Feature Mining of Shanghai Metro Commuters Based On K-Prototypes Clustering Method

Xin Yang *, Jinbing Ha and Yuting Tian

School of Economics and Management NUST, Nanjing, China

* Corresponding Author Email: yx123107123107@163.com

Abstract. Urban rail transit passenger travel behavior exhibits pronounced spatio-temporal heterogeneity, which poses significant challenges for accurately capturing passenger flow dynamics and enabling precise short-term travel forecasting. Unlike most existing studies that focus on station-level predictions, this research innovatively explores the spatio-temporal characteristics of travel behavior from the perspective of individual passengers. Leveraging one-card transaction data from Shanghai's rail transit system between April 1 and 30, 2015, we extract individual travel trajectory data in the form of OD chains through an OD relationship matching algorithm. A 13-dimensional clustering feature is constructed using a three-dimensional framework encompassing temporal, spatial, and travel intensity attributes. Notably, we introduce frequent travel OD share data as a critical metric to quantify the regularity of passenger travel patterns. To address the complexity of mixed-type data (including categorical and numerical variables), the K-Prototypes clustering algorithm is employed, demonstrating superior performance in handling heterogeneous datasets compared to traditional methods. The clustering results categorize passengers into four distinct groups, with the commuter category—accounting for 40% of the total—exhibiting the strongest spatial regularity. Further analysis of travel patterns across categories provides empirical evidence for identifying the primary sources of metro passenger flow, offering actionable insights for optimizing urban transit planning and demand management.

Keywords: Travelling features; K-prototypes algorithm; Passenger feature mining; OD chains.

1. Introduction

Urban rail transit, as an integral component of public transportation systems, has emerged as the primary mobility choice for urban residents due to its comprehensive advantages, including environmental sustainability, operational punctuality, enhanced safety, high passenger capacity, and energy efficiency. By 2023, China had established rail transit networks in 59 cities, operating 338 lines with a total national mileage exceeding 11,000 kilometers. As a cornerstone of urban public transport infrastructure, metro networks play a pivotal role in mass passenger transportation and traffic congestion mitigation.

However, the operational complexity of rail transit systems has intensified due to the dual pressures of continuous ridership growth and evolving urban lifestyle demands. Current operational models struggle to fully accommodate the diversified travel requirements of modern urban populations. Consequently, optimizing metro operational frameworks, enhancing operational efficiency, and leveraging advanced passenger flow analytics to better understand and guide travel behaviors have emerged as critical priorities for advancing urban rail transit development.

In passenger flow-based urban rail transit research, characterizing passenger flow patterns serves as a foundational step for many studies, with cluster analysis being a widely adopted method for this purpose. For instance, Wei et al [1]. employed clustering to segment groups whose members share learned embeddings, using inner product calculations to uncover flow patterns between urban functional zones, thereby enhancing spatial prediction accuracy. Liu et al. [2] proposed a two-step multi-station passenger flow prediction model integrating a Transformer encoder with K-Means clustering, wherein stations were categorized via an unsupervised approach. Wang et al. [3] adopted a multi-stage methodology to extract travel patterns from Beijing's bus and metro ridership data, classifying passengers through statistical analysis and unsupervised clustering. Li et al. [4] utilized

smart card data to cluster passengers via cluster analysis, further exploring correlations between classified passenger flow dynamics and influencing factors. These studies collectively underscore the methodological value of clustering in passenger flow characterization.

Common clustering techniques in urban rail transit research include K-Means, DBSCAN, Gaussian mixture models, and unsupervised deep learning frameworks. Xu et al. [5] applied K-Means clustering to group stations and paired it with a Random Forest model to predict bicycle lending volumes within clusters. Additionally, a multi-similarity inference model was deployed to calculate checkout probabilities and probabilistic occupancy predictions. Yue et al. [6] utilized DBSCAN to classify high-speed rail stations into three categories based on operational years. Guo et al. [7] compared similarity metrics within a K-Means framework, incorporating land-use and network characteristics to construct a clustering model. Lin et al. [8] analyzed spatiotemporal evolution of travel patterns using Shenzhen metro smart card data (2011–2017), employing a Gaussian mixture model with expectation-maximization (EM) clustering based on passenger trip frequency. Huang et al. [9] designed an unsupervised deep learning model to classify OD flow types by capturing trajectory shape features. These studies highlight the broad applicability of clustering methods in passenger flow analysis.

Despite its limited application in passenger flow prediction, the K-Prototypes algorithm demonstrates distinct advantages in handling mixed-type and large-scale datasets. Szepannek et al. [10] enhanced the Gower distance metric in K-Prototypes, validating its efficacy for numerical and categorical variables. Gao et al. [11] proposed an improved mixed-attribute K-Prototypes algorithm to cluster elderly populations based on functional capabilities, demonstrating robust performance for heterogeneous datasets. Hernandez et al. [12] applied a non-traditional K-Prototypes approach to define geostatistical domains for mineral grade estimation, proving its utility as an alternative to conventional methods. Kuo et al. [13] integrated K-Prototypes with genetic algorithms to optimize cluster centroids and weights, further enhancing classification performance via bagging-based ensemble methods.Shpigelman and colleagues[14]utilized the K-prototypes algorithm to categorize patients into four distinct clusters, leveraging clinical and radiological data collected upon hospital admission. This approach facilitated the identification of patient subgroups with unique characteristics, enabling more refined analyses and potential applications in clinical decision-making. These studies collectively affirm the algorithm's effectiveness in mixed-data contexts.

However, extant studies predominantly adopt a station-centric clustering perspective, with limited research focusing on individual passenger behavior. Zhao et al. [15] segregated temporal and spatial patterns independently, neglecting their interdependencies, while Li et al. [16] extracted passenger groups via soft constraints, limiting the scope of analysis. To address these gaps, this study constructs a 13-dimensional feature framework from an individual passenger-centric perspective, integrating temporal, spatial, and travel intensity dimensions. Notably, we introduce the frequent travel OD percentage as a spatial feature to quantify the regularity of metro passengers' route choices. By applying the K-Prototypes algorithm for passenger clustering, this research aims to unveil distinct travel patterns among passenger categories, providing a scientific foundation for optimizing metro operations and elevating service quality.

2. Data Sources and Preliminary Analysis

2.1. Basic Data

The dataset utilized in this study originates from the Shanghai Urban Rail Transit One Card Dataset, publicly released through the SODA Competition. This dataset encompasses automatic fare collection (AFC) swipe records from Shanghai's metro system between April 1–30, 2015, with a total of 234,596,521 records. In 2015, Shanghai's metro network comprised 14 operational lines, 337 stations, and an operational mileage of 548 kilometers, placing it among the most extensive systems in China. By the end of 2023, Shanghai Metro had expanded to 20 operational lines, with a 795.37-kilometer

network and significantly increased station numbers, reflecting its rapid development as a key urban transport infrastructure.

2.2. Data Cleaning Process

The primary objectives of data preprocessing were to eliminate outliers, duplicates, and incomplete records to ensure data integrity. The cleaning process followed three structured steps:

2.2.1 Temporal Filtering

Based on the official first and last train schedules for each line in 2015, travel records were restricted to the operational hours of 5:00–23:59. This step ensured that only valid passenger journeys within the designated service period were retained.

2.2.2 Duplicate Removal

Records with identical values across all fields (indicating duplicate entries) were identified and de-duplicated, retaining only one instance of each redundant record.

2.2.3 Missing Data Handling

Any record containing missing fields (e.g.incomplete swipe timestamps or station codes) was excluded to guarantee the completeness of retained data.

The original dataset of 234,596,521 records was reduced to 234,590,239 valid entries after preprocessing, with 6,282 records discarded. To illustrate the processed data structure, Table 1 presents a subset of cleaned AFC records from April 6, 2015, demonstrating the standardized format of retained entries.

Table. 1.Example of preprocessed passenger swipe data

Card Number	Transaction Date	Transaction Time	Subway Station Name	Transaction Amount
201530629	2015/4/6	21:25:29	Line 1 Gongkang Road	4
2004335807	2015/4/6	15:15:19	Line 3 Shanghai South Station	0
2004335807	2015/4/6	16:00:13	Line 3 Chifeng Road	5

2.3. Travelling Chain Identification

The origin-destination (OD) demand in a city or region serves as a critical input for numerous transportation applications,particularly in areas with high population mobility, where it plays a pivotal role in achieving efficient resource allocation.In this study, passenger OD chain data are extracted following a systematic process outlined below:

Step 1: Passenger Grouping

Passenger records are categorized into separate groups based on unique card numbers, ensuring that each passenger's travel history is independently processed. This step guarantees the integrity and independence of individual travel chains.

Step 2: Time Sorting

Within each passenger group, travel records are sorted in ascending order according to transaction timestamps. This chronological arrangement facilitates subsequent record pairing and ensures temporal consistency.

Step 3: Record Matching

During the traversal of passenger records, the nth record is paired with the subsequent (n+1) th record. The following checks are performed:

Verify whether the two records belong to the same passenger by confirming identical card numbers. If not, proceed to the next record.

If the card numbers match, continue to the next step of validation.

Step 4: Amount Condition Validation

For successfully paired records, the transaction amounts of the nth and (n+1) th records are examined:

Check whether the transaction amount of the nth record is zero, indicating an entry swipe.

Simultaneously, confirm that the transaction amount of the (n+1) th record is non-zero, indicating an exit swipe.

If these conditions are not satisfied concurrently, the nth record is ignored, and the process restarts from the (n+1) th record. Otherwise, proceed to the next step.

Step 5: Station Validation and Record Connection

After confirming that the two records belong to the same passenger and satisfy the amount condition, the station names in the paired records are further examined:

If the station names are identical, the nth record is skipped, and the process resumes from the next record.

If the station names differ, the two records are connected, forming a valid OD pair.

Upon completing the above steps, all travel records are matched based on their OD relationships. The integrated dataset includes key information such as card number, entry and exit times, entry and exit stations, and travel duration for each OD pair. OD pairs with excessively short travel durations are subsequently filtered out to eliminate anomalies, yielding the final processed underground passenger travel vector dataset.

The detailed algorithmic workflow is illustrated in Figure 1, providing a visual representation of the extraction process.

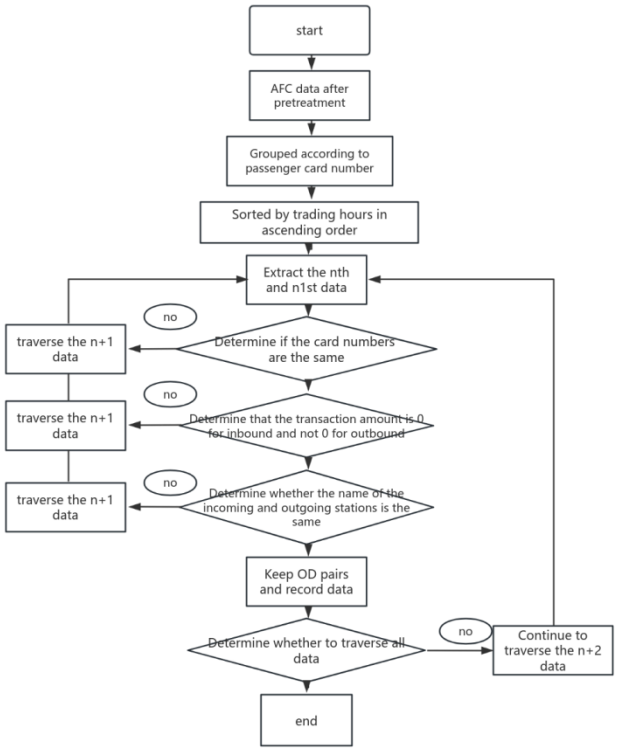

Fig. 1 Flowchart of the OD chain extraction algorithm for passenger travel

3. Passenger Travel Feature Extraction Based on AFC Data

The selection of passenger travel features is guided by the principle of maximizing the ability to distinguish different types of passengers. The primary objective of this study is to explore passenger travel behavior characteristics in depth through multidimensional analysis, thereby enabling the effective identification and classification of various passenger types.

To achieve this, the swipe card data from the Shanghai Metro system between April 1 and April 30, 2015, spanning a total of 30 days (21 working days and 9 non-working days), was selected as the research dataset. During this period, a total of 10,237,523 passengers traveled, accumulating 63,761,304 travel days, with an average of 6.23 travel days per capita. Both the number of travel days and the number of trips is critical indicators of travel frequency. Passengers who travel more frequently and over more days are considered more valuable for analyzing travel patterns.

However, daily trip counts exhibit significant fluctuations. For instance, some passengers may take the subway multiple times in a single day due to specific needs. While such behavior reflects high short-term travel frequency, it does not fully capture long-term travel patterns. Therefore, this study prioritizes the number of travel days as the primary criterion for screening and analyzing research data.

3.1. Travel Intensity Characteristics

3.1.1 Number of Travel Days

The number of travel days serves as a key indicator for assessing passengers' reliance on and demand for rail transit. A higher number of travel days typically reflects a stronger dependence on rail transit services. By analyzing the distribution of travel days across passengers, insights into their demand levels for rail transit can be obtained, as illustrated in Figure 2(a). This metric provides a foundation for understanding the intensity and regularity of passenger travel behavior.

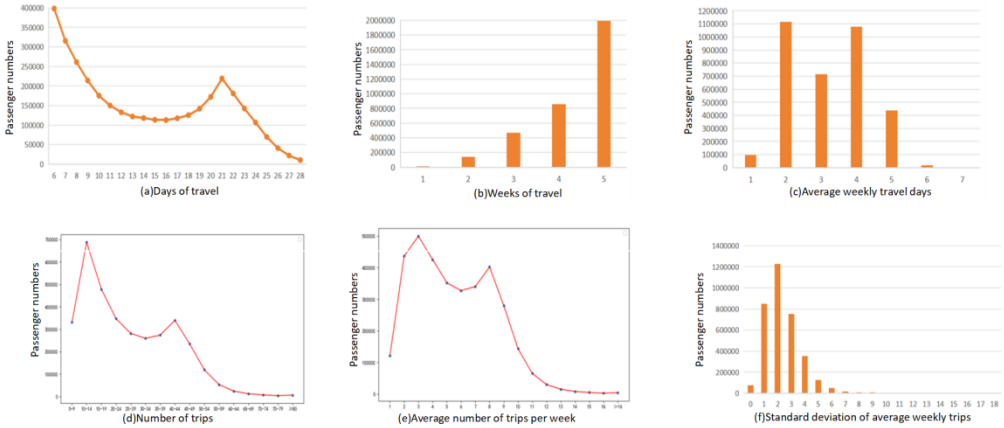

Fig. 2 Summary of passenger trip intensity characteristics

3.1.2 Weeks of Travel

The number of travel weeks serves as an indicator of the continuity of passenger behavior along the weekly time dimension. Commuting passengers typically adhere to a work-rest cycle, using the subway for repetitive trips to and from work or school on a weekly basis. Consequently, their travel behavior tends to exhibit greater continuity across weeks. In contrast, non-commuting passengers often display irregular patterns in their travel weeks. For instance, some non-commuting passengers may only use the subway during 1 or 2 weeks of the month, while others may have travel records every week.

It is important to note that the relationship between the number of travel weeks and the frequency of weekly trips is not always straightforward. For example, certain passengers may travel for all 5 weeks in a month but only register 1 or 2 trips per week. Such episodic travel behavior is illustrated in Figure 2(b). This highlights the need to differentiate between continuous and sporadic travel patterns when analyzing passenger behavior.

The average number of travel days per week serves as an indicator of the frequency of passenger trips within a weekly timeframe. A higher value signifies increased weekly travel frequency, as illustrated in Figure 2(c). To quantify this metric, the average weekly travel days for passenger i is calculated by dividing the total number of travel days (Da_i) by the corresponding number of travel weeks (We_i), as shown in Equation:

$$\mathrm{ADW}_i = \frac{\mathrm{Da}_i}{\mathrm{We}_i} \tag{1}$$

Here, $\mathrm{ADW_i}$ represents the average daily travel frequency for passenger i, with $\mathrm{Da_i}$ denoting the total number of days on which passenger i traveled and $\mathrm{We_i}$ representing the number of weeks during which these travels occurred. This calculation provides a standardized measure of weekly travel intensity, facilitating comparisons across different passenger groups.

3.1.3 Number of trips

As illustrated in Figure 2(d), passengers who travel 15 times or less account for 29.43% of the total, indicating sporadic travel behavior among this group. Overall, the number of passengers decreases as the number of trips increases. However, there is a slight increase in the number of passengers within the [40-44] trip range.

3.1.4 Average number of weekly trips

The average weekly travel frequency represents the average number of times a passenger uses rail transit per week. As shown in Figure 2(e) and expressed in Equation, if $\mathrm{Tr_i}$ denotes the total number of trips made by passenger i, then the average weekly travel frequency $\mathrm{ATW_i}$ can be calculated as:

$$\mathrm{ATW}_i = \frac{\mathrm{Tr}_i}{\mathrm{We}_i} \tag{2}$$

Here, $\mathrm{We_i}$ represents the number of weeks during which these travels occurred.

3.1.5 Standard deviation of average weekly number of trips

The standard deviation of the weekly travel frequency for passenger i, denoted as σ_{iwt}, is calculated using Equation:

$$\sigma_{iwt} = \sqrt{\frac{1}{\mathrm{We}_i}\sum_{n=1}^{\mathrm{We}_i}(\mathrm{TW}_{in}\text{-}\mathrm{ATW}_i)^2} \tag{3}$$

In this equation, $\mathrm{We_i}$ represents the number of travel weeks for passenger i, and $\mathrm{TW_{in}}$ indicates the number of trips taken by passenger i in each week, as shown in Figure 2(f).

3.2. Travel time characteristics

3.2.1 Average travel time consumption

Given the stable and punctual nature of rail transit, delays due to traffic congestion or weather conditions are rare. Therefore, analyzing the average travel duration for each trip can provide insights into passenger travel distances and patterns, as depicted in Figure 3(a).

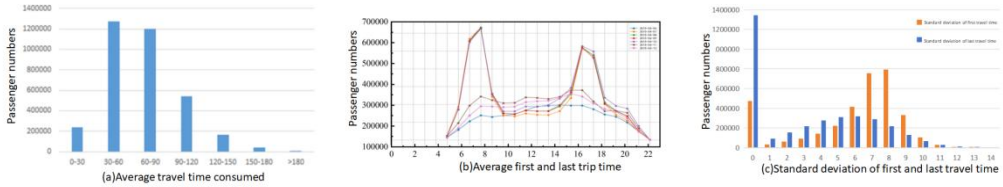

Fig. 3 Summary of passenger travel time characteristics

3.2.2 Average First and Last Travel Time

By statistically analyzing the travel time distribution of passengers during a specific week (for example, April 6-12) in April, both on weekdays and weekends, we can derive the regular travel patterns within a week, including peak and off-peak hours, as shown in Figure 3(b).

3.2.3 Standard Deviation of First and Last Travel Times

The standard deviation of first and last travel times for passengers can assess whether these times are relatively fixed within each day, indicating habits or routines. Combining this stability metric with the concentrated periods of first and last travel times provides a more comprehensive understanding of passenger travel type characteristics.

3.3. Spatial Travel Characteristics

Urban rail transit passengers exhibit spatial differences in terms of travel distance, selected routes, and origin-destination points. For regular travelers, their spatial travel characteristics often show certain stability. To conduct an in-depth analysis, travel records extracted from AFC data are coupled with station GIS information to derive metrics that reflect passenger spatial travel characteristics, including travel distance fluctuation coefficient, route entropy, transfer behavior, and frequent OD ratio.

3.3.1 Travel Distance Fluctuation Coefficient

The travel distance fluctuation coefficient quantifies the degree of variation in the entry and exit station locations for individual passengers, as illustrated in Figure 4(a).

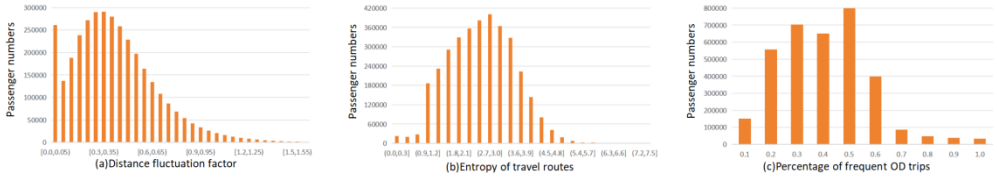

Fig. 4 Summary of spatial characteristics of passenger trips

3.3.2 Route Entropy

Route entropy is an indicator used to measure the diversity and uniformity of a passenger's spatial travel patterns. Based on the concept of information entropy, it reflects the distribution of a passenger's choices across different travel routes, as shown in Figure 4(b).

3.3.3 Transfer Behavior

When choosing rail transit for travel, passengers typically prefer origin-destination (OD) pairs located on the same line, allowing them to reach their destination without transferring. However, due to the actual distribution of metro stations, transfers between lines are sometimes necessary. To quantify this characteristic, we introduce a binary variable: if both the origin and destination of a passenger's trip are located on the same rail transit line (i.e., no transfer is required), the variable is assigned a value of 0. Conversely, if a transfer between different lines is needed to complete the journey, the variable is assigned a value of 1. For all passengers, the most frequently chosen travel route is analyzed to determine whether a transfer is required, with the results summarized in Table 2.

Table 2. Trip Interchange Statistics

Whether the trip is a transfer or not	Corresponding variable	Corresponding number of people	Corresponding ratio
Interchange	1	19695551	56.89%
No Interchange	0	1492455	43.11%

3.3.4 Frequent OD Pair Proportion

The frequent origin-destination (OD) pair proportion is an indicator used to characterize the regularity of subway passengers' travel routes based on the departure and arrival station data fields, as illustrated in Figure 4(c). Specifically, for each passenger, the most frequently occurring OD pairs among all travel vectors are selected as frequent OD pairs. The proportion of these frequent OD pairs relative to all travel vectors is referred to as the frequent OD pair proportion. A higher frequent OD pair proportion indicates greater stability in a passenger's travel routes, suggesting that they are more

likely to be commuting passengers. Conversely, a lower frequent OD pair proportion suggests more varied and possibly random travel behavior.

3.4. Summary of Multi-Dimensional Travel Characteristics

From the perspectives of travel intensity, travel time, and spatial travel characteristics, we have defined, extracted, and analyzed various travel features of passengers. This comprehensive analysis has enabled us to construct a detailed spatiotemporal feature indicator system that accurately depicts passenger travel behavior.

In summary, by combining AFC card data with station GIS information, we have extracted a total of 13 travel characteristic indicators for passengers, as summarized in Table 3.

Table 3. Feature Description Table

Feature Dimension	Feature Name	Value Field	Type
Intensity of travel	Days of travel	[6, 28]	Numeric type
	Weeks of travel	[1,2,3,4,5]	Sub type
	Average weekly travel days	[1,2,3,4,5,6,7]	Numeric type
	Number of trips	[1, 80]	Numeric type
	Average number of trips per week	[1, 16]	Numeric type
	Standard deviation of average weekly trips	[0, 18]	Numeric type
Travel time	Average travel time consumed	(0, 180]	Numeric type
	Average first and last trip time	[1,2,3,4,5,6,7,8]	Sub type
	Standard deviation of first and last travel time	[0, 14]	Numeric type
Travel space	Distance fluctuation factor	[0, 1.65]	Numeric type
	Entropy of travel routes	(0, 7.5]	Numeric type
	Whether the trip is transferring or not	[0, 1]	Sub type
	Percentage of frequent OD trips	[0, 1]	Numeric type

4. Construction of passenger clustering model

Clustering algorithms are unsupervised machine learning techniques that do not require predefined labels. They classify datasets based on different distance metrics, grouping objects with similar multi-dimensional features into the same clusters while maximizing differences between clusters. Clustering algorithms help uncover similarities and differences among data objects.

Currently, clustering analysis algorithms can be divided into two main categories: hierarchical clustering methods and dynamic clustering methods. Hierarchical clustering assigns samples to clusters in a single step without reassignment, which may lead to suboptimal results. In contrast, dynamic clustering algorithms transform the clustering problem into an optimization task involving sample combinations. By iteratively refining cluster assignments, these algorithms achieve optimal clustering results. Examples of dynamic clustering algorithms include k-means, k-modes, and k-

prototypes. Among these, k-means is suitable for numerical features, k-modes handles categorical features, and k-prototypes is optimized for mixed-type features, making it more versatile.

4.1. Basic Principle of K-Prototypes Clustering Algorithm

The k-prototypes algorithm is an effective hybrid clustering method that combines the strengths of k-means and k-modes, enabling simultaneous handling of both numerical and categorical features. This makes it highly adaptable to diverse data types.

4.1.1 k-means Algorithm

The k-means algorithm is one of the most widely used clustering techniques, applied across various fields. [17,18] Its basic principle involves randomly or strategically selecting k initial cluster centers, calculating the distances between each data point and the k centers, and assigning each point to the nearest cluster. The cluster centers are then updated by averaging the feature values within each cluster. This process is repeated iteratively until convergence or a maximum number of iterations is reached. The final cluster centers and data point assignments form the clustering result.

4.1.2 k-modes Algorithm

The k-modes algorithm is specifically designed for categorical (discrete) data and extends the principles of k-means. Unlike k-means, which use the mean to update cluster centers, k-modes uses the mode (the most frequent value) to update centroids, making it suitable for categorical data. The algorithm first selects k random initial cluster centers, calculates the Hamming distance between each data point and the centers, and assigns points to the nearest cluster. It then updates the centroids using the most frequent categorical values within each cluster. Iterations continue until the objective function converges.

4.1.3 k-prototypes Algorithm

With the increasing prevalence of mixed-type data (containing both numerical and categorical attributes), k-means and k-modes become inadequate due to their limitations in handling only one type of data.

The core idea of k-prototypes is to use Euclidean distance for numerical features and Hamming distance for categorical features.These distances are normalized and weighted to compute a combined distance metric. The algorithm begins by setting the number of clusters k and randomly selecting k initial cluster centers. Data points are assigned to clusters based on their combined distances. Cluster centers are updated iteratively: numerical features use the mean, while categorical features use the mode. The process continues until the loss function converges.

4.2. Distance Improvement Considering Feature Weights

The original k-prototypes algorithm measures the combined distance between samples by summing the Euclidean distance for numerical features and the Hamming distance for categorical features. However, it assumes equal weights for all features, which may not reflect the actual importance of different features. To address this limitation, scholars have proposed using the Entropy Weight Method (EWM) to assign weights to features.

4.2.1 Entropy Weight Method

The EWM is derived from information theory and is used to calculate feature weights in multi-criteria evaluation systems. For a given feature, its entropy reflects the degree of dispersion. A lower entropy indicates higher dispersion, increasing the feature's influence (weight) in the evaluation. Conversely, if all values of a feature are identical, its weight becomes zero.

4.2.2 Feature Weight Calculation

Let the passenger dataset contain n samples, each with m features (p numerical and $m - p$ categorical). The normalized passenger matrix is represented as:

$$P_{n\times m}=\begin{bmatrix} P_{11} & P_{12} & \cdots & P_{1p} & \cdots & P_{1m} \\ P_{21} & P_{22} & \cdots & P_{2p} & \cdots & P_{2m} \\ & & & & \vdots & \\ P_{n1} & P_{n2} & \cdots & P_{np} & \cdots & P_{nm} \end{bmatrix} \tag{4}$$

For the q-th numerical feature ($0 < q \leq m - p$) of passenger i, its proportion is calculated as:

$$y_{iq}=\frac{p_{iq}}{\sum_{i=1}^{n} p_{iq}} \tag{5}$$

The corresponding entropy is:

$$e_q=-k\sum_{i=1}^{n} y_{iq}\ln y_{iq}, \quad k=\frac{1}{\ln n} \tag{6}$$

The weight of the q-th feature is then:

$$w_q=\frac{1-e_q}{\sum_{j=1}^{q}(1-e_q)} \tag{7}$$

4.2.3 Weighted Sample Distance

The vector form of the i-th sample is $x_i = \{x_{i1}, \ldots, x_{iv}, x_{i(v+1)}, \ldots, x_{im}\}$, where the first v features are numerical. The weighted distance between samples i and j is:

$$d_{ij}=\sum_{l=1}^{v} w_l (x_{il}-x_{jl})^2+\sum_{l=v+1}^{m} w_l\, d(x_{il},x_{jl}) \tag{8}$$

Here, w_l represents the weight of the l-th feature. By incorporating feature weights, this approach prevents irrelevant attributes from influencing cluster center selection, improving clustering accuracy.

4.3. Optimization of Initial Cluster Center Selection

The k-prototypes algorithm exhibits significant advantages in handling mixed-type data, but it inherits the inherent drawbacks of k-means and k-modes algorithms during the initialization phase. In k-means, initial cluster centers are randomly selected from the dataset, which can lead to the selection of extreme or noisy data points, adversely affecting the subsequent calculation of cluster centroids and hindering the achievement of globally optimal clustering results. Similarly, in k-modes, low-frequency categorical attribute values might be chosen as initial cluster centers, resulting in clusters with few samples and weakening the overall clustering performance.

To address these shortcomings in the initial cluster center selection for the k-prototypes clustering algorithm and prevent the algorithm from getting trapped in local optima, we propose an optimized method for selecting initial cluster centers. For numerical features, cluster centers are chosen based on their weighted averages to reduce randomness and ensure that the selected centers more accurately reflect the true distribution of the data. For categorical features, we introduce the average dissimilarity metric to aid in selecting representative cluster centers. The average dissimilarity measures the differences between categories, helping to select the most representative categorical feature values as cluster centers.

4.4. Clustering Results Using the Elbow Method

The clustering result of the k-prototypes algorithm was evaluated using the elbow method to determine the optimal k value. By calculating the sum of distances between samples and cluster centers and plotting this against different k values, we identified the best k value.

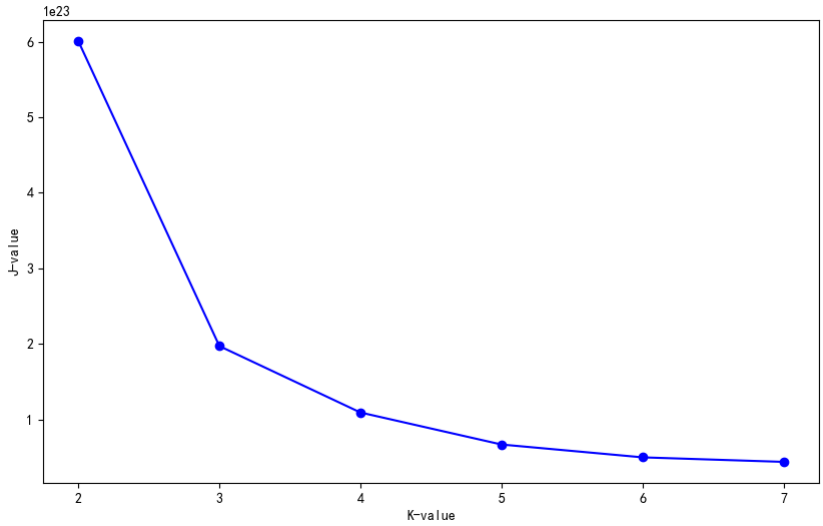

Fig. 5 Summary of spatial characteristics of passenger trips

As shown in Figure 5, as k increases, the value of J decreases. When $k = 4$, the rate of decrease in J becomes relatively slow, especially for changes from $k = 5$ to $k = 7$. Therefore, $k = 4$ is determined to be the optimal number of clusters for achieving the best clustering performance.

Based on the feature indicator system comprised of 13 indicators with low correlation utilized in this study, the clustering results for each type of passenger are summarized in Table 4. For numerical features, the cluster center is defined as the arithmetic mean of all values within the cluster, while for categorical features, the cluster center is identified as the value with the highest frequency within the cluster. These cluster centers effectively outline the travel behavior patterns characteristic of each passenger type. In this paper, for numerical features, the number of travel days, the number of trips, and the average travel time consumption have been rounded to the nearest whole number. All other numerical features are presented with two decimal places to facilitate clearer presentation and subsequent analysis.

Table 4. Passenger Clustering Centers by Category

Category	1	2	3	4
Days of travel	19	12	9	10
Weeks of travel	37	21	16	17
Average weekly travel days	2.54	2.13	2.33	2.23
Number of trips	70	55	63	64
Average number of trips per week	7.41	5.70	4.70	7.80
Standard deviation of average weekly trips	3.56	3.62	3.40	4.94
Average travel time consumed	2.36	2.80	2.88	2.35
Average first and last trip time	0.32	0.46	0.46	0.36
Standard deviation of first and last travel time	40.80%	30.94%	27.87%	39.00%
Distance fluctuation factor	5	5	4	4
Entropy of travel routes	1	0	1	1
Whether the trip is transferring or not	2	4	4	3
Percentage of frequent OD trips	6	5	6	5
Percentage of passengers	39.86%	29.55%	20.08%	10.52%

5. Characterization of Passenger Travel

By analyzing the clustering centers and their corresponding feature distributions for each passenger category, we can reasonably infer the travel behavior patterns of the four identified passenger categories.

Passengers in Category 1 exhibit a concentration of travel days around 21 days, which aligns perfectly with the 21 weekdays selected in April for this study. Their trip frequency is concentrated at 42 trips, averaging approximately 2 trips per day over 5 consecutive weeks. In terms of temporal characteristics, these passengers display typical morning and evening peak travel behaviors, with minimal standard deviations in their first and last travel times, indicating stable and regular travel patterns. These behaviors suggest that their activities are likely constrained by work or school schedules, consistent with commuting patterns. Spatially, both the travel distance fluctuation coefficient and travel route entropy are confined within a small range, indicating minor variations in travel distances and relatively fixed routes. Additionally, the peak percentage of frequent travel ODs (Origin-Destination) reaches 0.5, highlighting strong spatial regularity. Collectively, these findings suggest that the primary passenger group in Cluster 1 consists of typical commuters.

While some passengers in Category 2 show a narrower range of travel days, their average is 13 days, with the largest proportion traveling for five consecutive weeks. The standard deviation of their weekly trip counts is also relatively small, suggesting moderate travel intensity—second only to Category 1. Temporally, these passengers demonstrate centralized peak travel periods with low standard deviations in travel times, indicating stable travel time patterns but not exclusively for commuting purposes. Spatially, although they exhibit greater variability in travel distances and more diverse routes compared to Type 1, their peak OD share of frequent trips is also 0.5, indicating considerable spatial regularity. Notably, their average travel time consumption peaks at 50 minutes, lower than other types, and they predominantly engage in direct route travel without transfers. Based on this analysis, Category 2 passengers are classified as flexible commuters, characterized by significant travel intensity and spatial regularity but with more dispersed and variable first and last travel times.

Category 3 passengers, similar to Category 4, exhibit low trip frequencies and unstable spatiotemporal characteristics. However, their trip frequency is even lower, averaging 9 days of travel and 16 trips. Their average travel time consumption mirrors that of Category 1, around 63-64 minutes, with first trips concentrated midday and last trips in the evening. Spatiotemporally, their travel shows weaker regularity; the mean travel distance fluctuation coefficient is 0.46, indicating substantial fluctuations, and the mean travel route entropy is 2.9, reflecting the least stable travel routes among all categories. The percentage of frequent travel ODs is mainly concentrated between 0.1 and 0.2, underscoring highly unstable travel ODs. Consequently, Category 3 passengers are primarily identified as low-frequency travelers.

Category 4 passengers, compared to Category 1 and 2, exhibit low trip frequency and unstable spatiotemporal characteristics. Their average travel days total 10 days, with an average of 17 trips, peaking at 4 weeks of travel activity, implying an average of 2 travel days per week and weak travel intensity. Temporally, their first trips occur between 5-9 AM, and last trips between 10-3 PM, with more concentrated distributions of first and last trip times compared to Category 3. Spatially, the travel distance fluctuation coefficient for Category 4 is smaller, indicating less variability in travel distances and relatively stable routes. The percentage of frequent travel ODs is uniformly distributed between 0.1 and 0.5, suggesting a more regular spatial distribution than Category 3. Thus, the main passengers in Category 4 are inferred to be living passengers, characterized by moderate stability in travel patterns.

6. Concluseion

This study analyzes Shanghai Metro smart card data to reveal passenger travel characteristics. It shows that during morning and evening peak hours, there is temporal and spatial clustering of passenger flows, and travel regularity varies among passenger groups.Using time, space and travel intensity as dimensions, the study creates a 13-dimensional clustering feature, including a novel frequent trip origin-destination (OD) ratio metric to show route regularity.A K-Prototypes clustering algorithm is then applied to divide passengers into four groups. Typical commuters display clear

travel patterns during peak hours. Further analysis shows significant differences in travel repetition and regularity across passenger groups, offering important data for Metro traffic planning and management.The study proves the K-Prototypes model's effectiveness and accuracy in handling large-scale passenger flow data. It also offers scientific evidence for optimizing Metro operations and improving service quality.The study also provides a methodological framework for metropolitan transportation studies, with practical implications for peak-hour scheduling and personalized services in smart transit systems.Future studies could explore factors influencing travel behavior and passenger grouping methods to further support intelligent urban rail transit management.

Acknowledgements

The authors gratefully acknowledge the financial support from xxx funds.

References

[1] Wei, Q., Qiu, Y., Wen, Y. Cluster-based spatiotemporal dual self-adaptivenetwork for short-term subway passenger flow forecasting. Appl. Intell., 2022, 52: 14137–14152.

[2] Liu, L. J., Wu, M. X., Chen, R. C., Zhu, S. Z., Wang, Y. A hybrid deep learningmodel for multi-station classification and passenger flow prediction. Appl. Sci.-Basel, 2023, 13(5): 2899.

[3] Wang, L., Chen, Y., Wang, Y., Sun, X., Wu, Y., Peng, F., Song, G. Identificationand classification of bus and subway passenger travel patterns in beijing usingtransit smart card data. J. Adv. Transp., 2023: 6529819.

[4] Li, P., Wu, W., Pei, X. A separate modelling approach for short-term bus passen-ger flow prediction based on behavioural patterns: A hybrid decision tree method.Physica A, 2023, 616: 128567.

[5] Xu, H., Duan, F., Pu, P. Dynamic bicycle scheduling problem based on short-term demand prediction. Appl. Intell., 2019, 49: 1968–1981.

[6] Yue, Y. F., Chen, J., Feng, T., Wang, W., Wang, C. Y., Ma, X. W. New classifica-tion scheme and evolution characteristics analysis of high-speed railway stationsusing large-scale mobile phone data: A case study in jiangsu, china. J. Transp.Eng. Part A-Syst., 2023, 149(11): 04023108.

[7] Guo, Y. L., Zhu, Z. J., Jiang, X. H., Chen, T., Li, Q. Analyzing the impactsof land use and network features on passenger flow distribution at urban railstations from a classification perspective. Sustainability, 2024, 16(9): 3568.

[8] Lin, M., Huang, Z., Zhao, T., Zhang, Y., Wei, H. Spatiotemporal evolution oftravel pattern using smart card data. Sustainability, 2022, 14(15): 9564.

[9] Huang, Z. C., Zheng, H., Yang, K. Multitype origin-destination (od) passengerflow prediction for urban rail transit: A deep learning clustering first predictingsecond integrated framework. J. Adv. Transp., 2024: 6629500.

[10] Szepannek, G., Aschenbruck, R., Wilhelm, A. Clustering large mixed-type data with ordinal variables. Adv. Data Anal. Classif., 2024.

[11] Gao, Y., Hu, Y., Chu, Y. Elderly individuals with similar abilities are likely tohave similar care needs. Math. Probl. Eng., 2023: 7114343.

[12] Hern´andez, H., Alberdi, E., Goti, A., Oyarbide-Zubillaga, A. Application of thek-prototype clustering approach for the definition of geostatistical estimationdomains. Mathematics, 2023, 11(3): 740.

[13] Kuo, R. J., Wu, C. Y., Kuo, T. An ensemble method with a hybrid of geneticalgorithm and k-prototypes algorithm for mixed data classification. Comput. Ind.Eng., 2024, 190: 110066.

[14] Shpigelman, E., Hochstadt, A., Coster, D., et al. Clustering of clinical andechocardiographic phenotypes of covid-19 patients. Sci. Rep., 2023, 13: 8832.

[15] Zhao, J. J., Qu, Q., Zhang, F., Xu, C. Z., Liu, S. Y. Spatio-temporal analysis ofpassenger travel patterns in massive smart card data. IEEE Trans. Intell. Transp.Syst., 2017, 18(11): 3135–3146.

[16] Li, Y. C., Zhang, T., Lv, X. F., Lu, Y. X., Wang, W. S. Profiling public transitpassenger mobility using adversarial learning. ISPRS Int. J. Geo-Inf., 2023, 12(8): 338.