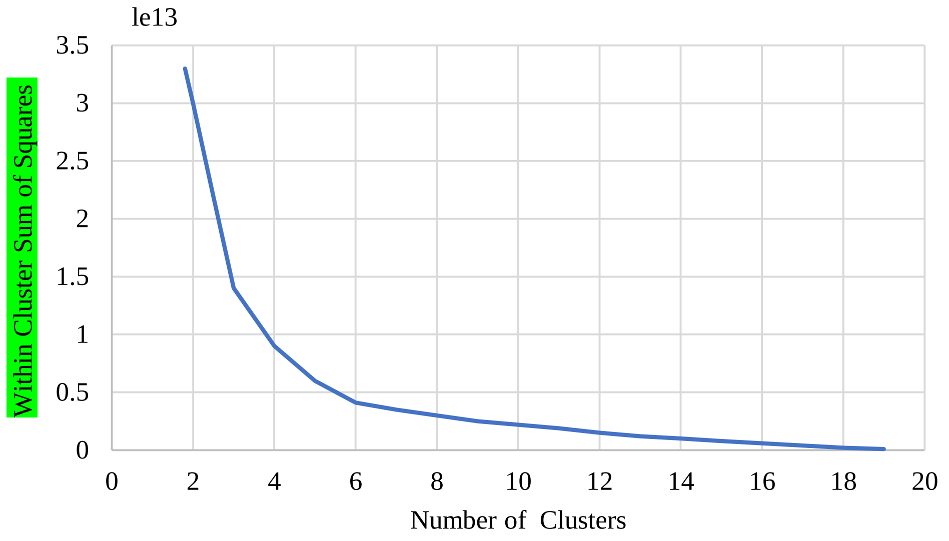


Figure (3) The results of elbow method for bus data

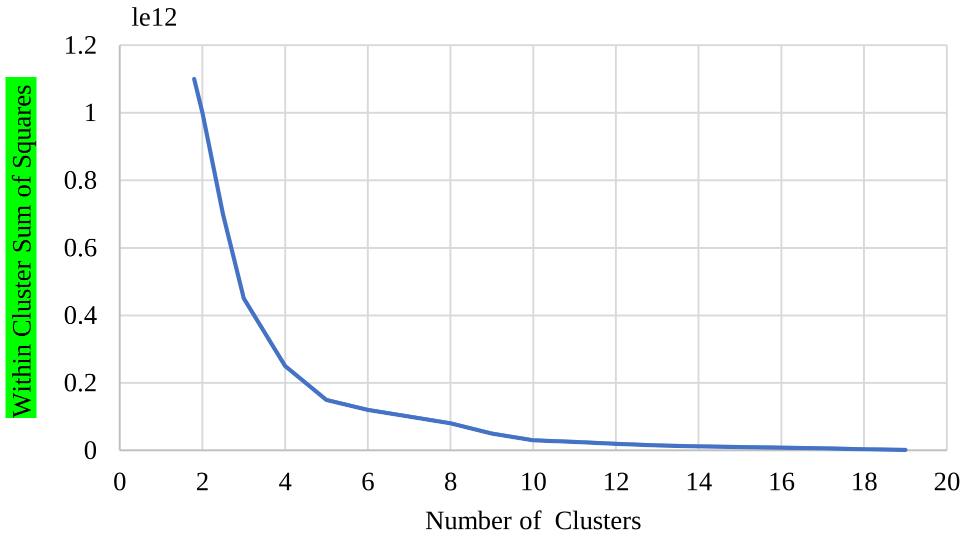


Figure (4) The results of elbow method for metro data

According to the obtained values, the K-Means algorithm divides the traffic zones for bus transportation into two clusters:

- *Cluster 0:* Low transaction (ranging from 290 to 592,300 transactions in March 2019 - 2020 for each time interval: morning, noon and evening).
- *Cluster 1:* High transaction (ranging from 595,900 to 2,680,000 transactions in March 2019 - 2020 for each time interval: morning, noon and evening).

For metro data, the K-Means algorithm divides the traffic zones into seven clusters, representing transaction intervals for morning, noon and evening time intervals. Table (5) specifies the transaction ranges for each metro cluster.

Table (5) Transaction intervals of the clusters formed for metro

Cluster	Cluster Quality range	Lower transaction limit	Upper transaction limit
0	very, very low transactions	19,791	110,345
1	very low transactions	150,709	269,507
2	low transactions	301,427	416,297
3	average transactions	431,967	568,987
4	high transactions	593,465	739,453
5	very high transactions	785,365	1,002,653
6	very, very high transactions	1,221,991	1,453,930

In Table (6) a part of the clustering results of the traffic zones based on the bus transportation data and in Table (7) a part of the clustering results based on the metro data can be seen and the areas not included in the clusters have been removed. In these tables, time "1" represents the morning time interval, time "2" represents the noon time interval, and time "3" represents the evening time interval.

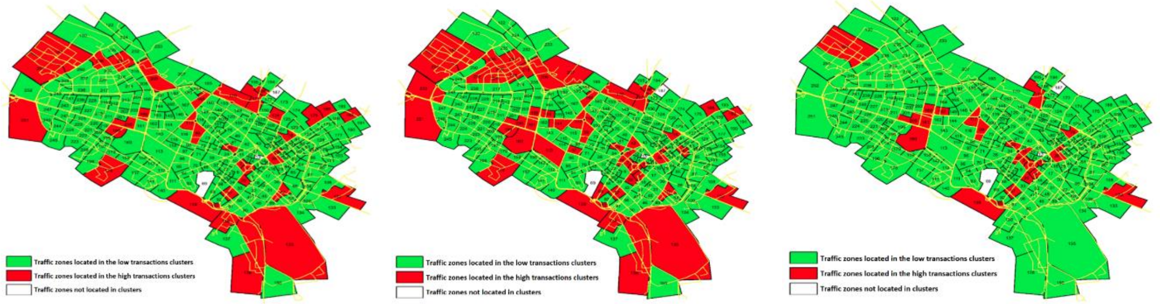
Table (6) A part of the clustering results of the districts based on the bus transportation data

Traffic zone	Time	Cluster	Traffic zone	Time	Cluster	Traffic zone	Time	Cluster
1	1	1	1	2	1	1	3	1
2	1	1	2	2	1	2	3	1
3	1	0	3	2	0	3	3	0
4	1	0	4	2	0	4	3	0
5	1	0	5	2	1	5	3	1
6	1	0	6	2	0	6	3	0
7	1	0	7	2	0	7	3	0
8	1	0	8	2	1	8	3	1
9	1	0	9	2	0	9	3	0
10	1	0	10	2	1	10	3	1
...	...	...	...	...	...	...	...	...

Table (7) A part of the clustering results of the districts based on the metro transportation data

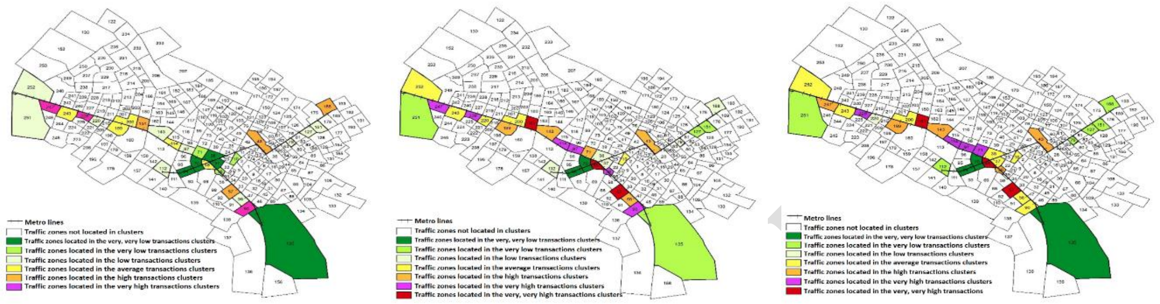
Traffic zone	Time	Cluster	Traffic zone	Time	Cluster	Traffic zone	Time	Cluster
10	1	2	10	2	4	10	3	4
36	1	2	36	2	6	36	3	5
37	1	1	37	2	3	37	3	4
38	1	1	38	2	3	38	3	4
43	1	5	43	2	5	43	3	5
48	1	4	48	2	7	48	3	7
56	1	3	56	2	5	56	3	4
57	1	5	57	2	7	57	3	7
70	1	1	70	2	1	70	3	1
...	...	...	...	...	...	...	...	...

The zonal clusters in the morning, noon and evening periods for bus and metro are displayed in **Error! Reference source not found.**) and (6).


Traffic zones located in the low transactions clusters
Traffic zones located in the high transactions clusters
Traffic zones not located in clusters

a. Morning period b. Noon period c. Evening period

Figure (5) The spatial distribution of clusters of traffic zones for the bus


Metro lines
Traffic zones not located in clusters
Traffic zones located in the very, very low transactions clusters
Traffic zones located in the very low transactions clusters
Traffic zones located in the low transactions clusters
Traffic zones located in the average transactions clusters
Traffic zones located in the high transactions clusters
Traffic zones located in the very high transactions clusters
Traffic zones located in the very, very high transactions clusters

a. Morning period b. Noon period c. Evening period

Figure (6) The spatial distribution of clusters of traffic zones for the metro

3-2- Results of clustering traffic zones based on spatial data and demand

This study employed Python to perform traffic zone clustering. The Mean Shift algorithm was chosen for this purpose. The optimal bandwidth parameter was calculated using the estimation bandwidth algorithm in python library and based on the specific data employed in this analysis to be 2.18. It was a crucial input for the Mean Shift algorithm.

The Mean Shift classified traffic zones into eight distinct clusters. Figure (7) illustrates the cluster of each traffic zone.

To assess the quality of the clustering results, two well-established evaluation indices were utilized: the Silhouette Index and the Davies-Bouldin Index. The Silhouette Index measures the resolution or quality of the clusters, with a positive value indicating good separation. The Davies-Bouldin Index scores better for well-separated and compact clusters, with lower values signifying better clustering outcomes.

The Silhouette Index obtained a value of 0.42, while the Davies-Bouldin Index reached a value of 0.73. Considering the positive Silhouette Index value and the Davies-Bouldin Index value close to zero, we can conclude that the Mean Shift algorithm achieved a satisfactory clustering of traffic zones. These evaluation metrics suggest that the identified clusters exhibit reasonable separation and compactness, supporting the validity of the clustering process.

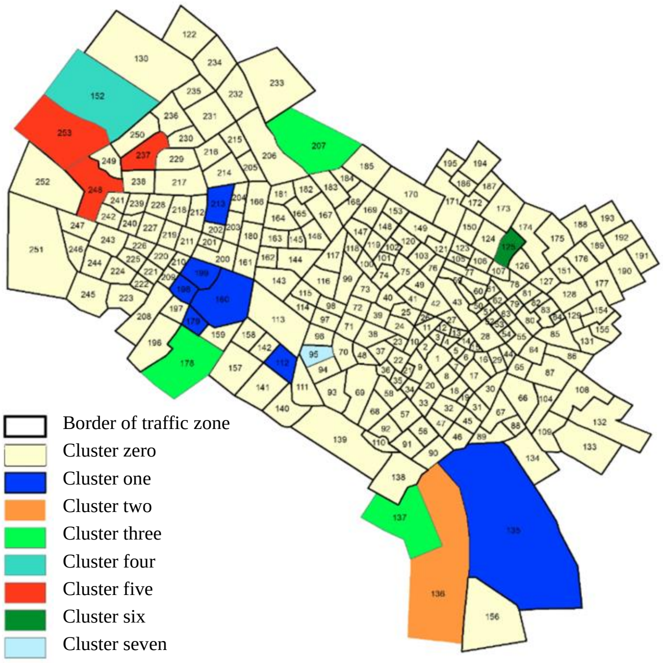


Figure (7) Spatial location of clustering of traffic zones

3-3- Discussion

This section explores the temporal and spatial patterns within the identified traffic zone clusters using demographic and land-use data, alongside annual origin-destination (OD) matrices.

3-3-1- Temporal patterns

Bus Trips: Analyzing bus transactions across three time intervals revealed distinct patterns:

- *Low Transaction:* 197 (78%) of the 253 traffic zones exhibited low transactions across all periods.
- *High Transaction:* 56 (22%) zones displayed high transactions in at least one time interval.
- *Variability:* 23 zones showed differing transaction levels between morning and afternoon.

A total of 35, 52, and 24 zones belonged to the "highly transactional" category in the morning, noon, and evening clusters, respectively. Notably, 17 zones consistently fell into the "highly transactional" category across all three time intervals. These zones were found to share at least one of the following characteristics:

- Presence of a Bus Rapid Transit (BRT) line offering high-speed travel through dedicated lanes.
- Location of a bus transfer stop facilitating passenger interchange from multiple routes.

Figure (8) visually depicts the location of these high-transaction zones.

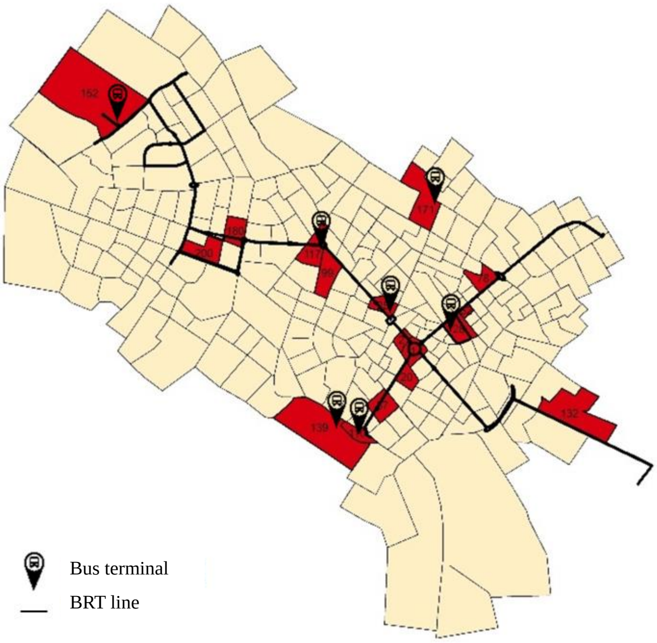


Figure (8) The location of high-transaction traffic zones

Further analysis focused on zones categorized in both the morning and evening bus clusters (16 zones, Figure (9)). These zones were primarily located in the city's peripheral areas and exhibited high resident populations (13 zones ranked in the top three deciles). This suggests their role as origin points for passengers traveling to other areas in the morning and return trip destinations in the evening.

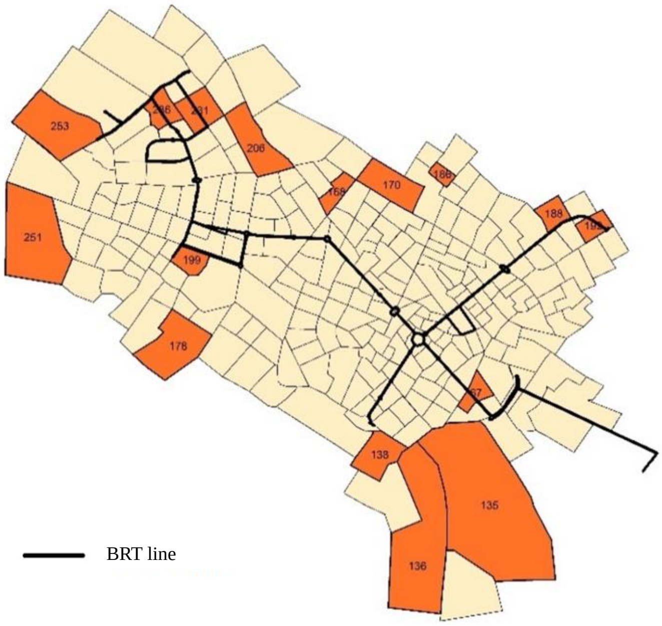


Figure (9) The location of bus high-transaction traffic zones in the morning and afternoon

Traffic zones 99, 132, 152, 171, 180, and 200 exemplified this pattern, demonstrating a strong correlation between population density and land uses in high travel production/demand, reflected in their morning cluster membership. Additionally, their top three deciles ranking in residential land use further supported this association. Conversely, educational or commercial land uses within these zones, also ranked in the top three deciles, likely contributed to high transactions during the noon and evening clusters, signifying trip ends associated with these purposes. Examining the annual OD matrix corroborated this analysis, indicating a higher frequency of intra-zonal trips within these zones compared to trips originating or terminating in other zones.

Metro Trips: All traffic zones situated along the metro route belonged to the top three deciles in terms of annual trip production. Interestingly, some of these zones might be categorized as "low transaction" in the bus service clusters, which typically offer broader zonal coverage.

Nine zones consistently exhibited above-average transaction levels across all three time periods (Figure (10)). No single defining characteristic emerged for all these zones. However, individual analysis is recommended to pinpoint the specific reasons for their high transaction patterns. For instance, traffic zones 43 and 161 possessed high rankings (top decile) in population and built residential, commercial, and educational land uses. Additionally, zone 43 benefited from the presence of a "suburban railway station" adjacent to the metro station, while zone 161 housed "Mellat Park" and a bus terminal near the metro station, potentially contributing to their consistently high transactions. The intersection of metro lines 1 and 2 in traffic zone 48 might also explain its inclusion in this set of high-transaction zones.

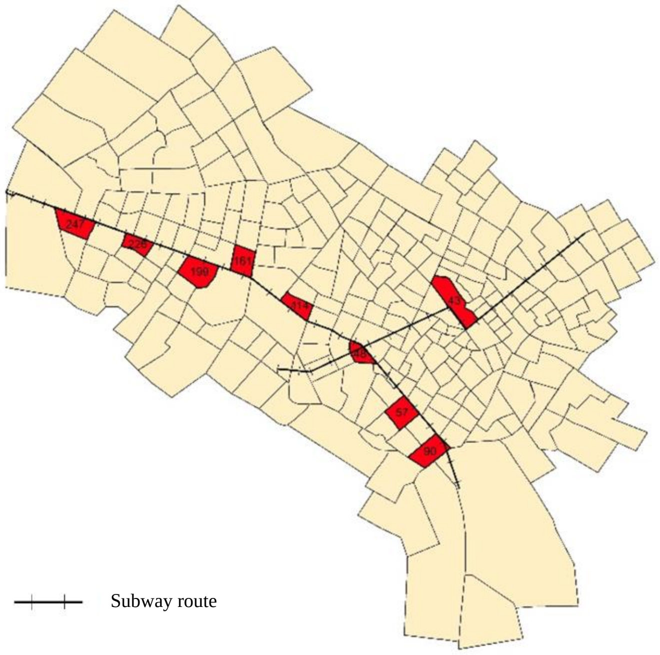


Figure (10) Location of metro traffic zones with more than average transaction

Four zones (36, 71, 97, and 143) displayed contrasting transaction patterns – lower than average in the morning cluster and higher than average in the noon and evening clusters. Examination of land uses revealed that the built commercial areas within each of these zones belonged to the top three deciles. The start and end of business hours in these commercial zones are the cause of the observed variations in transaction levels across different time clusters.

Notably, traffic zone 57 emerged as a unique case exhibiting consistently high transaction levels across all bus and metro temporal clusters. This zone ranked second in terms of travel production and first in annual travel attraction. While its population (6th decile) and residential/commercial land uses (9th decile) were not the most prominent factors, its strategic location played a critical role. Traffic zone 57 benefits from a segment of BRT line 1 passing through it, providing direct access to the city's most important tourist destination, "the holy Shrine of Imam Reza." Additionally, the zone's intersection with metro line 1 further strengthens its connectivity and contributes to its high travel demand. This finding highlights the significant influence of major public transportation infrastructure and key attractions on traffic patterns within specific zones.

3-3-2- Spatial patterns

The spatial analysis focused on identifying cluster characteristics based on built-up area, population, and travel demand. The clusters have the following characteristics:

- *Cluster 0 (236 members):* Diverse land uses, populations, and demands.
- *Cluster 1 (7 members):* High educational land use (10th decile).
- *Clusters 2 & 4 (marginal areas):* Located in the 10th decile for all investigated land uses and population, and also high travel demand.
- *Cluster 3 (3 members):* 10th decile for population and built residential and commercial use, with high moderate demand (6th and 8th deciles).
- *Cluster 5 (similar to Cluster 3):* High educational land use instead of commercial land use (10th decile).
- *Cluster 6 (1 member):* High commercial land use (10th decile), moderate population (9th decile), and moderate demand (7th decile).
- *Cluster 7 (1 member):* High population and built residential use (top 10th decile) but low demand (1st decile).

This analysis highlights the influence of factors like land use, population, BRT lines, and major attractions on the spatial and temporal patterns of traffic zone demand.

Applying the random forest algorithm to identify the most effective independent variables on public transport demand confirms the findings of Zhang and Xu (2022) for population and Sarkar and Mallikarjuna (2013), Hu et al. (2016), Kim et al. (2018), Zhou et al. (2019), Cai et al. (2020), Liu et al. (2021) and Cheng et al. (2024) for the most effective types of land use. Findings reported in a number of studies are consistent with the results of this research and confirm its findings, such that Ding & Lu (2016) and Kim et al. (2018) consider land use patterns to have a significant impact on people's travel behavior. They also believe that population and land use diversity will have positive effects on the number of trips. Ma et al. (2017) and Gao et al. (2018) attribute the imbalance between job opportunities in central city areas and residence in suburban areas, as well as low housing prices in these areas, to the generation of trips by transit passengers. Bautista-Hernández (2020) shows that for public transport users, job accessibility, population density, and mixed land use at the origin have a greater impact on the number of trips. Briand et al. (2016) interpreted areas where people mainly depart in the morning as residential and those that depart in the evening as workplaces. Qi et al. (2018) stated that the characteristics of an area and its travel attractors are an important and influential factor in the travel patterns of that zone.

4- Conclusion

This study employed zonal-based clustering analysis to investigate the temporal and spatial patterns of public transportation travel. Smart card data and the annual origin-destination (OD) matrix for public transportation provided the foundation for this analysis.

Smart card transaction data offered valuable insights into real-world network usage, capturing individual user trips at specific times and locations. This data facilitated the temporal clustering of traffic zones across morning, noon, and evening public transport periods. Clustering was based on entry transactions within each zone. K-Means algorithm has been used to perform temporal clustering of zones.

Spatial clustering of traffic zones utilized the Mean Shift algorithm and leveraged travel demand (attraction) alongside spatial variables. These spatial variables included population and built-up area data for various land uses (e.g. residential, commercial, and educational) within each zone.

Key Innovation: This research distinguished itself by shifting the focus of clustering from stations to traffic zones. Furthermore, instead of analyzing location characteristics around stations, it examined total population and built-up land uses within the traffic zones themselves. This approach broadens the scope of pattern recognition from the station level to the entire network and city level. Using one-year fare transaction data, inferred trip destination and public transport OD matrix were generated as one of the inputs of the clustering process.

Implications: The findings from this study hold significant value for transportation and land-use decision-makers. By revealing temporal and spatial demand patterns, the research can inform policies that:

- Optimize the distribution of land use types across different zones.
- Mitigate unnecessary trips.
- Allocate appropriate fleet sizes for public transportation.
- Ultimately aim to improve service delivery and user satisfaction.

This research contributes to the field by advancing the understanding of public transportation demand patterns at the traffic zone level, providing valuable insights for policymakers to optimize service delivery and network planning.

References

Bautista-Hernández, D. (2020). Urban Structure and Its Influence on Trip Chaining Complexity in the Mexico City Metropolitan Area. *Urban, Planning and Transport Research, 8*(1), 71-96. doi:10.1080/21650020.2019.1708784

Briand, A. S., Côme, E., El Mahrsi, M. K., & Oukhellou, L. (2016). A mixture model clustering approach for temporal passenger pattern characterization in public transport. *International Journal of Data Science and Analytics, 1*(1), 37-50. doi:10.1007/s41060-015-0002-x

Cai, Z., Li, T., Su, X., Guo, L., & Ding, Z. (2020). Research on Analysis Method of Characteristics Generation of Urban Rail Transit. *IEEE Transactions on Intelligent Transportation Systems, 21*(9), 3608-3620. doi:10.1109/TITS.2019.2929619

Chen, E., Ye, Z., Wang, C., & Zhang, W. (2019). Discovering the spatio-temporal impacts of built environment on metro ridership using smart card data. *Cities, 95*, 102359. doi:10.1016/j.cities.2019.05.028

Cheng, J., Liu, G., Gao, S., Raza, A., Li, J., & Juan, W. (2024). Short-Term Passenger Flow Prediction in Urban Rail Transit Based on Points of Interest. *IEEE Access, 12*, 95196 - 95208. doi:10.1109/ACCESS.2024.3425634

Cheng, Z., Trépanier, M., & Sun, L. (2020). Probabilistic model for destination inference and travel pattern mining from smart card data. *Transportation*. doi:10.1007/s11116-020-10120-0

Davies, D., & Bouldin, D. (1979). A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-1*(2), 224-227. doi:10.1109/TPAMI.1979.4766909

Ding, C., Cao, J., & Liu, C. (2019). How does the station-area built environment influence Metrorail ridership? Using gradient boosting decision trees to identify non-linear thresholds. *Journal of Transport Geography, 77*, 70-78. doi:10.1016/j.jtrangeo.2019.04.011

Ding, Y., & Lu, H. (2016). Activity participation as a mediating variable to analyze the effect of land use on travel behavior: A structural equation modeling approach. *Journal of Transport Geography, 52*, 23-28. doi:10.1016/j.jtrangeo.2016.02.009

El Mahrsi, M., Côme, E., Baro, J., & Oukhellou, L. (2014). Understanding Passenger Patterns in Public Transit Through Smart Card and Socioeconomic Data: A Case Study in Rennes, France. *The 3rd International Workshop on Urban Computing (UrbComp 2014)*, (p. 9). New York.

El Mahrsi, M., Côme, E., Oukhellou, L., & verleysen, M. (2017). Clustering Smart Card Data for Urban Mobility Analysis. *IEEE Transactions on Intelligent Transportation Systems , 18*(3), 712-728. doi:10.1109/TITS.2016.2600515

Faroqi, H., Mesbah, M., Kim, J., & Tavassoli, A. (2018). A model for measuring activity similarity between public transit passengers using smart card data. *Travel Behaviour and Society, 13*, 11-25. doi:10.1016/j.tbs.2018.05.004

Frutos-Bernal, E., del Rey, Á., Mariñas-Collado, I., & Santos-Martín, M. (2022). An Analysis of Travel Patterns in Barcelona Metro Using Tucker3 Decomposition. *Mathematics, 10*(7), 1122. doi:10.3390/math10071122

Hu, N., Legara, E. F., Lee, K. K., Hung, G. G., & Monterola, C. (2016). Impacts of land use and amenities on public transport use, urbanplanning and design. *Land Use Policy, 57*, 356-367. doi:10.1016/j.landusepol.2016.06.004

Huang, D., Yu, J., Shen, S., Li, Z., Zhao, L., & Gong, C. (2020). A Method for Bus OD Matrix Estimation Using Multisource Data. *Journal of Advanced Transportation, 2020*(1), 1-13. doi:10.1155/2020/5740521

Jin, X., & Han, J. (2017). Mean Shift. In C. Sammut, & G. Webb, *Encyclopedia of Machine* Learning and Data Mining (pp. 806-808). Boston, MA: Springer. doi:10.1007/978-1-4899-7687-1_532

Kim, M.-K., Kim, S., & Sohn, H.-G. (2018). Relationship between Spatio-Temporal Travel Patterns Derived from Smart-Card Data and Local Environmental Characteristics of Seoul, Korea. *sustainability, 10*(3), 787. doi:10.3390/su10030787

Kim, M.-K., Kim, S.-P., Heo, J., & Sohn, H.-G. (2017). Ridership patterns at subway stations of Seoul capital area and characteristics of station influence area. *KSCE Journal of Civil Engineering, 21*(3), 964–975. doi:10.1007/s12205-016-1099-8

Lee, S., & Hickman, M. (2014). Trip purpose inference using automated fare collection data. *Public Transport, 6*, 1-20. doi:10.1007/s12469-013-0077-5

Li, C., Bai, L., Liu, W., Yao, L., & Waller, S. (2021). Urban Mobility Analytics: A Deep Spatial-Temporal Product Neural Network for Traveler Attributes Inference. *Transportation Research Part C: Emerging Technologies, 124*. doi:10.1016/j.trc.2020.102921

Li, J., Kim, C., & Sang, S. (2018). Exploring impacts of land use characteristics in residential neighborhood and activity space on non-work travel behaviors. *Journal of Transport Geography, 70*, 141-147. doi:10.1016/j.jtrangeo.2018.06.001

Lin, P., Weng, J., Alivanistos, D., Ma, S., & Yin, B. (2020). Identifying and Segmenting Commuting Behavior Patterns Based on Smart Card Data and Travel Survey Data. *sustainability, 12,* 5010. doi:10.3390/su12125010

Liu, J., Shi, W., & Chen, P. (2020). Exploring Travel Patterns during the Holiday Season—A Case Study of Shenzhen Metro System During the Chinese Spring Festival. *International Journal of Geo-Information, 9*(11), 651. doi:10.3390/ijgi9110651

Liu, X., Wu, J., Huang, J., Zhang, J., Chen, B., & Chen, A. (2021). Spatial-interaction network analysis of built environmental influence on daily public transport demand. *Journal of Transport Geography, 92*. doi:10.1016/j.jtrangeo.2021.102991

Long, Y., & Thill, J.-C. (2015). Combining smart card data and household travel survey to analyze jobs–housing relationships in Beijing. *Computers, Environment and Urban Systems, 53*, 19-35. doi:10.1016/j.compenvurbsys.2015.02.005

Ma, X.-l., Liu, C., Wen, H., Wang, Y., & Yao-Jan, W. (2017). Understanding commuting patterns using transit smart card data. *Transport Geography*, 135-145. doi:10.1016/j.jtrangeo.2016.12.001

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* (Vol. 1, pp. 281-297).

Mariñas-Collado, I., Sipols, A., Santos-Martín, M., & Frutos-Bernal, E. (2022). Clustering and Forecasting Urban Bus Passenger Demand with a Combination of Time Series Models. *Mathematics, 10*(15), 2670. doi:10.3390/math10152670

Medina, S. (2018). Inferring weekly primary activity patterns using public transport smart card data and a household travel survey. *Travel Behaviour and Society , 12*, 93-101. doi:10.1016/j.tbs.2016.11.005

Mendonça, M., Netto, S., Diniz, P., & Theodoridis, S. (2024). Chapter 13 - Machine learning: Review and trends. In *Signal Processing and Machine Learning Theory* (pp. 869-959). doi:10.1016/B978-0-32-391772-8.00019-3

Netzer, M., Michelberger, J., & Fleischer, J. (2020). Intelligent Anomaly Detection of Machine Tools based on Mean Shift Clustering. *Procedia CIRP, 93*, 1448-1453. doi:10.1016/j.procir.2020.03.043

Qi, G., Huang, A., Guan, W., & Fan, L. (2018). Analysis and Prediction of Regional Mobility Patterns of Bus Travellers Using Smart Card Data and Points of Interest Data. *IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, 20*, 1197–1214. doi:10.1109/TITS.2018.2840122

Radfar, S., Koosha, H., Gholami, A., & Amindoust, A. (2025). A neuro-fuzzy and deep learning framework for accurate public transport demand forecasting: Leveraging spatial and temporal factors. *Journal of Transport Geography*, 126, 104207. doi:10.1016/j.jtrangeo.2025.104217

Saleem, S. , JAiswal, A. and G R, B. (2025). Evaluating factors influencing Active Transportation in Developing Metropolises. *Civil Engineering Infrastructures Journal*, doi:10.22059/ceij.2025.383151.2161

Sarkar, P., & Mallikarjuna, C. (2013). Effect of Land Use on Travel Behaviour: A Case Study of Agartala City. *Procedia - Social and Behavioral Sciences, 104*, 533-542. doi:10.1016/j.sbspro.2013.11.147

Shahapure, K., & Nicholas, C. (2020). Cluster Quality Analysis Using Silhouette Score. *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)* (pp. 747-748). Sydney, NSW, Australia: IEEE. doi:10.1109/DSAA49011.2020.00096

Shi, C., Wei, B., Wei, S., Wang, W., Liu, H., & Liu, J. (2021). A Quantitative Discriminant Method of Elbow Point for the Optimal Number of Clusters in Clustering Algorithm. *Journal on Wireless Communications and Networking, 31*(2021), 1-16. doi:10.1186/s13638-021-01910-w

Tang, J., Wang, X., Zong, F., & Hu, Z. (2020). Uncovering Spatio-temporal Travel Patterns Using a Tensor-based Model from Metro Smart Card Data in Shenzhen, China. *Sustainability, 12*(4), 1475. doi:10.3390/su12041475

Tin Kam Ho. (1995). Random decision forests. *Proceedings of 3rd International Conference on Document Analysis and Recognition, 1*, pp. 278-282. Montreal, QC, Canada. doi:10.1109/ICDAR.1995.598994

Wang, X., Yao, L., Liu, W., Li, C., Bai , L., & Waller, S. (2020). Mobility Irregularity Detection with Smart Transit Card Data. *Pacific-Asia Conference on Knowledge Discovery and Data Mining. PAKDD 2020. Lecture Notes in Computer Science. 12084*, pp. 541–552. Springer, Cham. doi:10.1007/978-3-030-47426-3_42

Yu, C., & He, Ph.D, Z.-C. (2017). Analysing the spatial-temporal characteristics of bus travel demand using the heat map. *Journal of Transport Geography, 58*, 247-255. doi:10.1016/j.jtrangeo.2016.11.009

Zhang, S., Yang, Y., Zhen, F., Lobsang, T., & Li, Z. (2021). Understanding the travel behaviors and activity patterns of the vulnerable population using smart card data: An activity space-based approach. *Journal of Transport Geography*. doi:10.1016/j.jtrangeo.2020.102938

Zhang, Y., & Xu, D. (2022). The bus is arriving: Population growth and public transportation ridership in rural America. *Journal of Rural Studies, 95*, 467-474. doi:10.1016/j.jrurstud.2022.09.018

Zhao, J., Ruyue, L., Zhang, F., Xu, C.-Z., & Feng, S. (2014). Understanding temporal and spatial travel patterns of individual passengers by mining smart card data. *2014 IEEE 17th International Conference on Intelligent Transportation Systems (ITSC)* (pp. 2991-2997). Qingdao: IEEE. doi:10.1109/ITSC.2014.6958170

Zhao, X., Wu, Y.-p., Ren, G., Ji, K., & Qian, W.-w. (2019). Clustering Analysis of Ridership Patterns at Subway Stations: A Case in Nanjing, China. *Journal of Urban Planning and Development , 145*(2). doi:10.1061/(ASCE)UP.1943-5444.0000501

Zhao, X., Zhang, Y., Hu, Y., Qian, S., & yin, b. (2021). Interactive Visual Exploration of Human Mobility Correlation Based on Smart Card Data. *IEEE Transactions on Intelligent Transportation Systems, 22*(8), 4825 - 4837. doi:10.1109/TITS.2020.2983853

Zhou, Y., Fang, Z., Zhan, Q., Huang, Y., & Fu, X. (2017). Inferring Social Functions Available in the Metro Station Area from Passengers' Staying Activities in Smart Card Data. *ISPRS International Journal of Geo-Information, 6*(12), 394. doi:10.3390/ijgi6120394

Zhou, Y., Qian, C., Xiao, H., Xin, J., Wei, Z., & Feng, Q. (2019). Coupling Research on Land Use and Travel Behaviors Along the Tram Based on Accessibility Measurement -Taking Nanjing Chilin Tram Line1 as an Example. *Sustainability, 11*(7), 2034. doi:10.3390/su11072034

Zhou, Y., Thill, J.-C., Xu, Y., & Fang, Z. (2021). Variability in individual home-work activity patterns. *Journal of Transport Geography, 90*, 102901. doi:10.1016/j.jtrangeo.2020.102901

Zhu, T.-L., Wang, X., Zhang, J., Yu, S., & Molotov, I. (2022). Mean-shift clustering approach to the tracklets association with angular measurements of resident space objects. *Astronomy and Computing, 40*, 100588. doi:10.1016/j.ascom.2022.100588