Analysis of Spatial Travel Association Rules for Rail Transit Based on AFC and POI Data

Yujie Yang, Hui Li, Qingsong Du, Zhenbo Liu, Zihao Feng

Xihua University, Chengdu 610039, China.

Abstract: In order to explore the spatial distribution rules and causes of urban rail transit passenger travel, this paper mines the spatial 1-frequent itemset and 2-frequent itemsets of weekdays and weekends metro passenger travel based on Apriori algorithm using the continuous week of Automatic Fare Collection System (AFC) swipe card. At the same time, the K-Means algorithm is used to cluster the subway stations and explore the causes of association rules by combining the Point of Interest (POI) data of the same period within the radiation range of the subway stations. The study shows that the spatial distribution pattern of inbound and outbound passenger flow of Shanghai rail transit is consistent between weekdays and weekends, and the outbound passenger flow is more concentrated than the inbound passenger flow, and the significance of weekends is higher; the spatial distribution of metro stations is "circled"; the analysis of the high-lift association rules show that a large passenger flow group centered on the type 3 station is formed in the spatial location, and the passenger flow within the group is mainly commuter flow with separation of employment and residence. The association rule mining of metro passenger travel data is beneficial to understanding the spatial distribution pattern and causes of metro ridership, which can provide reference for rail network planning and operation management.

Keywords: Urban Rail Transit; Association Rules; Spatial Travel Characteristics; AFC Data; POI Data

Introduction

With the acceleration of urbanization in China, the urban public transportation network with rail transit as the backbone has been increasingly improved. By the end of 2022, more than 40 cities in China have planned and built rail transit systems, and the density of rail transit line network is increasing [1]. In order to study urban rail transit travel behavior more deeply, we can grasp the dynamic changes of passenger flow between stations by analyzing the travel characteristics of passengers and using the passenger flow change law, and also optimize the subway travel strategy according to the association rules to maximize the utility of the rail transit system.

Currently, studies on urban rail travel behavior either use commuters with significant time characteristics as the entry point for data mining, or explore the relationship between passenger flow and stations based on AFC card swipe data and land use characteristics and POI data. Quan Liang firstly obtained commuters' travel chain data by matching RP survey with public transportation multi-source data, and based on FP-Growth algorithm, extracted commuters' stable population by constructing frequent pattern tree structure, and refined different levels of correlations and intrinsic patterns of commuters' travel [2]. Xin et al. matched and analyzed the travel chain of travelers from AFC swipe card data, divided the relevant rules between traditional nine-to-five commuting behavior and non-nine-to-five commuting behavior, analyzed commuting travel behavior based on spatial and temporal distribution characteristics, mined the interrelationships between highly relevant interchange services, and applied them to the operation strategies of metro networks [3]. Chu Fan uses the Apriori algorithm to carry out single-dimensional and multi-dimensional passenger flow feature mining in three dimensions: time dimension, spatial dimension, and route dimension, respectively, and focuses on analyzing the rail traffic travel characteristics on weekdays [4]. Yan Xu first defined land usage types of travel Origin-Destination (OD) locations using spatial clustering analysis of city's Points of interests (POI) data, and then discover the most representative travel scenarios using association

rules mining method with car-sharing records data, and analyzed the spatio-temporal consumption characteristics and competitive advantages of car-sharing in three travel scenarios [5]. Shen first clustered Shanghai metro stations according to the pattern of passenger flow over time, then investigated the travel time and surrounding land use characteristics for different types of stations, and found that the passenger flow pattern of metro stations is closely related to the station location conditions and its surrounding land use pattern [6]. Ziyi Wang took Open Street Map (OSM) road network data and POI data as the main data source, and proposed a method for accurate identification of urban functional areas based on kernel density estimation, functional area identification, mixed degree calculation and other technical means. This paper further identified urban single functional land, mixed and comprehensive functional land, and analyzed the distribution characteristics of the functional structure in the research area [7]. Xu Wei extracted four main factors related to rail transit stations by principal component analysis, and then used K-Means clustering method to classify the stations according to the extracted main factors, providing reference for the subsequent development of the area around the stations [8]. Starting from three latitudes of rail transit stations in urban space, rail line network and urban transportation network, Xue Xia used K-Means clustering algorithm to divide urban rail transit stations into five major categories, and summarized the universal laws of passenger flow of each type of stations based on the results of station clustering [9]. The above research focuses on two aspects: on the one hand, we seek the passenger flow variation pattern from the perspective of individual travel chains and explore the association rules and describe their travel characteristics; on the other hand, we summarize the relationship between the variation pattern of passenger flow and the spatial characteristics of metro stations after clustering at the macro level.

Based on the AFC swipe card data of Shanghai rail transit system for one consecutive week in April 2015, this paper explores the spatial association rules of stations and clustering analysis of metro stations combined with POI data of the same period to explore the possible reasons for the association rules among stations and reveal the spatial distribution pattern of rail transit passengers' travel. The research results can provide new ideas for rail transit line network planning, operation management, and inbound and outbound traffic control decisions.

1. Data Overview

The AFC data for this study runs from Monday, April 13, 2015 to Sunday, April 19, 2015 and covers 5 typical weekdays with 2 typical weekends. The raw data was collected from over 5.95 million anonymous card users with over 58.26 million transactions recorded. As of April 2015, Shanghai Metro has 14 lines, and each swipe data contains 7 attributes (as shown in Tab 1). The POI data contains information of all points of interest in Shanghai in 2015, divided into 14 categories totaling more than 1.2 million items, and each data records 19 attributes of the POI such as name, address, type, latitude and longitude under the geodetic coordinate system.

Tab 1. AFC Swipe Card Data Attributes

NO.	Listings	Description
1	Card Number	Card number, a unique value in the record
2	Swipe date	Swipe date
3	Swipe time	Swipe time
4	Site Information	Line name and station name
5	Transportation	Subway
6	Price	0 is inbound, non-0 is outbound
7	Whether the discount	Is there a discount for this trip

1.1 Data Cleaning

The ideal experimental results cannot be achieved without high-quality data, and cleaning the original data is the basis for the validity and accuracy of data mining analysis. The main task is to check whether there is "dirty data" in the original data, "dirty data" generally refers to the data that do not meet the requirements, as well as the data that cannot be directly analyzed accordingly. In common data mining work, dirty data include: missing values, abnormal values, inconsistent values, duplicate data and data containing special symbols.

(1) Fill in the missing values. Some data attributes are lost due to mechanical reasons such as failure of data collection equipment, failure of storage media, failure of transmission media, etc. Such data are filtered out and filled based on the same historical records.

(2) Find outlier values. The data in and out of the same station and the card swipe data during non-operating hours are both outliers, which may be generated by the metro staff to check the operation status of the equipment; for passengers holding Shanghai public transportation card or one-way metro ticket, the limit between the single entry and exit time of the metro is 4 hours, and the card swipe data for staying in the metro station for more than 4 hours or less than one station travel time (2 minutes) should also be considered as outliers. The data should also be considered as outliers.

(3) Consistency check. Data inconsistency refers to the contradiction and incompatibility of data, and direct mining of inconsistent data may produce mining results that are contrary to reality. This study mainly conducts consistency check on the overlapping parts of AFC data and POI data, such as station names.

(4) Delete duplicate values. Due to duplicate data records caused by data collection and other equipment due to unstable network signals or other reasons, only the first data information is retained if the values of each attribute are judged to be identical.

1.2 Data Pre-processing

Pudong International Airport, the terminal of Shanghai Metro Line 2, as one of the three major gateway complex hubs in China and the first hub airport in East China region, adopts an inter-day operation strategy to alleviate the huge daily passenger flow pressure. Meanwhile, Shanghai AFC original swipe card data is stored in days, and there is a situation that the inbound and outbound data of the same traveler is stored in two files separately. When performing passenger travel OD matching, such data may be mistakenly deleted, resulting in partial data loss for that time period, which has an impact on the mining results of subsequent association rules. To solve this problem, the card swipe data of weekdays and weekends are connected, and the date and time are sorted and then travel OD matching is performed according to the unique value of card number to ensure the integrity of the data, and the pre-processed travel OD chain is as shown in Tab 2.

Tab 2. Examples of Passenger Travel OD.

Card number	Inbound time	Inbound lines	Inbound stations	Outbound time	Outbound lines	Outbound stations
4200000141	2015-04-17 08:33:18	Line 11	Jiangsu Road	2015-04-17 09:01:21	Line 1	Huangpi South Road
4200000172	2015-04-14 11:44:29	Line 8	Jiangyue Road	2015-04-14 12:40:02	Line 2	Zhangjiang Hi-Tech
4200000189	2015-04-17 15:49:43	Line 2	Nanjing East Road	2015-04-17 16:14:27	Line 2	Weining Road

2. Algorithm Introduction

In this paper, the association rule algorithm is first used to study the spatial travel pattern of passengers, then the K-Means algorithm is used to cluster the urban rail transit stations, and the reasons for the association rule are further analyzed according to the clustering results.

2.1 Association Rule Algorithm

The concept of association rules was first introduced by Agrawal and other scholars in 1993 [10], and its purpose is to identify potential connections between large database items. It is a common research method in data mining.

2.1.1 Basic Concept of Association Rules

Let $I = \{I_1, I_2, ..., I_n\}$ be literal attributes called items. An itemset is a set of items, and an itemset containing *k* items is called a *k*-itemset. If the itemset $X \subseteq I$, $Y \subseteq I$ and $X \cap Y = \varnothing$, then the implication in the form $X \Rightarrow Y$ is called an association rule, where X is called is the pre-itemset of the rule and Y is the post-itemset of the rule, which indicates that the transaction containing the itemset X is also likely to contain the itemset Y.

2.1.2 Excavation Process

Association rule mining can be implemented in two steps:

① Given a transaction database and an itemset, finds the complete set of all frequent patterns in the database that are greater than or equal to the user-specified minimum support threshold (min_Sup). The frequency of occurrence of an itemset is the count of all transactions containing that set, also known as the absolute support or support count. The expression for calculating the support for an itemset X, which is defined as Eq.(1)[11].

$$Support(X) = \frac{n(X)}{N} \times 100\% \quad (1)$$

The formula n(X) is the number of transactions in the transaction dataset containing itemset X, and N is the total number of transactions in the database. The frequent itemsets satisfying Support(X) ≥ min_Sup condition is filtered out to reduce the generation of redundant association rules and improve the overall performance of the algorithm.

② An association rule X ⇒ Y is extracted from each frequent pattern of the complete set of frequent patterns and its confidence is calculated with the following formula (2)[11]. The rules that satisfy both Support(X⇒Y) ≥ min_Sup and Confidence(X ⇒ Y) ≥ min_Conf conditions are called (strong) association rules.

$$Confidence(X \Rightarrow Y) = \frac{Support\ (X \cup Y)}{Support\ (X)} \times 100\% \quad (2)$$

Support(X ∪ Y) is the support of association rule X⇒Y in the formula. The higher the confidence, the more likely the transaction Y is accompanied by the transaction X.

The Apriori algorithm is the original and more influential of the association rule mining algorithms [12]. The algorithm uses the anti-monotonic property to compress the search space, i.e., all non-empty subsets of frequent itemsets are frequent, in other words, if an itemset is not frequent, then all its supersets must also be infrequent (strong) association rules. The specific steps are as follows:

(1) Discovery of frequent item sets.

Step1. k = 1, scanning all transactions in the database;

Step2. Count the support of each (k-1)-term candidate set;

Step3. For every two frequent itemsets of length k with (k-1) common items are concatenated to obtain C'_k;

Step4. Prune C'_k to obtain C_k according to the inverse monotonicity;

Step5. Remove the itemsets with support lower than min_Sup to get the k-frequent itemset L_k .

(2) Generating association rules from frequent itemsets.

Step1. For each frequent itemset l, generate all non-empty subsets of l;

Step2. For each non-empty subset s of l, if Confidence(s⇒ l-s) ≥ min_Conf is satisfied, then output s ⇒ l-s.

If only two measures of interest, support and confidence, are used to mine association rules, it is easy to generate meaningless, redundant, or even misleading rules [13]. In order to make association rules achieve better results in practical applications, the quality of association rules is further improved by introducing the Lift metric to further improve the quality of association rules [14].

The lift of association rule X ⇒ Y reflects the correlation between the itemset X and the itemset Y. The calculation formula is shown in Eq.(3).

$$Lift\ (X \Rightarrow Y) = \frac{Support\ (X \cup Y)}{Support\ (X)\ Support\ (Y)} = \frac{Confidence\ (X \Rightarrow Y)}{Support\ (Y)} \quad (3)$$

Lift(X⇒Y) > 1 indicates that the occurrence of itemset X and itemset Y are positively correlated. Lift(X⇒Y) = 1 indicates that itemset X and itemset Y are independent of each other. Lift(X⇒Y) < 1 indicates that the occurrence of itemset X and itemset Y are negatively correlated.

2.2 K-Means Clustering Algorithm

The K-Means algorithm was first used by MAC Queen as a representative of unsupervised clustering algorithm [15], and its main function is to automatically group similar samples into a category. Compared with other algorithms, this algorithm

has the advantages of high robustness, easy to understand, and low computational cost.

The clustering process of the K-Means algorithm can be divided into four steps:

① Determine k initial clustering centroids C_i ($1 \le i \le k$);

② The Euclidean distance from each sample to C_i is calculated to find the centroid nearest to that sample and assign it to the cluster corresponding to the cluster center C_i. The Euclidean distance from the sample to the cluster centroid is calculated as Eq.(4).

$$dis(x,\ C_i) = \sqrt{\sum_{j=1}^{m} \left(x_j - C_{ij}\right)^2} \quad (4)$$

The formula x is the sample, C_i is the i-th cluster center, m is the dimension of the sample, x_j, C_{ij} denotes the j-th attribute value of the sample x and the cluster centroid C_i.

③ Calculate the mean of all samples in the k clusters as the k cluster centroids for the second iteration;

④ Repeat ② and ③ until convergence (the centroids no longer change or the specified number of iterations is reached), i.e., the error sum of squares no longer changes, and the clustering process ends. The error sum of squares SSE for the whole data set is calculated as Eq.(5)[16].

$$SSE = \sum_{i=1}^{k} \sum_{x \in C_i} |dis(x, C_i)|^2 \quad (5)$$

In the formula, k is the number of clusters. The size of SSE indicates the effectiveness clustering, and the smaller SSE indicates the better clustering effect.

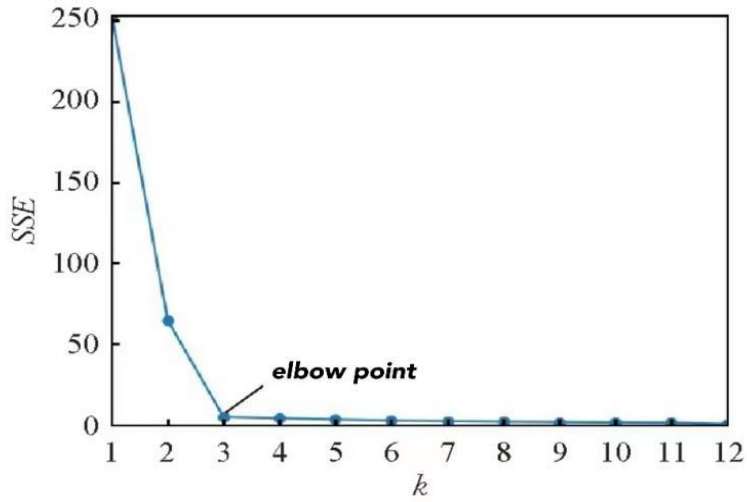

Fig 1. Relationship between SSE and k values

The key to the K-Means algorithm is the selection of k values, which is usually based on the trend graph of SSE values corresponding to different k values [16], namely “elbow method”, as shown in Fig 1. The SSE value corresponding to the k value before the "elbow point" decreases sharply with the increase of k, while the SSE value of k value after "elbow point" has no obvious change with the increase of k value. Therefore, the SSE value corresponding to "elbow point" is generally chosen as the optimal cluster number.

3. Example and Analysis

After pre-processing the AFC swipe data for one week, we finally obtained 22.58 million valid weekday swipe data and 6.3 million valid non-workday swipe data about 288 metro stations. Then the association rules between weekday and weekend stations were mined separately, and then the stations were clustered with POI data in the same period, and finally the reasons for the association rules were analyzed according to the clustering results.

3.1 Spatially Frequent Itemset Mining

By counting the average daily passenger flow in and out of each station on weekdays and weekends, a heat map of passenger flow distribution is drawn, as shown in Figs 2 and 3. The inbound and outbound stations of the weekdays overlap highly with each other and are mainly located at the junction of Huangpu, Jing'an, Hongkou, Xuhui, Changning and Putuo

districts. The distribution of inbound and outbound passenger flow on weekends is similar to that on weekdays, but the number of large passenger flow sites is reduced compared to weekdays.

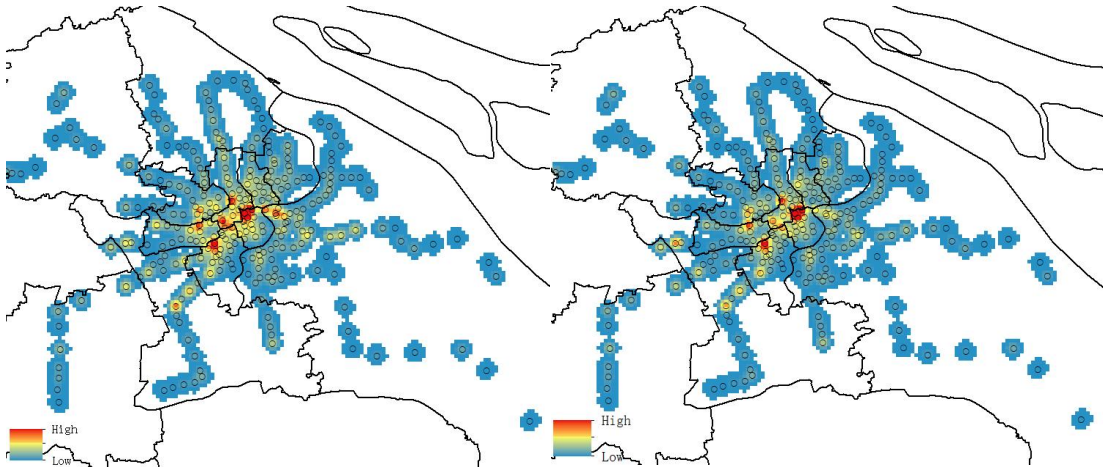

Fig 2. Heat map of daily average passenger flow to and from the station on weekdays

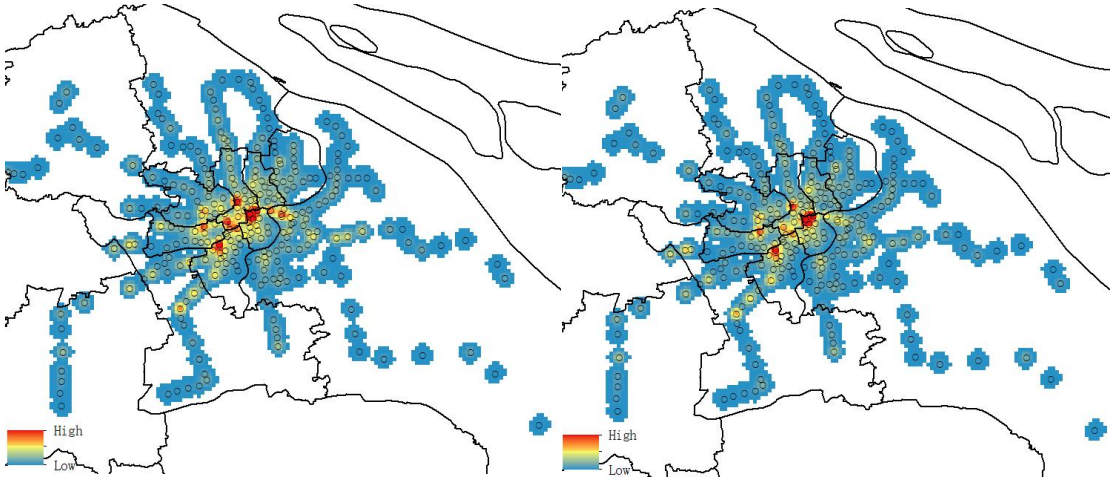

Fig 3. Heat map of the average daily inbound and outbound passenger flow on weekends

3.1.1 Weekday Inter-site Association Rule Mining

The 288 stations extracted from the swipe card data are used as the object of this paper, and the 1-itemset of inbound stations {O} and outbound stations {D} are analyzed first. The top ten itemsets of inbound stations {O} on weekdays are {People's Square}, {Xujiahui}, {Shanghai Railway Station}, {Jing'an Temple}, {Lujiazui}, {Zhongshan Park}, {Xinzhuang}, {East Nanjing Road}, {Shanxi South Road}, and {Lotus Road}, and the total support of these stations reaches 15.86%, in other words, 15.86% of the inbound passengers of Shanghai metro line network are concentrated in the above 10 stations around. The top ten concentrations of the collection of outbound stations {D} on weekdays are identical except for {Nanjing West Road} in the tenth place, which is different from {Lotus Road} in {O}, with a total support of 16.92%, that is, 16.92% of the rail travelers' activities are around these ten stations, and they bear a larger passenger flow than the inbound stations.

The Apriori algorithm is used to concatenate the 1-itemsets {O}, {D} to obtain 82,944 2-itemsets {O, D}. The top ten 2-frequent itemsets on weekdays are {Xujiahui, Xinzhuang}, {Tonghe New Village, Shanghai Railway Station}, {Jiuting, Caohejing Development Zone}, {Pengpu New Village, Shanghai Railway Station}, {Xinzhuang, Xujiahui}, {Xinzhuang, People's Square}, {People's Square, Xinzhuang}, {Lujiazui, Nanjing East Road}, {Caohejing Development Zone, Jiuting}, and {Nanjing East Road, Lujiazui}, with a sum of support is 0.74%. It is noteworthy that four of the top ten frequent itemsets {O, D} have a mirror image relationship, indicating that travelers tend to choose the same mode of transportation for both the return and onward journeys.

After determining the 1-frequent itemsets {O} and {D}, the 2-itemsets {O, D}, {D, O} are obtained by concatenation,

and the values of support, confidence and lift constitute different types of rules, and typical representatives are selected for display, as shown in Table 3 and Table 4.

The support and confidence of rule 1 in Table 3 meet the threshold of minimum support and minimum confidence, and the level of left is high, so this kind of association rule is the object of subsequent analysis. The support and left of rule 2 are high but the confidence level is low, which indicates that although the passenger flow between Shanghai Railway Station and Gongfu New Village Station is large, the accompanying relationship between the stations is not strong. Rules 3, 4, 7 and 8 have low lift levels, indicating that the positive correlation between the two stations is not significant and will not be discussed subsequently. Rule 5 and rule 6 have a general support level, but a high level of lift, indicating that although the passenger flow between the two stations is not large, the probability of the outbound station corresponding to the rule is significantly increased after the inbound station is determined.

Tab 3. {O}⇒{D} for Weekdays.

Number	Rules	Frequency	Support	Confidence	Lift
1	{Tonghe New Village} ⇒ {Shanghai Railway Station}	18034	0.08%	13.96%	6.43
2	{Shanghai Railway Station} ⇒ {Gongfu New Village}	11669	0.05%	2.98%	6.89
3	{Pengpu New Village} ⇒ {People's Square}	10790	0.05%	7.25%	2.93
4	{People's Square} ⇒ {Lujiazui}	9563	0.04%	1.78%	1.09
5	{Jinke Road} ⇒ {Tangzhen}	5200	0.02%	2.96%	9.37
6	{Tongji University} ⇒ {Wujiaochang}	4147	0.02%	6.02%	17.56
7	{Outside Ring Road} ⇒ {People's Square}	3844	0.02%	5.39%	2.17
8	{Yishan Road} ⇒ {Shanghai Railway Station}	3132	0.01%	1.41%	0.65

Taking Table 3 Rule 1 in Table 3 as an example for illustration, {Tonghe New Village}⇒{Shanghai Railway Station} indicates that when a traveler chooses Tonghe New Village as an inbound station, the probability of Shanghai Railway Station as an outbound station is 13.96%. Among the 22.58 million valid records, the travel chain with the inbound station as Tonghe New Village and the outbound station as Shanghai Railway Station appeared 18,034 times, accounting for 0.08% of the total number of trips. When the inbound station of a trip is determined to be Tonghe New Village, the probability that the outbound station is Shanghai Railway Station is 6.43 times higher than the probability that it is simply distributed in Shanghai Railway Station.

Tab 4. {D}⇒{O} for Weekdays

Number	Rules	Frequency	Support	Confidence	Lift
1	{Jiuting} ⇒ {Caohejing Development Zone}	17867	0.08%	10.94%	32.76
2	{Pengpu New Village} ⇒ {People's Square}	10790	0.05%	7.75%	2.30
3	{Xinzhuang} ⇒ {Dongchuan Road}	10353	0.05%	2.98%	8.97
4	{Xujiahui} ⇒ {Shanghai South Railway Station}	9312	0.04%	1.89%	1.35
5	{Hongqiao Terminal 2} ⇒ {Song Hong Road}	6307	0.03%	9.05%	12.98
6	{Qibao} ⇒ {Hechuan Road}	5840	0.03%	2.71%	9.04

7	{Xujiahui} ⇒ {Caoyang Road}	3535	0.02%	0.72%	0.82
8	{Xujing East} ⇒ {Lujiazui}	3166	0.01%	1.55%	1.31

The analysis of the rules shown in Tab 4 is similar to Tab 3.The inbound and outbound stations of Rule 3 in Tab 3 and Rule 2 in Tab 4 are mirror image of each other, and both rules have high support, high lift, and low confidence, which verifies the passenger flow relationship between the two stations.

3.1.2 Weekend Inter-site Association Rule Mining

The top ten itemsets of weekends inbound station collections {O} are {People's Square}, {Xujiahui}, {Shanghai Railway Station}, {Zhongshan Park}, {Nanjing East Road}, {Shanxi South Road}, {Hongqiao Railway Station}, {Jing'an Temple}, {Xinzhuang} and {Shanghai South Railway Station}, with a total support of 18.76%. Among them, the three large railway stations as transportation hubs are included, and the passenger flow of the line network is also more concentrated compared to weekdays. The top ten itemsets of weekends outbound station collection {D} are identical except for {Lujiazui} in the tenth position, which is different from {Hongqiao Railway Station} in {O}, with a total support of 19.0%. In general, the inbound and outbound station traffic is more stable on weekends than on weekdays.

The top ten 2-frequent itemsets for weekends are {Xujiahui, Xinzhuang}, {People's Square, Xinzhuang}, {Lujiazui, Nanjing East Road}, {Xinzhuang, People's Square}, {Xinzhuang, Xujiahui}, {Xujiahui, People's Square}, {Nanjing East Road, Lujiazui}, {People's Square, Lotus Road}, {Lotus Road, Xujiahui} and {Lotus Road, People's Square}, with a sum of support of 0.84%. It is easy to see that the top ten itemsets {O, D} are centered around the six stations of Xujiahui, People's Square, Xinzhuang, Lujiazui, Nanjing East Road and Lotus Road, and most of them also have mirror relationships. The typical association rules {O}⇒{D}, {D}⇒{O} that satisfy the minimum support and minimum confidence and have high lift are selected for display, as shown in Tab 5 and Tab 6.

Tab 5. Typical Association Rules {O}⇒{D} for Weekends

Number	Association Rules	Frequency	Support	Confidence	Lift
1	{Jiuting} ⇒ {Qibao}	4140	0.07%	8.58%	9.00
2	{Dongchuan Road} ⇒ {Xinzhuang}	3649	0.06%	17.46%	11.33
3	{Sheshan} ⇒ {Qibao}	3017	0.05%	10.07%	10.56
4	{Tongji University} ⇒ {Wujiaochang}	2206	0.04%	9.71%	20.21
5	{Fujin Road} ⇒ {Gongkang Road}	1726	0.03%	7.68%	16.30

Tab 6. Typical Association Rules {D}⇒{O} for Weekends.

Number	Association Rules	Frequency	Support	Confidence	Lift
1	{Dongchuan Road} ⇒ {Xinzhuang}	3649	0.06%	15.70%	10.42
2	{Sijing} ⇒ {Qibao}	2932	0.05%	11.01%	11.60
3	{Wujiaochang} ⇒ {Tongji University}	1875	0.03%	6.19%	17.17
4	{Fujin Road} ⇒ {Gongkang Road}	1726	0.03%	8.44%	18.30
5	{Gucun Park} ⇒ {Shanghai University Station}	1563	0.02%	5.94%	15.57

Rule 2 in Table 5 and Rule 1 in Table 6 are mirror image of each other, and the support, confidence and lift of these two association rules are all high, indicating that there is a certain scale of passenger flow active between Dongchuan Road Station and Xinzhuang Station.

3.2 Classification of metro station types

Based on the current situation of POI data in Shanghai in 2015, combined with the clear provisions on the rail influence zone in the "Guidelines for Planning and Designing Areas along Urban Railways" released by the Ministry of Housing and Urban-Rural Development in 2015 [17]. In this paper, the POI data within the 500m radial range around the rail stations are selected for K-Means clustering of metro stations. After standardization and normalization of POI data within the radius of the metro station, the "elbow diagram" is drawn to determine the value of *k*. As shown in Fig 4 , the metro stations in

Shanghai are finally divided into four categories.

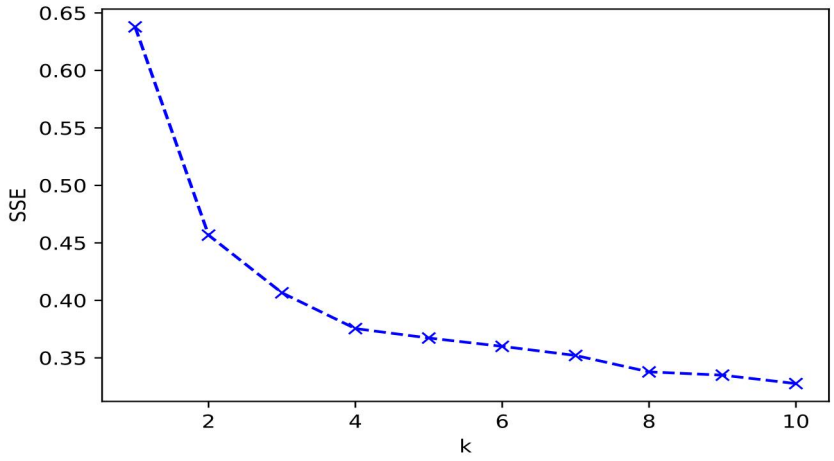

Fig 4. Correspondence plot between k and treated SSE values

The results of metro station clustering are shown in Tab 7, and the distribution of various types of stations in space is characterized by "circle", as shown in Fig 5.

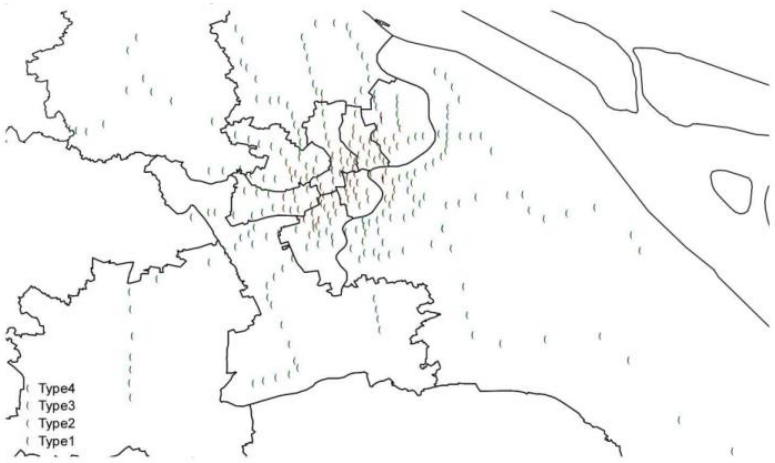

Fig 5. Spatial distribution of clustered sites

Tab 7. Stations Clustering Results

Station Type	Site Features	POI Characteristics within the Radius	Typical Stations	Number/pc
1	Concentrated in the downtown area, the overall distribution is inside the loop of Line 4, with more than half the number of interchange stations.	The average number of POIs within the radius of the station is 4229, and the top 3 POI categories are shopping services, catering services and corporate enterprises.	Xujiahui Station, People's Square Station	23
2	The stations are distributed on the No. 4 loop of the subway and its two sides, and are generally concentrated within the middle ring road, with 40% of the interchange stations.	The average number of POIs within the station's radius is 1940, and the top 3 POI categories are shopping services, corporate businesses and food and beverage services.	Jinshajiang Road Station, Shanghai Gymnasium Station	61

3	Starting from the No. 4 loop road and dispersing outward along each subway line, it is concentrated between the inner ring elevated road and the outer ring highway.	The average number of POIs within the radius of the station is 989, and the top 3 POI categories are shopping service, food service and living service.	Qibao Station, Songhong Road Station	80
4	It is mainly located at the beginning and end stations of each metro line, at the outermost part of the subway network.	The average number of POIs within the radius of the station is 283, and the top 3 POI categories are shopping service, food service, life service and company enterprise.	Sanlin Station, Sailong Road Station	124

The clustered stations are further classified into residential stations, employment stations, transportation stations and comprehensive stations according to the proportion of various types of POI data within the radiation range of each subway station [6], the reasons for the generation of large passenger flow clusters are described.

① Residential stations: They are mainly oriented to residential areas, serving the surrounding residents. The number of business-residential POIs within the radius is overall significantly larger than the number of company-enterprise POIs, and the number of residence-related types of living POIs, such as catering and shopping, and living services, accounts for a relatively high number.

② Employment-oriented stations: There are a large number of commuters, and they are usually located near industrial or commercial areas. A high percentage of companies and businesses within the radius, followed by the surrounding roads and transportation facilities services, and well-developed transportation provide convenience for the employed people.

③ Transit-oriented stations: Built to facilitate commuting, long-distance travel, or interchange between different modes of transportation, they are usually located near the center of the city or the start of the subway network. A high proportion of transportation-related POIs, such as road and transportation facility services, are located within the radius of the station, and a small number of companies are also located there, attracting travelers for employment or interchange. Transportation stations are subdivided into traditional and non-traditional categories. Traditional transportation stations serve as urban hubs, undertaking long-distance passenger transportation between two cities, while non-traditional transportation stations mainly undertake close-distance passenger transportation between the central city and the suburbs and major neighboring cities.

④ Comprehensive stations: They usually gather a variety of functions, such as dining and shopping, entertainment and leisure, medical and education, etc., aiming to provide more comprehensive and convenient services for passengers. There are a wide variety of POIs within the radiation area, but they are more balanced in terms of quantity distribution.

3.3 Analysis of Association Rules

According to the clustering results of metro stations in Section 3.2 to count the travel categories of high lift association rules, a total of 15 travel categories are obtained, except for {1, 1} category trips which do not exist. The travel categories of high lift association rules for working days and non-working days have great similarity, and the comparison between them is shown in Table 8.

Tab 8. Summary Comparison of Weekdays and Weekends OD pairs After Station Clustering.

Travel Category Number	Weekdays			Weekends		
	{O}⇒{D} Station Travel Category	Number of Statistics/Group	Average Travel Distance/Stop	{O}⇒{D} Station Travel Category	Number of Statistics/ Group	Average Travel Distance/ Stop

1	{3, 3}	24	3.36	{3, 4}	25	5.80
2	{3, 4}	22	5.14	{3, 3}	24	3.58
3	{3, 2}	21	4.29	{4, 3}	22	5.23
4	{4, 3}	19	5.42	{3, 2}	11	4.64
5	{2, 3}	15	4.20	{3, 1}	11	3.91
6	{3, 1}	12	5.50	{2, 3}	10	4.20
7	{2, 2}	9	4.11	{4, 4}	10	4.20
8	{1, 3}	8	4.75	{1, 3}	10	2.90
9	{4, 4}	7	2.43	{2, 2}	9	3.00
10	{4, 2}	4	6.75	{2, 4}	7	6.29
11	{2, 4}	3	8.00	{4, 2}	5	6.80
12	{4, 1}	3	9.33	{4, 1}	2	9.00
13	{1, 4}	2	9.00	{1, 4}	2	9.00
14	{2, 1}	1	5.00	{2, 1}	2	3.50
15	{1, 2}	1	1.00	{1, 2}	2	3.50

As can be seen from the Tab 8, the distribution of trips between weekdays and weekends categories is relatively stable, and the mirror relationship between travel categories is obvious, echoing the previous association rule mining. NO.1 to NO.4 categories of trips account for more than 50% of the 15 categories of trips, and they are all related to the stations of type 3; their distribution is mainly centered on Pengpu New Village Station of Line 1, Jinke Road Station and Songhong Road Station of Line 2, and Qibao Station of Line 9, with its surrounding stations constitute a large passenger flow group, as shown in Figure 6.

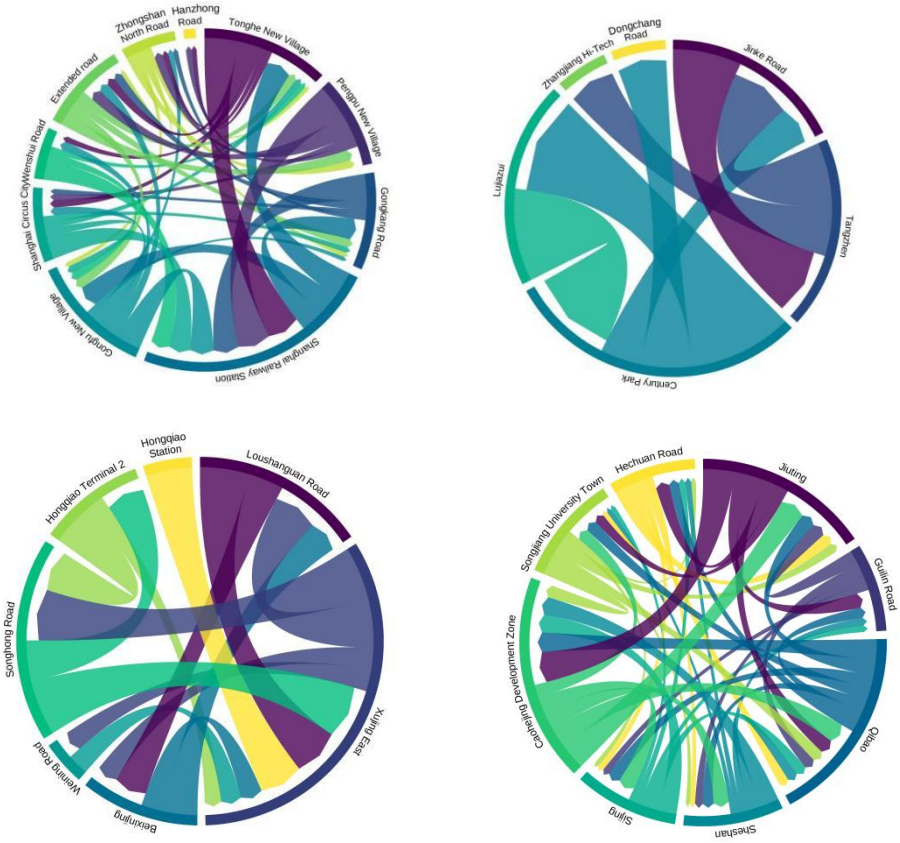

Fig 6. Chord diagram of passenger flow in high traffic groups

Pengpu New Village, as a comprehensive station, was the best developed workers' new village in Shanghai in the early

days. In addition to the commuter traffic attracted by corporate enterprises, services such as catering and shopping, medical science and education also attract productive and living traffic from other areas; at the same time, residents of business residential areas are also employed at the surrounding stations. Wenshui Road and Gongfu New Village in the cluster are built right in the middle of the overpass, and as employment-oriented stations, their convenient interchange services attract a large number of passengers to and from the station; comprehensive stations such as Shanghai Circus City and Gongkang Road Station provide certain employment and residential options; residential stations such as Tonghe New Village and Pengpu New Village are mostly commuting for the purpose of traveling. Shanghai Railway Station, as a traditional transportation station, naturally carries the passenger flow from all over the world.

Jinke Road and Lujiazui are typical employment-oriented stations, and the types of POIs within the radial area are mostly catering and shopping services and corporate enterprises, which attract people near the surrounding residential stations, such as Century Park and Tangzhen, which mostly go to the vicinity of the station for employment. It forms a commuter flow between Century Park and Lujiazui, Tangzhen and Jinke Road stations.

The unique geographical location of Songhong Road makes it a non-traditional transportation station. As the only channel connecting Line 2 to the vehicle section of Beizhai Road, Songhong Road carries a large number of passengers from the western region of Shanghai's central city to the suburban region and distant satellite cities as well as major surrounding cities; secondly, Hongqiao Terminal 2 adjacent to Songhong Road, as a traditional transportation hub station, carries passengers leaving and arriving at Shanghai on a daily basis. For these reasons, a large passenger flow group centered on this station has been formed around Songhong Road.

Within the radius of Qibao, there are many business residences and living services related to the daily activities of residents, which is a typical residential station. A large number of restaurants, shopping, sports and leisure attracts people from the neighboring stations for employment and entertainment, and the supporting scientific, educational, cultural and health care services increase the flow of passengers to and from school and medical care. In addition, Qibao Ancient Town, as a scenic spot, also attracts a part of tourists to come. Guilin Road, Caohejing Development Zone and Hechuan Road in Qibao Group are employment-oriented stations, while Jiuting, Sijing and Sheshan are typical residential stations, and there are a large amount of commuter flows due to the separation of employment and residence in this group.

NO.5 to NO.9 category of travel accounts for about 30%, basically in the above four large passenger flow group on the basis of the outer layer of the spread, eventually forming a larger group from Congfu New Village to Shanghai Railway Station, from Tangzhen to Lujiazui, from Loushanguan Road to Xujing East, from Songjiang University City to Guilin, as well as smaller groups under the same category. The 10 to 15 categories of trips only accounts for about 20%, although the category is the largest, but the travel distance is farther, and the cost of time consumed is higher, making part of the travelers give up rail transit to other means of transportation.

Overall, the average weekdays travel distance is slightly higher than the average weekends travel distance, with rigid travel in the commuting category dominating. Weekend trips are less time sensitive, so the distance traveled category increases.

4. Conclusion

Through association rule mining of Shanghai metro swipe card data and POI data of the same year, the following conclusions are obtained:

① The top ten 1-frequent itemset of inbound and outbound stations on weekdays have great similarity, and the outbound passenger flow is more concentrated than the inbound passenger flow; most of the 2-frequent itemsets have mirror relationships, indicating that travelers tend to choose the same type of transportation for both the inbound and outbound journeys. The spatial travel patterns of weekends are consistent with those of weekdays, but the inbound and outbound passenger flows of weekends are more concentrated compared with those of weekdays.

② The 288 stations are divided into four categories according to the type and number of POIs within the attraction area of each station, and the clustered stations show a "circle" distribution in space. The clustering areas with high lift association rules are all related to type 3 stations, forming four large passenger clusters, and the traffic within the clusters is

mainly commuter traffic between residential stations and employment stations.

③ The association rule analysis of Shanghai metro swipe card data and POI data by using data mining technology can effectively discover the hidden patterns among them. With the support of actual data to reveal the causes of frequent OD pairs and avoid the subjectivity of data acquisition, the research conclusions fit the current situation of urban rail transportation characteristics in China, and the research results have certain credibility.

There are still some shortcomings in this study, for example, the article only studied the association rules of the spatial dimension, and the subsequent research on the temporal dimension can be added; different types of POI data have different attraction strengths for travelers, and different weights can be applied to reflect the attractiveness of POI; as some types of POI data only present the number, but do not contain demographic information, the more intuitive AOI data can subsequently be selected as a substitute.

References

[1] Xu ZZ. Current situation and development trend of urban public transportation characteristics in China[J]. People's Public Transport, 2022, No.154(10):10-19.

[2] Liang Q, Weng JC, Zhou W, Rong J. Crowd identification of public transportation commuting stability based on association rules[J]. Journal of Jilin University (Engineering Edition), 2019, 49(05): 1484-1491.

[3] Guo X, David Z.W. Wang, Wu JJ, Sun HJ, Zhou L. Mining commuting behavior of urban rail transit network by using association rules[J]. Physica A: Statistical Mechanics and its Applications, 2020,559.

[4] Chu F. Research on the characteristics of Beijing rail transportation based on association rule mining[D]. Kunming University of Science and Technology, 2020.

[5] Xu Y, Ji X, Jin Z. What travel scenarios are the opportunities of car sharing? [J]. Plos one, 2021, 16(12): e0260605.

[6] Shen P, Ouyang L, Wang C, et al. Cluster and characteristic analysis of Shanghai metro stations based on metro card and land-use data[J]. Geo-spatial Information Science, 2020, 23(4): 352-361.

[7] Wang Z, Ma D, Sun D, et al. Identification and analysis of urban functional area in Hangzhou based on OSM and POI data[J]. PLoS one, 2021, 16(5): e0251988.

[8] Xu W, Zheng CJ, Ma GH. et al. Research on classification of urban rail transit stations based on k-means clustering[J]. Journal of Guizhou University (Natural Science Edition), 2018, 35(06): 106-111.

[9] Xia X, Gai JY. Classification and passenger flow characteristics analysis of urban rail transit stations based on K-Means clustering algorithm[J]. Modern Urban Rail Transit, 2021(04):112-118.

[10] Agrawal R, Imielinski T, Swami A. Mining association rules between sets of items in large databases[J].ACM SIGMOD Record, 1993, 22(2): 207-216.

[11] Cai WJ, Zhang XH, Zhu JQ. et al. A review of association rule mining[J]. Computer Engineering,2001(05):31-33+49.

[12] Agrawal R, Srikant R. Fast algorithms for mining association rules[C]//Proc. 20th int. conf. very large data bases, VLDB. 1994, 1215: 487-499.

[13] He Z, Huang HK, Tian SF. A discovery method for optimizing correlation rules[J]. Journal of Computer Science, 2006(06): 906-913.

[14] Brin S, Motwani R, Silverstein C. Beyond market baskets: generalizing association rules to correlations[C]// Proceedings of the 1997 ACM SIGMOD international conference on Management of data. 1997: 265-276.

[15] MAC Queen J. Some methods for classification and analysis of multivariate observations [C]// Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability. Berkeley: University of California Press, 1967:281-297.

[16] Yang JB, Zhao C. A review of research on K-Means clustering algorithm[J]. Computer Engineering and Applications,2019,55(23):7-14+63.

[17] Ministry of Housing and Urban-Rural Development of the People's Republic of China. Letter of Construction Regulations [2015] 276 Notice of the Ministry of Housing and Urban-Rural Development on the Issuance of Planning and Design Guidelines for Areas Along Urban Railways [Z]. 2015.

About authors:

Yujie Yang(August 1999), female, Han Nationality, Luohe, Henan province, master's student, Xihua University, Chengdu city, Sichuan Province, 610039, research direction: Urban rail transit, travel behavior.

Hui Li (August 1976), male, Han Nationality, Qiyang, Hunan Province, associate professor, master, Xihua University, Chengdu city, Sichuan Province, 610039, research direction: Transportation planning.

Qingsong Du(December 1998), male, Han Nationality, Guangyuan, Sichuan province, master's student, Xihua University, Chengdu city, Sichuan Province, 610039, research direction: Traffic Information Engineering and Control.

Zhenbo Liu (July 1998), male, Han Nationality, Dazhou, Sichuan province, master's student, Xihua University, Chengdu city, Sichuan Province, 610039, research direction: Traffic Information Engineering and Control.

Zihao Feng(September 1998), male, Han Nationality, Chengdu, Sichuan Province, master's student, Xihua University, Chengdu city, Sichuan Province, 610039, research direction: Transportation and Logistics optimization.