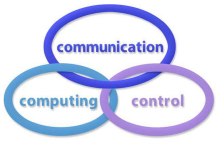
INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL
Online ISSN 1841-9844, ISSN-L 1841-9836, Volume: 19, Issue: 2, Month: April, Year: 2024
Article Number: 6422, https://doi.org/10.15837/ijccc.2024.2.6422



CCC Publications

Unsupervised Learning-Based Exploration of Urban Rail Transit Passenger Flow Characteristics and Travel Pattern Mining

M. C. Tang, J. Cao, D.Q. Gong, G. Xue

Mincong Tang*
1.Xuzhou University of Technology, China
2.Industrial University of Ho Chi Minh City, Vietnam
3.International Center for Informatics Research, Beijing Jiaotong University, China
*Corresponding author: mincong@bjtu.edu.cn

Jie Cao*
Xuzhou University of Technology, China
*Corresponding author: cj@nuist.edu.cn

Daqing Gong
International Center for Informatics Research, Beijing Jiaotong University, China
dqgong@bjtu.edu.cn

Gang Xue
School of Economics and Management, Tsinghua University, China
xuegang@sem.tsinghua.edu.cn

Abstract

This study delves into the realm of urban rail transit systems, leveraging unsupervised learning techniques to analyze passenger flow characteristics and unearth travel patterns. Focused on the dynamic and complex nature of urban rail networks, the research utilizes extensive datasets, primarily derived from Automated Fare Collection (AFC) systems, to provide a comprehensive analysis of passenger behaviors and movement trends. Employing advanced algorithms like DBSCAN, the study categorizes passengers into distinct groups, including tourists, shoppers, thieves, commuters, and station staff. These classifications reveal intricate patterns in travel behaviors, significantly contributing to a deeper understanding of urban transit dynamics. The findings offer valuable insights into peak travel times, popular routes, and station congestion, highlighting potential areas for operational improvements and infrastructure development. The study's application of unsupervised learning in analyzing vast, unstructured data sets a precedent in urban transportation research, showcasing the potential of artificial intelligence in enhancing the efficiency and sustainability of urban transit systems. The insights garnered are pivotal not only for optimizing current operations but also for shaping future expansion and adaptation strategies, ensuring urban rail systems continue to meet the evolving needs of growing urban populations.

Keywords: Urban Rail Transit, Unsupervised Learning, Travel Pattern Mining, DBSCAN.

1 Introduction

Urban rail transit, as the lifeline of modern urban life, plays a crucial role in supporting city development and facilitating daily commutes. With the rapid growth of the global economy and accelerated urbanization, urban rail transit worldwide has encountered unprecedented opportunities and challenges [1, 2, 3]. Particularly in China, the swift expansion of urban rail networks is a globally recognized phenomenon. As of the end of 2020, mainland China boasted 44 cities operating 233 urban rail transit lines, covering a total of 7545.5 kilometers, encompassing 4660 stations, and serving an astonishing annual passenger volume of 17.59 billion [3]. Behind these staggering figures lies not only the central role of urban rail transit in urban transportation systems but also numerous challenges and issues such as carriage overcrowding, insufficient transport capacity, and long waiting times for passengers [4, 5].

The emergence of these issues puts significant pressure and challenges on the operation and planning departments of urban rail transit systems. There is an urgent need for effective planning and management to enhance operational efficiency and improve passengers' commuting experience. In this context, the analysis of passenger flow characteristics becomes particularly critical. By conducting in-depth analysis of passengers' travel habits, commute durations, and other behavioral data, operators can better manage peak-time passenger flows, alleviate carriage congestion, and more rationally allocate resources during regular operations, thereby enhancing operational efficiency. Additionally, this analysis provides essential data support for the planning of new routes [3, 4, 5, 6].

On the technological front, the development of big data and artificial intelligence has made the analysis of urban rail transit passenger flows more feasible and efficient [7, 8, 9]. The data sources are extensive, including usage data from smart transit cards and video surveillance data from stations, all of which form the basis for analysis. By employing advanced data mining technologies and algorithms such as K-Means++, DBSCAN, etc., these massive datasets can be effectively processed to extract passengers' travel characteristics, thereby offering more precise operational decision-making support [9, 10, 11, 12].

Furthermore, the study of individual passenger flow characteristics is not only beneficial for the optimization of existing networks but also provides crucial insights for the planning and design of new routes. By analyzing the travel patterns of current passengers, planning departments can make more scientifically informed decisions about the direction of new lines, station layouts, and train compositions, thereby preventing potential future issues such as traffic congestion and inadequate station design.

Therefore, this study aims to conduct an in-depth analysis of passenger flow data from the Beijing subway system to categorize individual passenger flow characteristics, including both regular and irregular passengers. Through the analysis and visualization of this data, we aspire to aid operators in better understanding their passengers and in providing superior services. Simultaneously, this study will offer crucial data support for the future planning and optimization of urban rail transit systems. The integration of advanced technologies such as artificial intelligence in this research will contribute significantly to enhancing the operational efficiency and service levels of urban rail systems. It will provide robust support in addressing issues like carriage congestion and inadequate transport capacity, thereby elevating the commuting experience for citizens and promoting the sustainable development of urban rail transit systems.

In conclusion, urban rail transit systems are more than just transportation networks; they are integral to the sustainable growth of cities and the well-being of their inhabitants. The insights gained from this study will not only optimize current operations but also shape future developments, ensuring that urban rail transit remains a cornerstone of urban infrastructure, supporting the dynamic life of modern cities and the ever-evolving needs of their residents. The application of cutting-edge data analysis in understanding and improving urban rail systems marks a significant step towards smarter, more responsive urban environments, where technology and human-centered design converge to create seamless, efficient, and enjoyable transit experiences.

2 Related works

2.1 Passenger Travel Pattern Mining

Urban rail transit systems, as vital arteries of modern urban life, carry the immense daily travel demand of cities. Recent studies underscore the importance of leveraging advanced data mining techniques to understand and optimize the passenger flow characteristics in these complex systems. In this regard, the study by Jiang et al. [3] unveils periodic frequent travel patterns of metro passengers by considering different time granularities and station attributes, offering an in-depth understanding of passenger behavior. Similarly, Daneshvar et al. [4] analyze the behavioral patterns of bus passengers using data mining methods, providing valuable perspectives on passenger behaviors in different transit systems. Furthermore, Li et al. [5] engage in individualized passenger travel pattern multi-clustering based on graph regularized tensor latent Dirichlet allocation, demonstrating how to extract useful insights from big data to predict passenger behavior. Ye and Ma [6] focus on using transit smart card data for clustering-based travel pattern prediction of frequent passengers, a critical approach to understanding regular commuter flows. In the realm of real-time big data processing, Shi et al. [7] illustrate how to adaptively detect anomalous paths in floating vehicle trajectories, crucial for understanding and managing urban traffic flows. Finally, Kong et al. [8] explore human mobility for multi-pattern passenger prediction using a graph learning framework, again emphasizing the application of advanced analytical techniques in traffic system management.

Collectively, these studies reveal the complexity and diversity of passenger flows in urban rail transit systems. By analyzing extensive datasets, these works not only enhance our understanding of passenger behavior patterns but also provide valuable insights for the planning and operation of rail transit systems. These insights are particularly significant for managing peak times and crowded routes, planning new lines, and handling emergency situations. In summary, these studies showcase the potential of big data and machine learning techniques in the management of modern urban transit, pointing the way for future research and practice. This body of work underscores a paradigm shift in urban transit analysis: from traditional methods reliant on manual surveys and limited data, to dynamic, AI-driven approaches that harness the power of large-scale, diverse datasets. This transition is critical not only for optimizing current operations but also for shaping future transit infrastructure, ensuring it meets the evolving demands of urban populations. The integration of AI and data science in urban transit represents a significant leap towards smarter, more efficient urban mobility solutions.

2.2 Unsupervised Learning Based Pattern Mining

Unsupervised learning, a subset of machine learning techniques that operates without labeled outcomes, is increasingly applied across various domains to extract meaningful patterns from data [9, 10]. This section synthesizes insights from recent studies [11, 12, 13, 14, 15] to demonstrate the versatility and impact of unsupervised learning methodologies. Li et al. [11] delve into maritime traffic pattern extraction using an unsupervised hierarchical methodology. This approach in maritime logistics underscores the potential of unsupervised learning in enhancing the understanding of complex, large-scale traffic movements, pivotal for optimizing maritime operations and safety. Lefoane et al. [12] explore unsupervised learning for feature selection in botnet detection within 5G networks. Their work illustrates the crucial role of unsupervised learning in cybersecurity, particularly in the context of emerging 5G technology, where traditional security measures may fall short against sophisticated cyber threats. In the field of computer vision, Zhang et al. [13] present a novel approach to unsupervised 3D action representation learning through contrastive positive mining. This study highlights the adaptability of unsupervised learning in extracting rich, nuanced features from complex visual data, essential for advancing automated recognition systems. Jindal & Singh [14] focus on detecting malicious transactions in databases using a hybrid metaheuristic clustering and frequent sequential pattern mining approach. This research showcases the application of unsupervised learning in database security, emphasizing its effectiveness in identifying subtle, anomalous patterns that may indicate security breaches. Lastly, Vignesh et al. [15] propose a framework for analyzing crime datasets using an unsupervised optimized K-means clustering technique. Their study demonstrates the application

of unsupervised learning in social science, specifically in criminology, to decipher patterns and trends that can inform law enforcement strategies.

Collectively, these studies reveal the expansive utility of unsupervised learning across diverse fields. Whether it's enhancing maritime traffic safety, bolstering cybersecurity in next-generation networks, advancing visual recognition systems, securing databases, or aiding crime analysis, unsupervised learning proves to be a powerful tool [16, 17, 18]. This versatility not only paves the way for innovative applications but also poses significant implications for future technological advancements. By enabling the extraction of deep insights from vast, unstructured datasets, unsupervised learning methodologies are reshaping the way we analyze data and solve complex problems in various domains [19, 20, 21].

3 Data

This article focuses on the Beijing Subway as its research subject. Beijing is the first city in China to develop urban rail transit, with its first line operational since 1971. As of March 2022, the Beijing Subway has 27 lines, covering a total of 783 kilometers with 459 stations, including 72 transfer stations. It is the second-largest urban rail transit system in the world, second only to Shanghai. Despite its vast size, the Beijing Subway continues to expand. Currently, there are 11 subway lines under construction, totaling 235.6 kilometers. By 2025, the Beijing Subway is expected to form a network of 30 operating lines, spanning a total length of 1177 kilometers. Figure 3-1 shows the changes in passenger traffic and operational mileage of the Beijing Subway from 2000 to 2019. It is evident that in the past nearly 20 years, the annual total passenger traffic of the Beijing Subway has increased nearly sevenfold, and the operational mileage has increased nearly sixfold. Simultaneously, with the rapid expansion of the network, the management challenges it brings are also immense, which is rare for cities around the world.

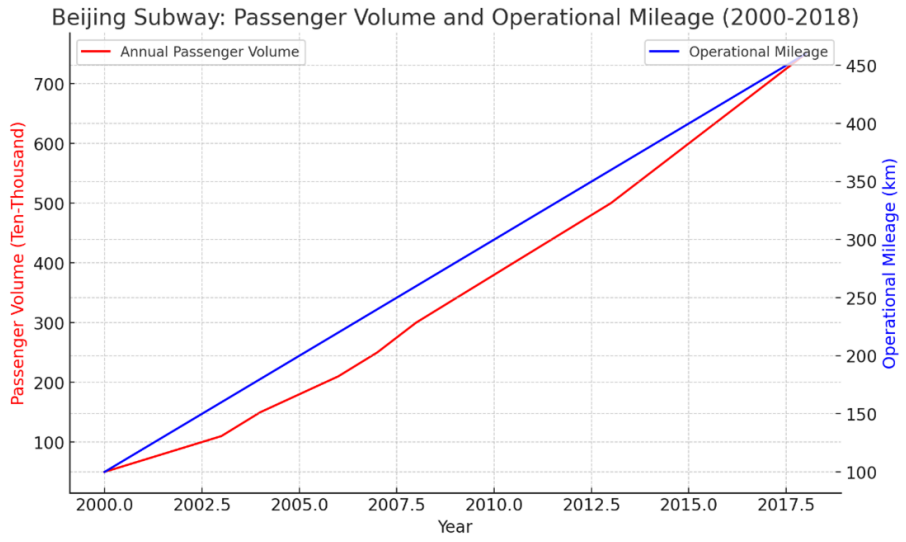


Figure 1: Changes in passenger flow and operating mileage of Beijing subway over the years (2000-2019)

In preparation for the experimental part of this study, it is necessary to preprocess various types of data involved to meet the requirements of the experiments. This chapter will discuss the data involved in this experiment and how to process and merge them to reach a state ready for experimentation. This study involves multiple types of data, including: Beijing Subway AFC (Automated Fare Collection) system data, geographic coordinates of Beijing subway stations, and information on Beijing subway ticket types. The AFC system, an integral part of urban subway systems, is encountered by passengers as the ticket gates that process payments upon entry and exit. This system records crucial information such as the time and specific stations of passenger entry and exit, the type of ticket used, and the amount spent. Therefore, for the purpose of this study, Beijing Subway's AFC data is used as the primary research object.

Table 1: Forms of collated data and their corresponding relevance

Exit Time	Arrival Station Code	Card Number	Origin Station Code	Entry Time	Card Type	Card Subtype
TXN_DATE _TIME	DEVICE_ LOCATION	CARD_SERIAL	TRIP_ORIGIN _ LOCATION	ENTRY_TIME	CARD_ISSUER _ID	PRODUCT_ TYPE

In May 2006, the Beijing Transportation Company began issuing smart cards compatible with both the Beijing bus and subway systems. There are two types of fare systems in the Beijing AFC: a flat fare and a distance-based fare. However, due to design flaws in the smart card scanning system, the AFC system on buses with a flat fare does not save any boarding location information. Although it stores the boarding and alighting locations for distance-based fare buses, it does not record the times. This presents additional challenges in data processing.

As shown in Figure 2, the data style used in this research project is sampled from the OD (Origin-Destination) data in the Beijing Subway AFC system from October 27th to October 31st, 2017. "O" stands for Origin, the starting point, and "D" for Destination, the end point, thus OD data refers to records of passenger entry and exit at stations. The study focuses on analyzing characteristics of passenger riding habits, using entry and exit times for temporal feature clustering analysis, station codes for spatial feature analysis, and a combination of both for comprehensive analysis of individual passenger flow characteristics and anomaly detection. For example, the same station code for entry and exit indicates a same-station entry and exit event. Since this study does not involve data on card balances, gate equipment, or subway operators, unnecessary data will be deleted using Python, retaining only the first nine columns, and the data will be organized as shown in Table 1 for research purposes.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	TXN_DATE_TIME	DEVICE_LOCATION	CARD_SERIAL	TRIP_ORIGIN_	ENTRY_TIME	PRODUCT_I	PRODUCT_T	PAYMENT_T	CARD_LIFFE	RECONCILIAT	SETTLEMEN	SAM_ID	DEVICE_ID	SOURCE_PA	PURSE_REM	PTSN
2	2017/5/9 5:11	150995208	12548460	150995208	2017/5/9 4:46	99	7	1	0	2017/5/9	2017/5/9	10000723	520161284	41	170	0
3	2017/5/9 5:07	150995461	12554364	150995461	2017/5/8 22:20	99	7	1	0	2017/5/9	2017/5/9	20000683	520226058	42	134	0
4	2017/5/9 5:08	150995463	12517707	150995463	2017/5/9 5:08	99	7	1	0	2017/5/9	2017/5/9	20000752	520226573	42	183	0
5	2017/5/9 4:55	150995463	12517707	150995463	2017/5/9 4:55	99	7	1	0	2017/5/9	2017/5/9	20000759	520226578	42	184	0
6	2017/5/9 5:01	150998315	12554552	150998315	2017/5/8 17:48	99	7	0	0	2017/5/9	2017/5/9	30000269	520956723	53	173	27
7	2017/5/9 5:06	150998315	12545166	150998315	2017/5/8 21:39	99	7	0	0	2017/5/9	2017/5/9	30000267	520956679	53	181	19
8	2017/5/9 4:57	150995463	12517677	150995463	2017/5/8 22:52	99	7	1	0	2017/5/9	2017/5/9	20000752	520226573	42	167	0
9	2017/5/9 4:53	150995462	12540710	150995462	2017/5/8 22:38	99	7	1	0	2017/5/9	2017/5/9	20000731	520226324	42	147	0
10	2017/5/9 5:02	150995462	12513306	150995462	2017/5/8 21:24	99	7	1	0	2017/5/9	2017/5/9	20000703	520226315	42	149	0

Figure 2: Data Sample

4 Method

4.1 DBSCAN

DBSCAN, short for Density-Based Spatial Clustering of Applications with Noise, is a data clustering algorithm that is considered to be one of the most significant advancements in the field of data mining. Proposed by Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu in 1996, it represents a paradigm shift from traditional centroid-based clustering algorithms such as k-means. The distinctive feature of DBSCAN is its capacity to identify clusters of arbitrary shape and size in a data set, which is particularly effective in the presence of noise and outliers.

The algorithm operates on two main parameters: epsilon (eps) and minimum points (minPts). Epsilon is a spatial distance parameter that determines the neighborhood radius around each data point, while minPts specifies the minimum number of points required to form a dense region. The classification of points into core, border, and noise points is pivotal to the clustering process. A core point is one that has at least minPts within its epsilon neighborhood. A border point is in the neighborhood of a core point but has fewer than minPts within its epsilon range. Noise points are those that are neither core nor border points.

The core of the DBSCAN algorithm is the concept of density reachability and density connectivity. A point A is density reachable from point B if there is a chain of points P1, P2, ..., Pn where each Pi is within epsilon distance from Pi+1 and there are at least minPts within the epsilon neighborhood

of each Pi. Density connectedness extends this concept by linking points that are density reachable from a common core point.

DBSCAN begins with an unvisited point and determines if it is a core point. If it is, the algorithm proceeds to recursively find all points density-reachable from this point, which forms a cluster. This process continues until all points are either assigned to a cluster or marked as noise. The ability to incrementally build up clusters gives DBSCAN its robustness against noise and its capability to discover clusters of complex shapes. Mathematically, the DBSCAN algorithm can be described using a set of points in a metric space and a density estimation technique. The algorithm iterates through the dataset, computing the epsilon neighborhood of every point and counting the number of points within this neighborhood. If the count exceeds minPts, a new cluster is initiated. Otherwise, the point is labeled as noise, although it may later be found to be part of a cluster as the algorithm processes other points.

From an implementation perspective, DBSCAN can be efficiently realized using data structures such as R-trees or k-d trees that support efficient range querying, which is necessary to find the neighbors of a point within the epsilon radius. The general steps involve marking all points as unvisited, randomly selecting points to grow the clusters by adding all density-reachable points to the cluster, and iterating until all points have been processed.

In terms of computational complexity, the algorithm is generally O(n*log(n)) when a spatial index is used, which is particularly efficient for large datasets. However, the performance and the outcome of the clustering process are highly sensitive to the settings of eps and minPts. This sensitivity necessitates a careful selection of parameters, which may require domain knowledge or adaptive methods.

In practical terms, DBSCAN has been applied across various disciplines and industries, from geospatial analysis to market segmentation. Its ability to handle outliers and identify clusters of varying densities and shapes without the need for specifying the number of clusters makes it a versatile tool in the arsenal of data analysts. Despite its age, DBSCAN remains a highly relevant and widely employed clustering algorithm, owing to its simplicity, efficacy, and the intuition it offers in understanding the underlying structure of complex datasets.

4.2 Key mathematical concepts and formulas involved

Epsilon Neighborhood: For any point p in the dataset D, the epsilon neighborhood is defined as the set of points within a distance ϵ from p:

$$N\epsilon(p) = \{q \in D \mid dist(p,q) \leq \epsilon\} \text{.....(1)}$$

where $dist(p,q)$ is a distance measure (such as the Euclidean distance) between points p and q.

Core Point: A point p is a core point if its epsilon neighborhood contains at least a minimum number of points minPts:

$$\mid N\epsilon(p) \mid \geq \min Pts \text{.....(2)}$$

Directly Density-Reachable: A point p is directly density-reachable from point q if p is within the epsilon neighborhood of q and q is a core point:

$$p \in N\epsilon(q) \wedge \mid N\epsilon(q) \mid \geq \min Pts \text{.....(3)}$$

Density-Reachable: A point p is density-reachable from point q if there is a chain of points p_1 ,p_2 ,...,p_n such that $p_1 = q$, $p_n = p$, and $p_i + 1$ is directly density-reachable from p_i .

Density-Connected: A point p is density-connected to point q if there exists a point o such that both p and q are density-reachable from o.

These definitions set the stage for the clustering process, which is essentially the application of these concepts to the dataset to discover clusters. The mathematical aspects of DBSCAN are embedded in its core operations, which identify points satisfying these conditions and group them into clusters accordingly.

The algorithm can be summarized by the following pseudo-mathematical process:

Initialize all points as unclassified.

Table 2: Silhouette Coefficient Results for Different Eps and MinPts Values

MinPts	Eps =2	Eps =3	Eps =4	Eps =5
5	0.039	0.068	0.611	0.863
6	0.062	0.067	0.607	0.895
7	0.062	0.069	0.605	0.9
8	0.067	0.067	0.605	0.9
9	0.047	0.068	0.605	0.931
10	0.046	0.068	0.605	0.91

For each unclassified point p:

If $\mid N\epsilon(p) \mid \geq \min Pts$, classify p as a core point and start a new cluster C.

Expand C by recursively adding all points that are density-reachable from p to C.

Points not belonging to any cluster are classified as noise.

The algorithm's effectiveness depends on the distance function used, which in the standard implementation is the Euclidean distance for numerical data:

$$dist(p,q) = \sqrt{(\sum_i^n p_i - q_i^2} .....(4)$$

where p_i and q_i are the $i-th$ coordinates of points p and q, respectively.

DBSCAN's implementation usually involves data structures that can efficiently support the neighborhood queries, like k-d trees for low-dimensional data or metric trees for high-dimensional data. This efficiency is critical for maintaining the algorithm's computational complexity at $O(n \log n)$.

We employed the silhouette coefficient as a measure to determine the effectiveness of clustering at various Eps and MinPts values. This coefficient helps in assessing how close each point in one cluster is to points in the neighboring clusters, thus providing a clear indication of the clustering performance.

Parameter Selection: Based on our tests, we found that an Eps of 5 and MinPts of 9 yielded the highest silhouette coefficient of 0.931, suggesting highly effective clustering. This combination was hence selected for the main clustering experiment.

Balancing Noise Points: While adjusting these parameters, we also considered the impact on noise points. A higher Eps generally reduced the number of noise points, but a trade-off had to be made to ensure accurate clustering and not merely minimizing noise.

We acknowledge that setting hyperparameters can be challenging due to their sensitivity and the variability of datasets. Our approach was therefore iterative, testing various combinations to find the most effective settings for our specific dataset.

5 Experimental results

5.1 Determining the parameters Eps and MinPts

To determine the optimal values for the Eps radius and MinPts within the range of 5 to 10, this experiment utilized the DBSCAN function and the Silhouette_score function from the "sklearn.cluster" module in Python to calculate silhouette coefficients. The specific experimental data is displayed in the following table:

It is evident that when Eps is set to 5 and MinPts to 9, the silhouette coefficient reaches 0.931, which is remarkably close to 1 and the highest among all the data in the experiment. Therefore, this parameter will be used for the clustering experiment conducted from October 27 to October 31. The table also reveals that with a fixed MinPts value, a larger Eps radius results in better clustering outcomes; and as Eps increases, changing MinPts does not significantly impact the effectiveness of clustering.

Another factor considered in this experiment is the number of noise points under different parameters, where noise points are those not assigned to any cluster. More noise points within the clustering results suggest poorer performance, indicating a larger number of data points not categorized. In urban rail transit data, there are instances where passengers with atypical yet non-anomalous commuting patterns are categorized as noise points. To minimize this error, clustering parameters resulting in

fewer noise points are naturally preferred. Tables 3 through 8 showcase the variations in the number of noise points caused by different Eps values for the same MinPts value.

Table 3: Number of Noise Points for Different Eps Values (MinPts=5)

Eps(MinPts=5)	2	3	4	5
Noise	327	226	163	124

Table 4: Number of Noise Points for Different Eps Values (MinPts=6)

Eps(MinPts=6)	2	3	4	5
Noise	404	247	196	146

Table 5: Number of Noise Points for Different Eps Values (MinPts=7)

Eps(MinPts=7)	2	3	4	5
Noise	448	289	207	172

Table 6: Number of Noise Points for Different Eps Values (MinPts=8)

Eps(MinPts=8)	2	3	4	5
Noise	494	318	243	200

Table 7: Number of Noise Points for Different Eps Values (MinPts=9)

Eps(MinPts=9)	2	3	4	5
Noise	533	358	273	223

Table 8: Number of Noise Points for Different Eps Values (MinPts=10)

Eps(MinPts=10)	2	3	4	5
Noise	580	391	298	229

From the above tables, it can be observed that a larger Eps gradually reduces the number of noise points, and for the same Eps radius, a larger MinPts results in more noise points. By this reasoning, one might consider adopting the parameters Eps=5 and MinPts=5. However, given that the silhouette coefficient more accurately reflects the quality of clustering results, and considering that the few hundred noise points are negligible compared to the tens of millions of data points in urban rail transit data, this study will continue to use the parameters Eps=5 and MinPts=9 for further analysis.

5.2 Clustering results

Using the parameters of Eps=5 and MinPts=9, as identified in the previous section, the DBSCAN algorithm was applied to cluster analysis on the Beijing Subway AFC data from October 27 to October 31, 2017. The clustering analysis was based on travel time and the straight-line distance between origin and destination stations. In addition to calculating the silhouette coefficient, the number of clusters and noise points were also determined.

As shown in figure 3, apart from October 30, where the silhouette coefficient was relatively low at 0.637, the silhouette coefficients for the other four days exceeded 0.9, approaching the ideal value of 1, indicating highly effective clustering results. The average silhouette coefficient over the five days was 0.865, suggesting the clustering was particularly robust. Furthermore, the distribution of noise points during the five-day experiment ranged from 210 to 260, with an average of 238.6, indicating a stable occurrence of noise points relative to the millions of passenger flow data points.

Most passengers were grouped into a single cluster, displaying a clear upward-right trend. This pattern indicates that the time and space characteristics of these passengers are directly proportional,

that is, the longer the travel distance, the longer the travel time, which is consistent with typical commuting patterns. Hence, these passengers can be classified as exhibiting no anomalous behavior.

However, a subset of passengers exhibited travel times ranging from 250 to 600 minutes with travel distances close to zero, suggesting potential anomalies. Two plausible explanations were considered: firstly, subway company employees might need to spend extended periods within the fare area for work purposes; secondly, there might be passengers exhibiting abnormal behavior, such as loitering for extended periods, which could be indicative of theft, begging, or other activities. The noise points, not classified within the normal range of passenger flow, could also represent anomalies. These seemingly anomalous data points will be further investigated in the subsequent chapter on user profile construction

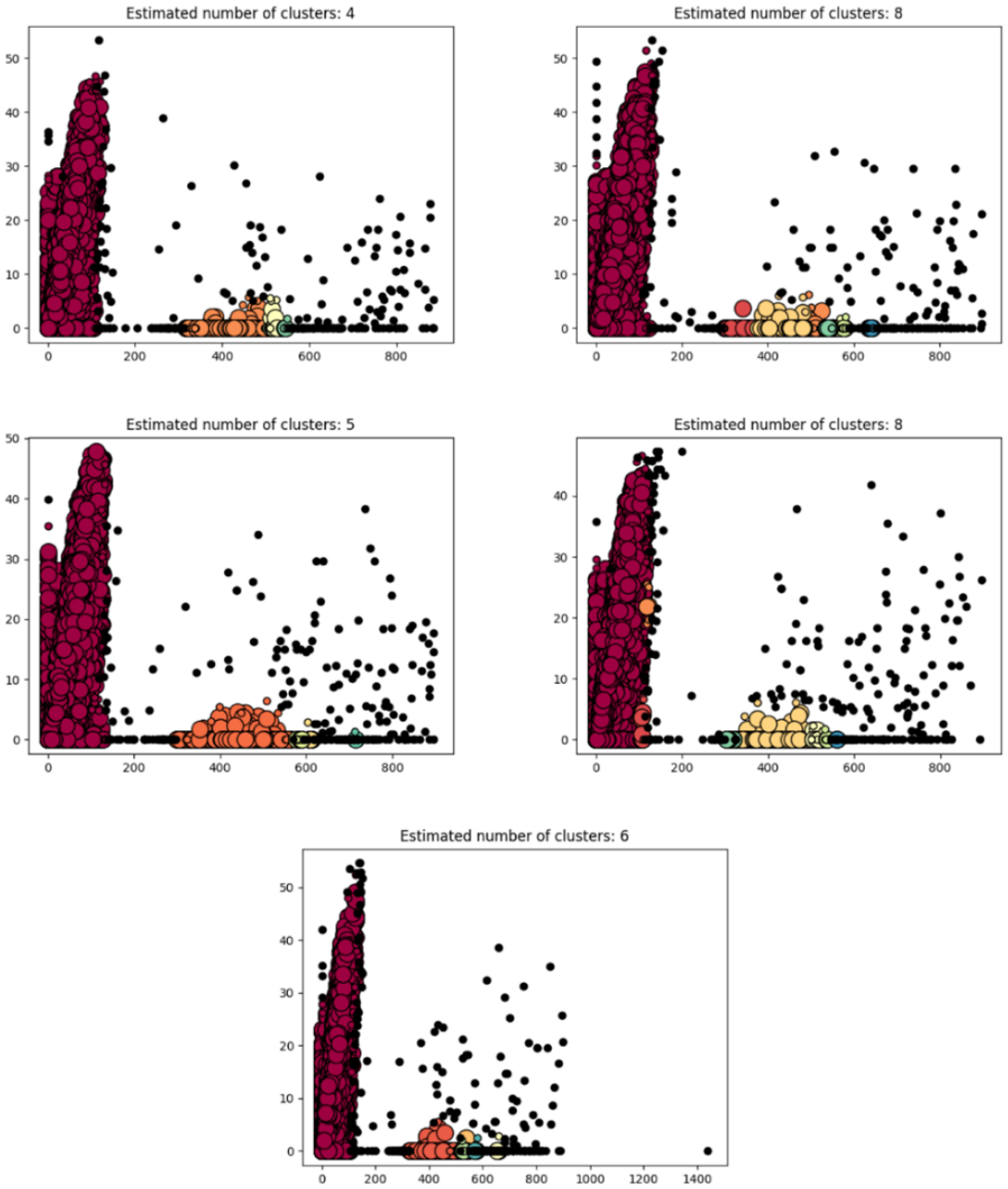


Figure 3: DBSCAN results

5.3 Individual User Profiles

The concept of user profiling was first introduced by Alan Cooper. Current research on user profiles primarily follows three directions:

User Attributes: The main goal here is to understand users by collecting characteristic information through a Social Annotation System. The core task of user profiling is data labeling, involving cleaning and organizing various raw data, refining user attributes, and eventually extracting user tags from these attributes.

User Preferences: Aimed at enhancing the quality of personalized recommendations by measuring users' interest levels.

User Behavior: Focused on predicting user behavior trends to prevent customer loss and design appropriate measures. User attributes can be categorized into three types: basic attributes, domain attributes, and specific attributes. Basic attributes are natural attributes of users, such as age, gender, and other demographic properties. Domain attributes apply knowledge from a specific professional field to analyze theoretical aspects of the researched problem, deriving user attributes needed for the study, including behavioral attributes and interest attributes. Specific attributes are identified based on specific research needs, often representing the innovative aspects of various studies.

This paper first needs to complete the collection of user data. Next, mining and filtering of user data take place, which was also implemented in the data processing stage of Chapter 2. The final step involves the extraction and recombination of user profile tags. User profiling reflects the data of each real individual in the virtual world. For a massive dataset like urban rail transit passenger flow data, it's obviously not feasible to study and create user profiles for every individual passenger flow. Passenger flow can be divided into three types: visitors, shoppers, and thieves. Visitors and thieves exhibit similar behaviors in data, as they both visit many different places without a clear pattern. However, visitors generally head to tourist attractions or commercially prosperous areas with longer intervals between destinations. In contrast, thieves tend to wander randomly, frequently disembarking and with unclear destinations. They are prone to visiting multiple tourist or commercial areas in a short period, unlike most regular passengers who only visit one area in a short time.

Based on these classification criteria, three types of passengers can be identified. Additionally, this study will incorporate two more categories: commuters and station staff. The classification criteria for commuters are relatively straightforward: typically, they have commuting records during morning and evening peak hours on workdays, traveling between two fixed points, often represented on the map as two points and a line, the "two points, one line" phenomenon. Station staff, classified by their card number and entry-exit times mostly in the early morning non-operational hours, can be further divided into station service personnel and technical maintenance personnel. The difference lies in that service personnel only enter and exit at the same station (represented on the map as one point), whereas maintenance personnel's records appear at multiple stations, all during pre-operation hours in the early morning. To summarize, this study will select six types of passengers for profiling: visitors, shoppers, thieves, commuters, station service personnel, and technical maintenance personnel, building their user profiles using data and OD (origin-destination) record connection maps. Based on the six passenger types mentioned above, data representing each type will be filtered, and OD record connection maps will be drawn using algorithms. Additionally, due to the potentially pejorative nature of some categorizations, it is declared that the results are speculative, and the following qualitative results will be described using the term "suspected."

User Profile - Tourist Type

Case Study 1: Suspected tourist-type passenger. Characterized by long intervals between exiting and entering the metro, indicating infrequent use, in line with the pattern of tourists visiting attractions. Destination analysis based on latitude and longitude coordinates of their origin stations reveals visits to tourist attractions. The OD data connection map shows faint lines, suggesting rare travel along each path, aligning with the typical one-time visit pattern of tourists. Hence, the user of this card number is suspected to be a tourist-type passenger.

Case Study 2: Another suspected tourist-type passenger. The time intervals between station entries and exits are typically one to two hours, aligning with the typical tourist behavior of staying at attractions. Destination analysis indicates visits to tourist sites, and the OD data connection map's

mostly pale, gray lines suggest non-repetitive travel routes, supporting the tourist visitation pattern. Therefore, this card number is also suspected to belong to a tourist-type passenger.

Case Study 3: Similar to the previous cases, this passenger exhibits characteristics of long intervals between station use, visits to tourist sites without repetition, and an OD data connection map showing travel from one point to various areas, with non-repetitive, faint path colors. Thus, this user is also suspected to be a tourist-type passenger.

Shopper Type

Suspected shopper-type passenger. This user's activities on Friday and weekends (October 27 was a Friday, 28th and 29th were Saturday and Sunday) are primarily around major shopping areas in Beijing. The passenger might have visited shopping centers like Huamao Skp near Dawanglu Station, Sanlitun Taikooli near Tuanjiehu, and the bustling Guomao CBD. The long intervals between station entries and exits suggest the purposeful visitation of these areas. The activities are relatively concentrated, with clear destinations, mostly radiating from a fixed point (presumed residence) to specific locations.

User Profile - Thief Type

Case Study 1: Suspected thief-type passenger. Characterized by multiple metro rides over five days without clear destinations, frequent station entries and exits, mostly at crowded tourist spots and transport hubs. The aimless wandering, frequent disembarkation, and unclear destinations align with thief behavior. The connection map indicates multiple travels along certain paths, with some white sections suggesting frequent use of these routes during the five days. The focus on tourist spots increases the likelihood of a thief identity. Therefore, this user is marked as a "suspected thief."

Case Study 2: Another suspected thief-type passenger. This user frequently used the metro network, totaling 28 entries and exits in five days. The chaotic travel pattern in the OD data connection map, lacking specific destinations, aligns with thief characteristics of random wandering and unclear objectives.

User Profile - Commuter Type

Case Study 1: Suspected commuter-type passenger. Typically travels during peak commuting hours on weekdays, with fixed destination patterns evident in the data. Therefore, it is likely that this user is a regular commuter.

Case Study 2: Card number "15979436" is suspected to be a commuter-type passenger. The data and map show "two points, one line" style travel on October 27, 30, and 31, commuting between two coordinates during peak hours, and a visit to a commercial area station on October 28. Thus, this user is also suspected to be a commuter-type passenger.

User Profile - Station Service Personnel

Case Study 1: Suspected station service personnel. Characterized by multiple entries and exits at the same station within a day, mostly in the early morning just before metro operation starts. Travel times are either exceptionally long or immediate exits. Given the card type is "1-2" employee card, the user is likely a station service staff member. The OD data connection map for station service shows only one point, offering little additional insight, so only data is presented.

User Profile - Technical Maintenance Personnel

Case Study 1: Suspected technical maintenance personnel. Entries and exits mostly occur between 5 and 7 AM, before metro operations begin, with a "99-7" employee card type. The OD data connection map shows chaotic destinations with faint colors for paths taken during this short period, suggesting likely maintenance work to ensure safe metro operations before the start of the day.

The above user profiles are developed for six passenger types: tourists, shoppers, thieves, commuters, station service personnel, and technical maintenance personnel, constructed using data and OD record connection maps. The qualitative nature of these profiles is speculative, using the term "suspected" for identification.

6 Discussion

6.1 Theoretical Implications

The comprehensive analysis of individual passenger flow characteristics in urban rail transit systems, particularly in the context of big data, offers several theoretical implications. First, this study underscores the significance of integrating big data analytics into public transportation research. The use of advanced data mining techniques, such as DBSCAN, in analyzing complex and large-scale datasets, demonstrates the potential of big data in uncovering hidden patterns and trends in passenger behavior. This approach challenges traditional methods that often rely on manual surveys or limited sample sizes, offering a more nuanced and detailed understanding of passenger flows.

Secondly, the study contributes to the field of transportation planning and management by providing a framework for categorizing passengers into distinct profiles based on their travel patterns. This classification not only enriches the existing literature on passenger behavior analysis but also introduces a novel perspective in understanding the dynamics of urban rail transit systems. By identifying specific groups such as tourists, shoppers, and commuters, the study offers a more granular view of passenger needs and preferences, which is critical for effective transit planning.

Additionally, the research highlights the importance of context in data analysis. The differentiation between various passenger types, such as distinguishing between tourists and thieves who exhibit similar travel patterns, emphasizes the need for contextual understanding in interpreting data. This insight is valuable for developing more sophisticated models and algorithms that can accurately interpret and predict passenger behavior in different scenarios.

Lastly, this study contributes to the broader discourse on the role of technology in urban development. By demonstrating how big data can be harnessed to enhance the efficiency and effectiveness of urban rail transit systems, the research aligns with the growing emphasis on smart city initiatives. It underscores the potential of technology in fostering sustainable urban growth and improving the quality of urban life.

6.2 Practical Implications

From a practical standpoint, this study offers several key takeaways for urban rail transit operators and city planners. Firstly, the detailed analysis of passenger flow characteristics can inform more effective management strategies. For instance, identifying peak travel times and high-demand routes enables operators to optimize train schedules and frequency, reducing wait times and alleviating carriage overcrowding. This approach not only enhances passenger satisfaction but also improves the overall efficiency of the transit system.

Moreover, the categorization of passengers into distinct profiles provides valuable insights for targeted service improvements. Understanding the specific needs and preferences of different groups, such as tourists versus commuters, allows for the customization of services. For example, tourist-heavy routes could benefit from enhanced navigation aids and information services, while commuter routes might prioritize speed and frequency.

The study also has implications for future transit network planning. The analysis of passenger flow patterns can guide the development of new lines and stations, ensuring that they effectively meet the needs of the city's population. Data-driven insights can help predict future demand, prevent overcapacity issues, and ensure that new infrastructure investments are both efficient and cost-effective.

Additionally, the findings of this study can be instrumental in crisis management and emergency planning. Understanding passenger flow dynamics is crucial in scenarios such as evacuations, service disruptions, or implementing health and safety measures, as seen during the COVID-19 pandemic.

In conclusion, the application of big data analytics in analyzing urban rail transit passenger flows offers both theoretical and practical benefits. It not only advances academic understanding in the field of urban transportation but also provides actionable insights for improving transit services, planning future developments, and ensuring the sustainable growth of urban rail systems.

7 Conclusions

This study, centered on the analysis of individual passenger flow characteristics in urban rail transit using big data, marks a significant stride in understanding and optimizing urban transportation systems. By leveraging advanced data analytics, particularly the DBSCAN algorithm, the research successfully categorized passengers into distinct profiles: tourists, shoppers, thieves, commuters, station service personnel, and technical maintenance personnel. This categorization illuminated the diverse needs and behaviors within the urban rail network, providing a nuanced understanding of passenger dynamics. The findings revealed key patterns in travel behaviors, such as the unique movement patterns of tourists compared to regular commuters or the distinctive travel signatures of potential thieves. This deeper insight into passenger behavior is instrumental for enhancing operational efficiency, improving passenger experience, and guiding the strategic planning of urban rail transit systems. The use of big data analytics in this context demonstrates its value in extracting meaningful insights from large and complex datasets, proving essential for contemporary urban transit management and planning.

However, the study is not without its limitations. One of the primary constraints lies in the reliance on historical data, which may not fully capture the rapidly changing dynamics of urban rail transit systems, especially in the face of emerging challenges like pandemics or significant urban developments. Additionally, the methodological approach, while robust, might oversimplify the complexity of human behavior by categorizing passengers into distinct groups, potentially overlooking the fluidity and intersectionality of passenger characteristics and travel purposes. Another limitation is the geographic focus on the Beijing subway system, which may limit the generalizability of the findings to other urban contexts with different cultural, economic, and infrastructural backgrounds.

Looking ahead, future research should aim to address these limitations by incorporating real-time data analysis to capture the evolving nature of urban rail systems. Expanding the study to include diverse urban contexts and transit systems globally could enhance the generalizability and applicability of the findings. Moreover, integrating more sophisticated machine learning and artificial intelligence techniques could provide a more dynamic and nuanced understanding of passenger behavior, accommodating its inherent complexity and variability. Future studies could also explore the integration of predictive analytics to not only analyze but also forecast passenger flow trends, thereby proactively informing operational and strategic decisions. Ultimately, the goal should be to continue harnessing the power of big data and advanced analytics to drive the evolution of urban rail transit systems, making them more efficient, responsive, and attuned to the needs of their diverse user base.

Funding

This study is partially supported by the project: "Jiangsu Provincial Talent Work Leadership Group: Study on Public Safety Risks in Smart Cities Based on Big Data (BRA2017396)"

Author contributions

The authors contributed equally to this work.

Conflict of interest

The authors declare no conflict of interest.

References

[1] Youn, J., Kim, T., Lee, J.K.D. (2022). Impacts of changes in traffic conditions on preference for public apartments, *Journal of System and Management Sciences*, 12(2), 378-390.https://doi.org/10.33168/JSMS.2022.0220

[2] Dakak, S., Wahbeh, F. (2020). Designing fast transportation network in Damascus: an approach using flow capturing location allocation model, *Journal of Logistics, Informatics and Service Science*, 7(1), 58-66. https://doi.org/10.33168/LISS.2020.0105

[3] Jiang, Z., Tang, Y., Gu, J., Zhang, Z., & Liu, W. (2023). Discovering periodic frequent travel patterns of individual metro passengers considering different time granularities and station attributes, *International Journal of Transportation Science and Technology*, https://doi.org/10.1016/j.ijtst.2023.03.003

[4] [4] Daneshvar, A., Salahi, F., Ebrahimi, M., & Nahavandi, B. (2023). Analyzing behavioral patterns of bus passengers using data mining methods (case study: rapid transportation systems), *Journal of applied research on industrial engineering*, 10(1), 11-24.

[5] Li, Z., Yan, H., Zhang, C., & Tsung, F. (2022). Individualized passenger travel pattern multi-clustering based on graph regularized tensor latent dirichlet allocation, *Data Mining and Knowledge Discovery*, 36(4), 1247-1278. https://doi.org/10.1007/s10618-022-00842-3

[6] Ye, P., & Ma, Y. (2023). Clustering-Based Travel Pattern for Individual Travel Prediction of Frequent Passengers by Using Transit Smart Card, *Transportation Research Record*, 2677(2), 1278-1287. https://doi.org/10.1177/03611981221111355

[7] Shi, Y., Wang, D., Ni, Z., Liu, H., Liu, B., & Deng, M. (2022). A Sequential Pattern Mining Based Approach to Adaptively Detect Anomalous Paths in Floating Vehicle Trajectories, *IEEE Transactions on Intelligent Transportation Systems*, 23(10), 18186-18199.

[8] Kong, X., Wang, K., Hou, M., Xia, F., Karmakar, G., & Li, J. (2022). Exploring human mobility for multi-pattern passenger prediction: A graph learning framework, *IEEE Transactions on Intelligent Transportation Systems*,23(9), 16148-16160. https://doi.org/10.1109/TITS.2022.3165066

[9] Zhang, L., Ma, J., Liu, X., Zhang, M., Duan, X. & Wang, Z. (2022). A Novel Support Vector Machine Model of Traffic State Identification of Urban Expressway Integrating Parallel Genetic and C-Means Clustering Algorithm, *Tehnički vjesnik*, 29 (3), 731-741. https://doi.org/10.17559/TV-20211201014622

[10] Balan, N. & Ila, V. (2022). A Novel Biometric Key Security System with Clustering and Convolutional Neural Network for WSN, *Tehnički vjesnik*, 29 (5), 1483-1490. https://doi.org/10.17559/TV-20211109073558

[11] Li, H., Lam, J. S. L., Yang, Z., Liu, J., Liu, R. W., Liang, M., & Li, Y. (2022). Unsupervised hierarchical methodology of maritime traffic pattern extraction for knowledge discovery, *Transportation Research Part C: Emerging Technologies*, 143, 103856. https://doi.org/10.1016/j.trc.2022.103856

[12] Lefoane, M., Ghafir, I., Kabir, S., & Awan, I. U. (2022). Unsupervised learning for feature selection: A proposed solution for botnet detection in 5g networks, *IEEE Transactions on Industrial Informatics*, 19(1), 921-929. https://doi.org/10.1109/tii.2022.3192044

[13] Zhang, H., Hou, Y., Zhang, W., & Li, W. (2022, October). Contrastive positive mining for unsupervised 3d action representation learning. In*European Conference on Computer Vision*, (pp. 36-51). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-19772-7_3

[14] Jindal, R., & Singh, I. (2022). Detecting malicious transactions in database using hybrid meta-heuristic clustering and frequent sequential pattern mining, *Cluster Computing*, 25(6), 3937-3959.

[15] Vignesh, K., Nagaraj, P., Muneeswaran, V., Selva Birunda, S., Ishwarya Lakshmi, S., & Aishwarya, R. (2022, July). A framework for analyzing crime dataset in R using unsupervised optimized K-means clustering technique, In*Congress on Intelligent Systems: Proceedings of CIS 2021, Volume 1*, (pp. 593-607). Singapore: Springer Nature Singapore. https://doi.org/10.1007/978-981-16-9416-5_43

[16] Feng, J., Zhang, R., Chen, D., Shi, L., & Li, Z. (2024). Automated generation of ICD-11 cluster codes for Precision Medical Record Classification, *International Journal of Computers Communications & Control*, 19(1). https://doi.org/10.15837/ijccc.2024.1.6251

[17] Zhang, L., Zhang, Y., Wei, Y., Zhang, T., Zhang, J. & Xu, J. (2023). Unveiling Patterns and Colors in Architectural Paintings: An Analysis by K-Means++ Clustering and Color Ratio Analysis, *Tehnički vjesnik*, 30 (6), 1870-1879. https://doi.org/10.17559/TV-20230514000634

[18] Yeh, J., Tsai, C.:(2022). A Graph-based Feature Selection Method for Learning to Rank Using Spectral Clustering for Redundancy Minimization and Biased PageRank for Relevance Analysis, *Computer Science and Information Systems*, 19(1), 141-164. https://doi.org/10.2298/CSIS201220042Y

[19] Zeng, M., Ning, B., Gu, Q., Hu, C., Li, S.(2022). Hyper-graph Regularized Subspace Clustering With Skip Connections for Band Selection of Hyperspectral Image, *Computer Science and Information Systems*, 19(2), 783–801. https://doi.org/10.2298/CSIS210830005Z

[20] Wang, A. & Gao, X. (2023). A Two-Stage Variable-Scale Clustering Method for Brand Story Marketing of Time-Honored Enterprises, *Tehnički vjesnik*, 30 (2), 373-380. https://doi.org/10.17559/TV-20230120000250

[21] Aka, A. C., Atta, A. F., Keupondjo, S. G., & Oumtanaga, S. (2023). An efficient anchor-free localization algorithm for all cluster topologies in a wireless sensor network. *International Journal of Computers Communications & Control*, 18(3). https://doi.org/10.15837/ijccc.2023.3.4961

Journal's webpage: http://univagora.ro/jour/index.php/ijccc/

This journal is a member of, and subscribes to the principles of,
the Committee on Publication Ethics (COPE).
https://publicationethics.org/members/international-journal-computers-communications-and-control

Cite this paper as:

Tang, M. C.; Cao, J.; Gong, D.Q.; Xue, G.(2024). Unsupervised Learning-Based Exploration of Urban Rail Transit Passenger Flow Characteristics and Travel Pattern Mining, *International Journal of Computers Communications & Control*, 19(2), 6422, 2024.
https://doi.org/10.15837/ijccc.2024.2.6422