Research on Hierarchical Clustering Method of Urban Rail Transit Passengers Based on Individual Portrait

Jingwen XUE1,a, Bingfeng SI2,Hongmeng CUI3, Shiguang ZHU4
1 School of Traffic and Transportation, Beijing Jiaotong University, Beijing, China
2 Institute of Systems Sciences, Beijing Jiaotong University, Beijing, China
3 School of Traffic and Transportation, Beijing Jiaotong University, Beijing, China
4 School of Traffic and Transportation, Beijing Jiaotong University, Beijing, China
a email: 19120925@bjtu.edu.cn

Abstract. The travel behavior of urban rail transit passengers has spatial-temporal differences, which increases difficulties of accurately grasping the evolution of passenger flow and achieving accurate short-term travel forecasts. Therefore, it’s particularly significant to deeply analyze passenger travel spatial-temporal characteristics and to achieve refined passenger classification. Using the AFC data of the urban rail transit system, this paper focuses on the construction of individual travel spatial-temporal portraits. An improved DBSCAN-based method for extracting individual travel spatial-temporal characteristics is proposed, and Python software is used to realize the visual drawing of individual portraits, so as to dig out potential passenger travel rules. Furthermore, a hierarchical clustering method of passengers based on individual portraits is proposed, and the quantified extracted individual travel feature values are clustered through the two-step clustering method, and the refined classification of passengers is realized layer by layer. Finally, the passengers are divided into four categories: definite passengers, flexible passengers, random passengers and one-way ticket passengers. The rationality and accuracy of the classification results are verified from multiple angles. The study clarifies the travel spatial-temporal characteristics of different types of passengers, improves the objectivity of passenger classification, and lays the foundation for short-term travel prediction.

1. Introduction
As the travel subject, the individual economic attributes and travel preferences[1] of urban rail transit passengers has spatial-temporal differences, which increases difficulties of accurately grasping the evolution of passenger flow and achieving accurate short-term travel forecasts. Short-term travel prediction is of great significance to realize the rational planning and management of urban traffic. So it is particularly important to deeply analyze passenger travel spatial-temporal characteristics and realize the fine classification of passengers.

Traveling passengers have significant heterogeneity, which can not be simply regarded as a group to analysis[2]. The research objects of passenger travel spatial-temporal characteristics mainly include two categories: Firstly, study travel rules of passengers with different attributes according to the classification of passengers' socioeconomic attributes; Secondly, based on classifications of passengers' travel spatial-temporal information, study travel rules of different types of passengers.

Travel analysis and passenger classification based on passengers' socioeconomic attributes mainly focus on SP/RP questionnaire data: for example, based on passengers' economic attributes and travel
range obtained by sampling survey, Mohamed K. El Mahrsi et al. [3] carried out the mixed clustering of individual travelers by using the one-dimensional mixed model, and the spatial unit clustering analysis was carried out by using the hidden Markov random field model. Li Yanjin et al. [4] designed SP/RP questionnaire to obtain the economic attributes and travel purposes of high-speed rail passengers. By stratified sampling and constructing a classification model based on Logistic regression method, the passengers were divided into three categories. Although these studies have made some progress, there are some shortcomings such as small data coverage, high investigation cost and long data update period, which is not conducive to further short-term travel prediction.

In recent years, AFC data presents the characteristics of massive database, persistent storage and full sample records, which makes it more perfect to analyze travel characteristics and realize passenger classification from an individual perspective. LEMK et al. [5] took the starting and ending points and departure time as indicators, and passengers were divided into four categories by density clustering. Based on the AFC data of Beijing Metro, Zou Qingru et al. [6] constructed the objective classification index from the perspective of "consumption behavior", and passengers were divided into five categories. The existing researches on passenger classification based on AFC data mostly adopt density clustering, but few researches have realized the hierarchical clustering process of passengers based on multiple clustering indexes, and the classification precision needs to be improved.

Based on AFC data of urban rail transit, this paper focuses on the construction of individual travel portraits and proposes an improved DBSCAN-based extraction method of individual characteristics. Clustering indicators are selected reasonably by the difference of passenger travel rules, and Python is used to realize the visual drawing of individual portraits, visually displaying the potential passenger travel rules. Furthermore, a hierarchical clustering method of passengers based on individual portraits is proposed, which uses two-step clustering to cluster the quantitatively extracted individual spatial-temporal characteristic values, sets different clustering variables in stages, and realizes refined classification of passengers layer by layer. Finally, passengers are divided into four categories: deterministic passengers, elastic passengers, random passengers and one-way ticket passengers, providing support for the development of short-term travel prediction.

2. Passenger individual spatial-temporal portrait construction

Constructing individual spatial-temporal portrait means mining passenger travel data, quantitatively extracting travel characteristics, and realizing visual description of individual travel laws [7]. To accurately construct individual portraits, it’s significant to effectively extract reliable individual travel spatial-temporal characteristic values. Individual travel spatial-temporal characteristics of passengers refer to certain regularity with statistical characteristics when a single passenger travels at the same time and between the same OD. Because passengers' travel preferences are different, and their travel behaviors are influenced by many factors, the individual travel time and space characteristics of different passengers are quite different. AFC system can obtain and store the travel records of a single passenger in real time, which provides an accurate data basis for in-depth excavation of individual travel spatial-temporal characteristics.

The individual travel characteristics can be described by mathematical method as follows:

\[ E_{ni}^{i} = \{T_{ni}^{i}, O_{ni}^{i}, D_{ni}^{i}, f_{ni}^{i}\} \]  

Among them, \( E_{ni}^{i} \) represents the category \( i \) travel characteristic value of passenger \( n \), \( T_{ni}^{i} \) indicates the occurrence period of the category \( i \) travel characteristics of passenger \( n \), \( O_{ni}^{i} \) and \( D_{ni}^{i} \) respectively indicate the inbound and outbound stations of the passenger's category \( i \) travel characteristics, \( f_{ni}^{i} \) indicates the number of historical trips of the passenger's category \( i \) travel characteristics.

There may also be differences in individual passengers' travel behavior every time they take urban rail transit. Therefore, there may be various types of individual travel characteristics in history travel records. The individual travel characteristic matrix of passenger \( n \) can be expressed as follows:
Constructing individual portrait includes two steps: extraction of travel spatial-temporal features and visual drawing of travel rules. Because the number of individual travel spatial-temporal features is unknown, the clustering method based on the number of categories is not suitable for the extraction of travel eigenvalues. Density-based Spatial Clustering of Application with Noise (DBSCAN) method does not need to preset the initial number of cores or clusters [8]. Therefore, this paper proposes an improved DBSCAN-based method for extracting individual travel spatial-temporal features. Based on the quantitative extraction results of individual travel time and space features, Python can be used to realize the visual expression of individual portraits.

2.1. Improved DBSCAN-based extraction method of individual travel characteristics

DBSCAN algorithm is a clustering method based on density and noise. It assumes that the categories of research objects are determined by their close distribution, and similar research objects will be divided into a cluster. There are two global parameters: maximum density distance \( \text{EPS} \), and minimum similarity points \( \text{MinPts} \), and their value range can be adjusted according to different research scenarios, and there is no fixed value.

In this paper, DBSCAN algorithm is applied to the extraction of individual travel features, and the historical travel stations of passengers are taken as research objects. \( \text{EPS} \) indicates the ideal time interval for passengers to travel between stations, used to measure whether there is time regularity for individuals to travel between two stations; \( \text{MinPts} \) indicates the minimum travel times of passengers in a certain period of time and a certain OD, used to judge whether there is spatial regularity in passenger travel. If the passenger's historical trip times \( f < \text{MinPts} \) in a certain OD in a certain period of time, it is considered that the passenger travels randomly between the OD in this period; On the contrary, the passenger travels regularly between the OD in this period.

AFC data preconditioning: Time is converted into minute format, date information is removed, and clustering distance is judged according to differences; For different OD codes, in order to ensure that different ODs do not interfere with each other during clustering, the interval after assignment of each OD code should be greater than \( \text{EPS} \).

The algorithm flow is as follows:

- Step 1: Set global parameters \( \text{EPS} \) and \( \text{MinPts} \).
- Step 2: Enter the historical travel record set \( H^i_n = (t_i, O_i, D_i) \) of passenger \( n \), in which \( t_i \), \( O_i \), \( D_i \) respectively represent the inbound time, inbound and outbound stop of the category \( i \) historical travel record of passenger \( n \). Mark all travel records \( i_nH \) that the initial state is On.
- Step 3: Randomly select one record \( H^\sigma_n \) from all records marked On, and change its mark to Off. Circle with \( t_{\sigma} \) as the center point and traverse all On objects of the passenger data set \( i_nH \). For any cluster \( iE \), take the minimum inbound time \( t_{i_{\text{min}}} \) for \( t_O \), the maximum inbound time \( t_{i_{\text{max}}} \) for \( t_D \) in the cluster, then \( t_O \) and \( t_D \) form the period \( T^i_n \); if \( f < \text{MinPts} \), the next record \( H^\mu_n \) is randomly selected, then go to step 3; If all the history records \( H^i_n \) have been traversed (all records are marked Off), then go to step 5.
• Step 5: If the total number of core points marked by passenger \( n \) is 0, which means that the passenger has no fixed travel rule, and directly turn to step 8. If the total number of core points marked by passenger \( n \) is not 0, go to step 6.

• Step 6: For any two core points under a certain OD, if their inbound time interval is less than \( \varepsilon_{\text{PS}} \), they are connected, and all the interconnected core points and the points in their neighborhood form a combined cluster. If a core point is not connected with any other core point, the core point and the points in its neighborhood form a cluster separately. Travel records without clusters are recorded as noise points, indicating random travel.

• Step 7: Output all clusters \( E_n^i = \{T_n^i, O_n^i, D_n^i, f_n^i\} \) generated by passenger \( n \).

• Step 8: Enter all the historical travel record sets of the next passenger ID, and go to Step 2. Until all passenger IDs are traversed, the algorithm ends.

2.2. Visualization of individual travel spatial-temporal portrait
Taking individual passengers as analysis units, the specific process is as follows:

• Step 1: Integrate the individual travel spatial-temporal characteristic matrix \( E_n \) of passenger \( n \) as the basic input data for drawing individual portraits.

• Step 2: Determine the coordinate axis of individual portrait. The time series of passengers swiping their cards (min) is taken as the abscissa, and the names of passengers' historical inbound stations are called the ordinate.

• Step 3: Determine individual portrait elements. The node represents a single historical travel record, that is, the travel record from the starting station to the terminal station within a certain period of time; Different colors are used to distinguish the individual spatial-temporal characteristics of different categories, in which black dots represent random travel records.

• Step 4: Based on the travel spatial-temporal characteristic matrix \( E_n \), draw and output all individual travel portraits of passenger \( n \) one by one.

3. Passenger hierarchical clustering method based on individual portrait
This paper puts forward a hierarchical clustering method of passengers based on individual portraits, which selects different clustering indexes in stages, clusters the spatial-temporal characteristic values of passengers layer by layer, and gradually realizes the refined classification of passengers.

The hierarchical clustering process is as follows:

• Stage 1: Set the passenger ticket card types in AFC data as cluster index 1 to initially divide passengers. Screen out all the one-card passengers and continue to divide them in stage 2, then define and output other passenger categories directly according to the ticket card types.

• Stage 2: Set \( \text{MinPts} \) as cluster index 2 to roughly classify all the input one-card passengers. If historical travel times \( f < \text{MinPts} \) of one-card passengers, the individual travel characteristics can not be obtained by DBSCAN clustering, and they can be divided into one class separately; The rest of all one-card passengers are transferred to stage 3 to continue division.

• Stage 3: Set \( \varepsilon_{\text{PS}} \) as cluster index 3 to cluster the rest of the input one-card passengers in two steps, and the more detailed passenger classification results are defined and output according to the individual travel spatial-temporal characteristic values.

4. Case analysis

4.1. AFC data preconditioning
In this paper, the whole month's AFC data of Suzhou Metro in July, 2019 is taken as an example. There are 14,538,006 travel records, including information such as arrival and departure time, arrival and departure station, arrival and departure route, card type, passenger ID, equipment number, etc. Repeated and abnormal data were cleaned, and 3650897 records were eliminated.
4.2. Constructing individual travel portraits
Set $EPS$ and $MinPts$ values. The passengers who travel regularly are defined as those who travel more than or equal to 4 times in the same OD for whole months, that is, $MinPts = 4$. In order to get reasonable $EPS$ value, the number of individual travel features were statistically analyzed, and it can be known that: (1) Suppose $MinPts = 4$, when $EPS = 5\text{min}$, the travel features can not be extracted completely because of the short inbound time interval; when $EPS$ was set to 15min, 30min and 60min in turn, all the travel features can be extracted, and there is no significant difference; (2) The maximum time interval between inbound stations and $EPS$ have a significant positive correlation between values, if $EPS$ value is too large, the maximum time interval will multiply and the clustering accuracy will be affected. Table 1 shows the extraction of travel features for a passenger in $MinPts = 4$, different $EPS$. To extract the complete number of travel features and ensure the clustering accuracy, this paper chose the intermediate value $EPS = 15\text{min}$ and $EPS = 30\text{min}$ to construct individual portraits firstly.

| MinPts | EPS   | Number of Travel Characteristics | Maximum Travel Time Interval |
|--------|-------|---------------------------------|-----------------------------|
| 4      | 5min  | 5                               | 8:08–8:16 (8min)            |
| 4      | 15min | 9                               | 8:00–8:23 (23min)           |
| 4      | 30min | 9                               | 10:04–10:53 (49min)         |
| 4      | 60min | 9                               | 14:39–16:23 (104min)        |

Taking a passenger A as an example, the individual travel feature matrix $E_A$ is as follows:

$$E_A = \left( E^1_A, E^2_A, \ldots, E^7_A \right)^T$$

8:38 – 9:42, Yuexi, Guangjinan Road.8
10:03 – 10:52, TayuanRoad, XinghuStr et.6
12:12 – 13:07, Yuexi, Guangjinan Road.5
13:25 – 14:31, Yuexi, Guangjinan Road.7
14:17 – 14:58, LeBridge, XujiangRoa d.7
15:03 – 16:32, Yuexi, Guangjinan Road.9
16:11 – 17:36, LeBridge, XujiangRoa d.6

Fig.1 is individual portrait visualization of passenger A.

Fig.1 Individual Portrait Visualization of Passenger A
4.3. Passenger hierarchical clustering process

4.3.1. Preliminary classification and rough classification of passengers
There are 14.53 million records in Suzhou data selected of the study, including 10.88 million of one-card passengers, involving about 1.34 million passenger IDs; Data of one-way ticket and other special ticket card types has no corresponding fixed passenger ID, which is impossible to statistically analyze individual travel rules. Therefore, they are collectively called one-way ticket passengers.

Select $\text{MinPts} = 4$ as a clustering index. Among the 1.34 million passenger IDs, there are 670,000 passengers’ total trips frequency $f < 4$ times in July, which are defined as inactive random passengers; The remaining 670,000 one-card passengers’ total trip frequency $f \geq 4$ times, which are classified by two-step clustering.

4.3.2. The first two-step clustering of passengers
Select $\text{EPS} = 15\text{min}$ as a clustering index, and the quantitative information table of travel characteristics of the remaining 670000 one-card passengers was obtained. Five items were initially selected as cluster variables, including the total number of historical trips, the total number of ODs, the average OD trip times, the noise ratio (the ratio of random trips to total trips), and the number of travel characteristics. SPSS and Euclidean Distance were used to evaluate the rationality of them. The value range is between 0-1, and the closer to 1, the more important the variable is. The evaluation results are shown in Fig.2. It can be seen that the influence of average OD trip times and noise ratio on clustering results is much higher than the other three items. Therefore, the first two-step clustering was conducted for these passengers according to these two items. The results are shown in Table 2.

SPSS evaluates the clustering effect according to the contour coefficient. The closer the contour coefficient is to 1, the better the clustering effect is. It can be seen from the clustering quality chart that the clustering quality is close to 1, and the effect is good. Analysis of Table 2 shows that 99% of class 1 passengers and 68% of class 2 passengers travel regularly; Although there are obvious differences between them, the noise ratio is high, the travel rules are uncertain, and the classification results do not meet the ideal standards.

In order to optimize the clustering results, this paper added the noise ratio dispersion degree (the distance from each ID noise ratio to the average value of all ID noise ratios) to re-cluster these passengers.
It can be seen from the clustering quality chart (Fig.3) that the improved clustering quality is closer to 1 than the first clustering, and the clustering effect is better. Analysis of Table 4 shows that 82% of class 2 travel regularly, which can be defined as deterministic passengers, accounting for 13.5%; While the remaining 86.5% passengers will be regarded as uncertain passengers.

### Table 3 Improved Two-step Clustering Results

| Cluster Number | Average OD Trip Times | Noise Ratio | Noise Ratio Dispersion Degree | Proportion |
|----------------|-----------------------|-------------|-------------------------------|------------|
| 1              | 3.28                  | 0.55        | 0.13                          | 14.60%     |
| 2              | 8.01                  | 0.18        | 0.42                          | 13.50%     |
| 3              | 1.46                  | 1           | 0.4                           | 71.90%     |

Fig.3 Evaluation of Improved Cluster Results

4.3.3. Second two-step clustering of passengers

Keep $MinPts = 4$, relax $EPS$ value range, setting $EPS = 30$min as a new clustering index, the travel characteristic of all uncertain passengers was obtained. The clustering results are shown in Table 4. It can be seen that the average noise ratio of class 1 passengers is 0.37 (lower than class 1 at the first clustering, 0.55), and 63% of them travel regularly, but a small number travel randomly, regarded as elastic passengers; 98% of class 2 passengers travel irregularly, regarded as active random passengers.

### Table 4 Clustering Results Of Non-deterministic Passengers

| Cluster Number | Average OD Trip Times | Noise Ratio | Noise Ratio Dispersion Degree | Proportion |
|----------------|-----------------------|-------------|-------------------------------|------------|
| 1              | 3.42                  | 0.37        | 0.23                          | 15.79%     |
| 2              | 1.45                  | 0.98        | 0.37                          | 84.21%     |

So far, all the traveling passengers are divided into four categories: one-way ticket passengers, deterministic passengers, flexible passengers and random passengers.

4.3.4. Analysis of clustering results

Fig.4 shows the relationship between the first travel time and the number of passengers on working and non-working days. It can be seen that the first travel time of flexible passengers on working days is mostly concentrated in the morning peaks, while on weekends is not relatively fixed; The travel rules of deterministic and flexible passengers are similar, but deterministic passengers account for a larger proportion in the morning peak; There is no obvious regularity of active random passengers. Different from the first two categories, a considerable number of passengers will choose to travel in the afternoon.

Fig.4 Relationship between the First Travel Time and The Number of The Three Types of Passengers
Fig. 5 is a visualization result obtained according to OD flow direction of first trip of three types of passengers on Monday, combined with latitude and longitude of the station. The curve represents the OD flow direction, and different colors represent different levels of OD flow. Green line indicates that the flow rate is between 50 and 100 times; The yellow line indicates that the flow rate is between 100 and 200 times; The red line indicates that the traffic is over 200 times.

It can be seen from the figure that elastic passengers are mostly green lines, and deterministic passengers are mostly yellow and red lines. They tend to travel regularly in multiple fixed rooms. Active random passengers are mostly green and yellow lines, and the total number is obviously more. They tend to travel irregularly among multiple unfixed OD, and the traveling area is also larger.

To sum up, the rationality and stability of the proposed hierarchical clustering method are verified through the comparative analysis of the first trip time and OD flow rules of elastic, deterministic and active random passengers, which realizes the refined passenger classification.

5. Conclusion

Based on the AFC data of Suzhou Metro in July 2019, this study digs the passenger travel rules and constructs the individual travel portrait: An improved DBSCAN-based method for extracting individual travel spatial-temporal characteristics is proposed; The quantitative extraction results of travel features are analyzed, and the visual drawing of individual portraits is realized by Python. Furthermore, a hierarchical clustering method of passengers based on individual portraits is put. Different clustering indexes are set in stages, and individual trip feature values are clustered in two steps layer by layer, thus realizing the refined classification of passengers. Finally, passengers are divided into four categories: deterministic, elastic, random and one-way ticket passengers. The travel rules of various passengers are analyzed from multiple angles, and the rationality and accuracy of the classification results are verified. However, how to calculate the temporal and spatial distribution of passenger flow based on classification and put forward reasonable short-term travel prediction methods will be the goal of further study.

References

[1] Liu X.M., Lu X.Y., Sun Q.X. (2017) Study on Travelers’ Route Selection Behavior Based on Different Preferences. J. Journal of Chongqing Jiaotong University (Natural Science Edition). 36: 102-106.

[2] CHENC Y-H, CHEN S-Y. (2015) Perceived Accessibility, Mobility, and Connectivity of Public Transportation Systems. J. Transportation Research Part A: Policy and Practice. 77: 386-403.

[3] El Mahrsi M K, Côme E, Oukhellou L, etc. (2017) Clustering Smart Card Data for Urban Mobility Analysis. J. IEEE Transactions on Intelligent Transportation Systems. 18: 712-728.

[4] Li Y.J., Luo X, Liu Y, etc. (2018) High-speed Rail Passenger Transport Market Segmentation based on Logistic Regression Model. J. Journal of Transportation Engineering and Information. 16: 32-37.

[5] LEMK, BHASKARA, CHUNGE. (2015) Passenger Segmentation Using Smart Card Data. J.
IEEE Transactions on Intelligent Transportation Systems. 16: 1537-1548.

[6] Zou Q.R., Zhao P, Yao X.M.. (2018) Passenger Classification of Urban Rail Transit based on Fare Collection Data. J. Journal of Transportation Systems Engineering and Information Technology. 18: 223-230.

[7] Liang Q, Weng J.C., Zhou W, etc. (2020) Construction of Individual Travel Characteristic Map of Bus Commuters and Similarity Discrimination. J. Journal of Chongqing Jiaotong University (Natural Science Edition). 39: 6-13.

[8] He N.Z., An J.C.. (2020) Improving Chameleon's Clustering Algorithm DN-KChameleon with DBSCAN. J. Computer Engineering and Science. 11: 1-7.