Analysis of subway passenger flow based on smart card data

Yi Wang[1], Weilin Zhang[2], Fan Zhang[2], Ling Yin[2], Jun Zhang[2], Chen Tian[3], Wei Jiang[4]
[1] Post Big Data Technology and Application Engineering Research Center of Jiangsu Province, Nanjing University of Posts and Telecommunications, China
[2]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences,China
[3]Nanjing University, China
[4]Nanjing Sino-German Institute of Digital Twin Smart Cities Co., Ltd.

***Abstract*—In recent years, local governments have paid more and more attention to the construction of urban rail transit represented by subways. The subway can greatly improve the ground traffic congestion in the city and the urban traffic environment, making people's travel more convenient. The analysis and accurate prediction of subway passenger flow has always been one of the key tasks of urban rail transit management departments. Especially in the context of the rapid growth of rail transit capacity and the changing needs of passengers, the need for passenger flow analysis is even more significant. This article completes the analysis of Shenzhen's subway passenger flow through big data analysis. It is of great significance to solve urban congestion, optimize the traffic network, and protect public transportation.**

Keywords—subway passenger flow, smart card data, Travel characteristics

I. Introduction

The vigorous construction of urban rail transit can greatly improve the ground traffic congestion in the city and the urban traffic environment, and can promote the development of the city, making people's travel more convenient. Therefore, in recent years, local governments have increasingly paid attention to the construction of urban rail transit. As a representative of rail transit, the subway has the advantages of low failure rate, large transportation volume, stability and safety, and is an important means to meet the basic travel needs of the public. It is irreplaceable. By establishing a complete transportation network, the dilemma of increasing passenger flow and mismatching public transportation capacity due to the development of the city can be improved [1]. For example, in 2020, there will be 17 subway lines (including extension lines) under construction in Shenzhen. This is only a part of the national rail transit construction. From the national perspective, from the statistics of the expected construction of rail transit in each city, it is estimated that by 2020, the number of new urban rail mileage in the country will exceed 5,500 kilometers. In addition to the construction of subways in first-tier cities, the rise of second- and third-tier cities will also contribute to the increase of urban rail transportation.

In recent years, with the maturity of sensing technology and computing environment, various kinds of data related to urban traffic have emerged quietly, such as GPS traffic flow, smart card data, mobile phone data, and so on. Using data mining, artificial intelligence and other technologies to analyze people's travel patterns is of great significance for effectively improving public transport services. In the research based on smart card data, Mariko et al. analyzed the frequency and consistency of daily travel patterns of passengers[3]. Mousumi et al. use smart card data to estimate passenger transfer rates, average trips, etc[4]. Chu et al. based on smart card swiping records to simulate bus passengers' journeys and their spatiotemporal characteristics[5]. Bruno et al. Analyzed the space-time regularity of passengers and daily travel patterns, and they found that passenger travel patterns are related to card types[6]. Ma et al. based on the similarity of trips, the regularity of individual passenger travel is analyzed[7]. Sun et al. look for "familiar strangers" based on the similarity of passenger travel time-space patterns[8]. Ceapa et al.Estimate the degree of congestion at different stations in the transportation network[9]. [30] Morency et al. evaluated the variability of passenger travel behavior on different days of the week [10]. Ticiana L et al. studies have shown that people's travel patterns are largely spatiotemporal and predictable[11].

Compared with the above research results, this paper completes the calculation of Shenzhen's subway passenger flow based on passenger flow related data such as smart card data, and analyzes the daily characteristics of subway passenger flow from the time and space level. As subway passengers have their own characteristics, this research has a good reference value for rail companies to effectively understand customers and formulate related strategies to improve service.

II. Studying Case

In 2017, Shenzhen ’s daily average passenger flow reached 8.88 million passengers / day, ranking fourth in the country, behind Beijing, Shanghai, and Guangzhou. Rail transit carried most of the passenger flow, and the average daily passenger flow reached 6.62 million passengers / day, accounting for about 74% of the total passenger flow. The daily average daily passenger flow during the morning peak period reached 1.25 million passengers / day, and the average daily passenger flow during the morning peak hours of the rail transit reached 940,000 passengers / day, accounting for about 75% of the total passenger flow; The average daily passenger flow of the night peak of rail transit reached 860,000 passengers / day, accounting for about 77% of the total passenger flow. It can be seen that research on rail transit is extremely important for public transportation throughout the city.

978-1-7281-8275-9/20/$31.00 ©2020 IEEE
DOI 10.1109/BigCom51056.2020.00034

As of December 31, 2017, the Shenzhen subway has a total of 8 lines and 168 subway stations (the transfer stations are not superimposed). The operating mileage of rail transit reaches 286 kilometers, the length of rail transit ranks sixth in the country, and there are 27 interchange stations located in Nanshan District, Baoan District, Futian District, Luohu District, Longgang District, and Longhua New District. The subway lines are shown in Fig.1.

The data used in this article is the ticket data collected by Shenzhen in 2017 through the AFC system (Automatic Fare Collection System). As with the smart card ticket data in most cities, these records contain information about passengers entering and leaving stations. Using this information, we can dig out the changes in passenger trajectories, the time and space passenger flow of the stations, and passengers. Travel characteristics and more.

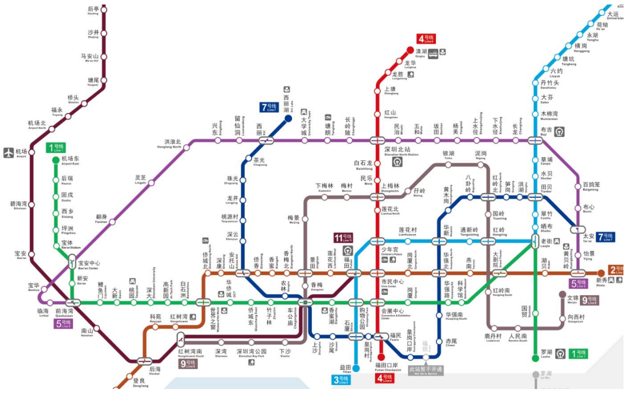

Fig. 1. Map of Shenzhen subway Lines

TABLE I. FIELDS OF SMART CARD RAW TRANSACTION DATA

	Field	Format
1	Record code	int
2	Card logical code	Long
3	Swipe terminal code	Long
4	Company code	String
5	Transaction Type	String
6	Transaction amount	Double
7	Card balance	Double
8	Swipe time	String
9	Success sign	String
10	Company (line)	String
11	Line (station)	String
12	gate number	String

As shown in Table I, it is the introduction of the original transaction data fields of the smart card. It can be seen that the ticket data we collected also includes consumption data, etc., because in many cities, the function of smart IC cards not only includes bus or subway, but also can be used to spend at designated places. Shenzhen is exactly in this way.

In the later processing, we also need to rely on the static information table of the subway station to make the OD connection. The fields are shown in TABLE II.

TABLE II. INFORMATION OF SUBWAY STATION

	Field	Format
1	Device terminal code	int
2	station	String
3	Station’s line number	String
4	Line name	String
5	longitude	Double
6	latitude	Double

III. SPATIO-TEMPORAL ANALYSIS OF PASSENGER FLOW

A. Solution Framework

We designed a system. First, through the data of ticket card and mobile terminal characteristics, the Spark computing framework based on Hadoop platform was used to calculate the subway line passenger flow, station inbound and outbound passenger flow, and regional passenger flow. The statistics and analysis of passenger flow characteristics were completed. In the following, a deep neural network will be established through the TensorFlow framework to predict the short-term passenger flow of lines, stations and areas within the station, and compare with the prediction results of other commonly used short-time passenger flow prediction algorithms. The framework structure is divided into the following sections:

1) Data layer

The data layer contains the main data sets used in this article, including subway ticket data, station static information, train schedules, maps and holidays data for a city, and mobile terminal characteristic data will also be collected.

2) Platform layer

All passenger flow calculations in this paper rely on big data platforms, so as to ultimately implement a passenger flow analysis and prediction system based on big data combined with deep learning. Because our traditional computing methods cannot meet the current computing needs, we need to use cloud-based big data platforms for computing. For the calculation of historical passenger flow, the Spark computing framework of the Hadoop platform is mainly used, and because the Spark computing framework is based on elastic distributed data sets, it has natural advantages for processing machine learning algorithms that need to continuously iterate and update parameters, so here The Spark framework is used; the TensorFlow framework is used for subsequent deep learning-based prediction algorithms.

3) Algorithm layer

The algorithm layer mainly contains various algorithms. We have designed a clearing algorithm for the calculation of passenger flow on the line. An algorithm based on multi-source data fusion is used for passenger flow calculation in a certain area of the station. At the same time, for the subsequent passenger flow prediction, the system also includes a short-term passenger flow prediction algorithm based on deep learning.

4) Application layer

The application layer contains applications that can be generated based on the above framework and algorithms. This part can include the statistical analysis of passenger flow, the calculation and prediction of passenger flow at the station, the calculation and prediction of passenger flow at the line, and the calculation and prediction of passenger flow at the station area.

B. Data cleaning and preprocessing

As mentioned earlier, the functions of smart IC cards in many cities include not only swiping buses or taking subways, but also consumption at designated places, which makes the smart card transaction data we collect also include consumption data. Therefore, before using the smart card transaction data to calculate passenger flow, it is necessary to perform data cleaning and preprocessing on it first.

The following is the pre-processing algorithm we used, which mainly includes formatting the time of smart card transaction data, extraction of main fields, separation of bus and subway and other data, and restoration of subway stations.

```
Transaction data preprocessing algorithm

1. Enter smart card transaction data for two consecutive days
2. Retain data from 03:00 on the first day to 03:00 on the second day
3. Differentiate bus credit card and subway credit card according to transaction type
4. Take out the bus card and store it in useful fields
5. Read in the subway terminal credit card device code and the station correspondence table, and combine it with the subway data to complete the data items with incomplete information about subway data in and out of the station
6. Take out subway card swipe data and store it in useful fields
```

After the above data preprocessing algorithm, the separated subway data format fields are shown in TABLE III:

TABLE III. DATA FIELDS OF SMART CARD SUBWAY CARD SWIPE

	Field	Format
1	Record code	int
2	Record code	Long
3	Terminal code	Long
4	Transaction Type	String
5	Transaction Time	String
6	Line name	String
7	Station name	String
8	Gate number	String

After the data is cleaned, the subway card data is obtained by filtering, and the data is divided into days, the abnormal data is filtered, and some field completion and field formatting are performed.

C. Calculation of subway OD

Subway OD refers to the travel records of passengers within a specific time range from the origin station to the destination station. It can reflect dynamic characteristics such as real-time changes or historical cross-section passenger flow and spatial distribution changes. Making OD connections is an important process for analyzing rail transit data. Since the subway is a relatively closed environment, generally speaking, as long as passengers swipe their cards at one station to enter the station, they will inevitably swipe their cards out of the station again. In this way, an OD pair necessarily corresponds to a complete travel of the passenger, so here the OD pair matching and calculation are performed on the smart card subway swipe data after cleaning. The following is an introduction to the subway OD connection algorithm.

```
Subway OD algorithm Error! Reference source not found.

1. Sort the swipe records of subway passengers on the day by time
2. Combine two adjacent records into one record
3. Keep the first two records as inbound and outbound records
4. Keep data with time difference between inbound and outbound less than 2.5 hours
5. Keep different records of inbound and outbound sites
6. Save the data
```

The format of the connected OD data is as follows:

TABLE IV. SUBWAY OD FIELDS AFTER CONNECTION

	Field	Format
1	Card code	int
2	Transaction Time	String

3	Line Name	String
4	Gate Number	String
5	In/Out Flag	int
6	Station ID	String
7	Station Name	String
8	Station Number	String
9	Longitude	Double
10	Latitude	Double

After the OD connection algorithm is used to connect the subway card data with the OD connection algorithm, it can calculate the dynamic indicators such as the spatial distribution of passengers, travel rules, and cross-section passenger flow.

IV. EXPERIMENTS AND RESULTS

A. Analysis of time distribution characteristics

From the perspective of the passenger flow distribution of the subway in the week, the passenger flow of the subway shows a clear law with the change of working days and weekends during the week. Judging from the credit card data of a city subway from September 17, 2018 to September 23, 2018, for Shenzhen North Station, more people will travel to surrounding cities on weekends, so weekends (2018/09 / 22 and 2018/09/23) significantly more foot traffic than weekdays (2018/09 / 17-2018 / 09/21). At the same time, for the Shenzhen University Station, which is close to the Science and Technology Park, there are many high-tech industry companies, so the passenger flow during the working day is obviously more than during the weekend.

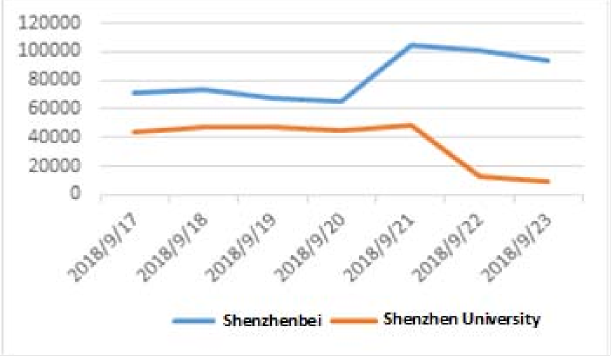

Fig. 2. Time distribution of passenger flow at different stations

The subway's time-shared passenger flow also has a very obvious law. There are significant morning and evening peaks on weekdays, with morning peaks concentrated between 7: 00-9: 00 and evening peaks concentrated between 17: 00-20: 00. During the peak hours, the hourly passenger flow of the Shenzhen University subway Station exceeds 5,000 passengers / hour, and the peak passenger hours do not exceed 2,000 passengers / hour. The double-peak phenomenon on weekends disappears, and the passenger flow in each period is similar to that in working days, and the overall passenger flow is less than that in working days, as shown in the figure below:

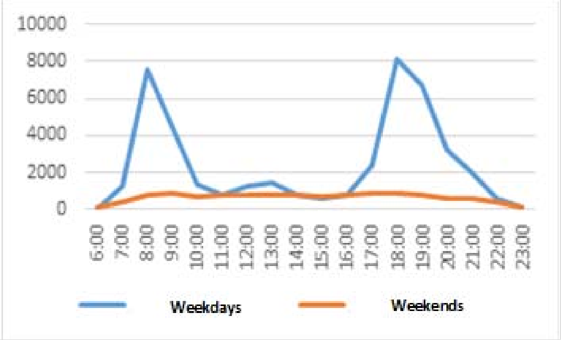

Fig. 3. Passenger flow of Shenzhen University Station on weekdays & weekends

B. Analysis of space distribution characteristics

By analyzing the regional passenger flow in a certain city, the passenger flow law at subway stations basically corresponds to the actual regional passenger flow distribution. The figure below shows the top ten daily average passenger traffic of a station in a city line network. Shenzhen North Station is not only a hub connecting a city to other cities, but also a transfer station for subway lines 4 and 5. So whether it is work Days and holidays, the passenger volume is very large; Chegong Temple is the subway station of Line 1, 7, 9 and 11, so the passenger flow can be ranked second; around the Grand Theater Station There are important landmarks in a certain city, such as Diwang Building and Ping'an Building. It is a financial center of a city with a large passenger flow and ranks third. Laojie Station is adjacent to major commercial areas due to the east gate “shopping paradise” Square, ranked fourth.

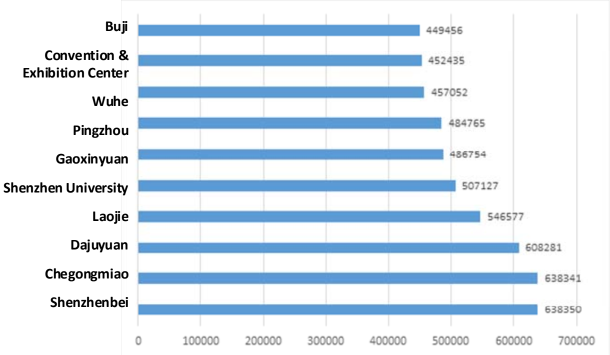

Fig. 4. Daily average passenger flow of different stations

The picture below shows that a certain city is the top ten in terms of OD passenger flow. Because Pengzhou Station is located in Baoan District of a city, the house price is not high, and Shenzhen University Station is a gathering place for many technology companies. The transit time is relatively short, so many office workers choose to live in Pengzhou and work in Shenzhen, so the daily average OD passenger flow is from Pengzhou to Shenzhen. Similarly, due to the "tidal phenomenon" of passenger flow, the OD passenger flow from Shenzhen University Station to Pengzhou Station ranks second.

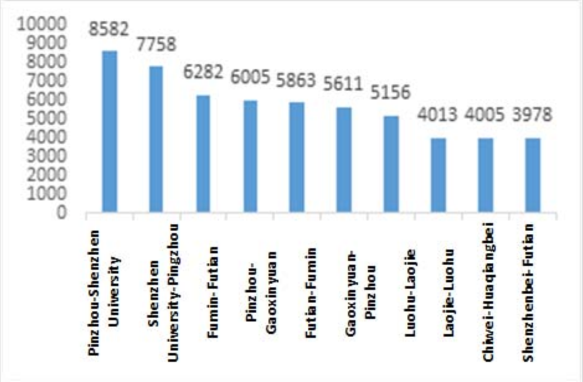

Fig. 5. Daily average passenger flow top10

In order to analyze the regional characteristics of the passenger flow of rail transit in a certain city, a city is divided into 10 administrative regions. The following figure is the distribution of the average daily departure and arrival passenger flow of the administrative region. It can be seen from the figure that from the perspective of the whole day, Futian District, as the center of the city, has the highest number of departures and arrivals in all administrative regions, with more than 800,000 passengers. From the perspective of the amount of departure, the second is Longgang District. Due to the relatively low housing prices in Longgang and the proximity to the city center, most citizens choose to live in Longgang and work in the city center. Since Dapeng, Pingshan, and Guangming New District are far from the city center, both the amount of departure and the amount of arrival are very low.

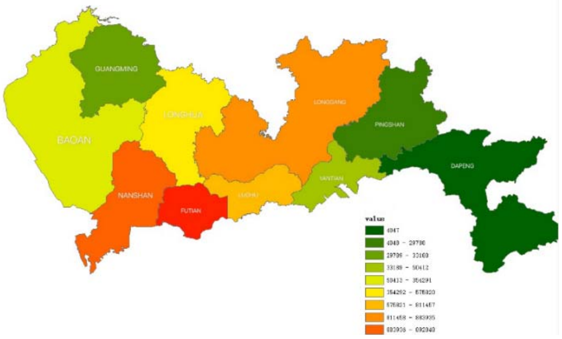

Fig. 6. Distribution map of departures by day

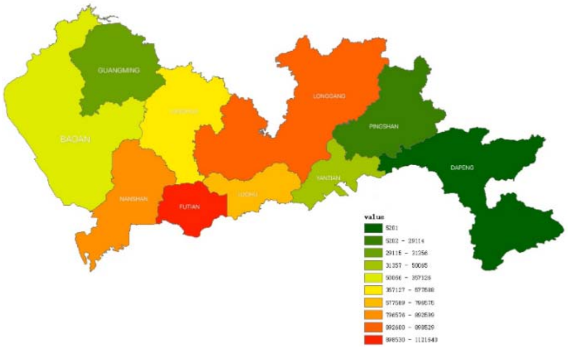

Fig. 7. Distribution map of all-day arrivals by administrative district

V. CONCLUSION

This paper collects the ticket data of Shenzhen Automatic Fare Collection System, and uses the information of passengers entering and leaving the station to dig out the trajectory changes of passengers, and then analyzes the time and space passenger flow rules of subway stations, and the characteristics of passengers' travel. The mining of this information is of great significance for solving urban traffic congestion, optimizing the existing traffic network, further planning new traffic routes, and protecting the safety of urban public transportation. Practical significance.

VI. ACKNOWLEDGMENT

The authors would like to thank anonymous reviewers for their valuable comments.

This research is supported by the National Natural Science Foundation of China under Grant Numbers 61672299, and 61972208.

REFERENCES

[1] Huang Zirong. Study on Passenger Flow Forecast Method of Rail Transit Network [D].

[2] Zou Dong, Liu Qiong, Huang Zirong. Prediction of Spatiotemporal Correlation Passenger Flows in Urban Rail Transit Networks [J]. Urban Rail Transit Research, 2016, 19 (3): 32-37.

[3] Mariko Utsunomiya, John Attanucci, and Nigel H M Wilson. Potential uses of transit smart card registration and transaction data to improve transit planning. Transportation Research Record, 2006.

[4] Mousumi Bagchi and PR White. The potential of public transport smart card. Transport Policy, 12(5):464-474, 2005.

[5] Ka Kee Alfred Chu, R Chapleau, and Martin Trepanier. Driver-assisted bus interview: Passive transit travel survey with smart card automatic fare collection system and applications. Transportation Research Record, 2009

[6] Bruno Agard, Catherine Morency, and Martin Trepanier. Mining Public transport user behavior from smart card data. In 12th IFAC Symposium on Information control Problems in Manufacturing-INCOM, pages 17-19, 2006

[7] Xiaolei Ma, Yaojan Wu, Yinhai Wang, Feng Chen, and Jianfeng Liu. Mining smart card data for transit riders travel patterns. Transportation Research Part C-emerging Technologies, 36:1-12, 2013

[8] Lijun Sun, Kay W Axhausen, Der-horng Lee, and xianfeng Huang. Understanding metropoliatan patterns of daily encounters. Proceedings of the National Academy of Sciences of the United States of America, 110(34)13774-13779, 2013

[9] Irina Ceapa, Chris Smith, and Licia Capra. Avoiding the crowds: understanding tube station congestion patterns from trip data. In Proceedings of the ACM SIGKDD International Workshop on Urban Computing, pages 134-141. ACM, 2012.

[10] Catherine Morency, Martin Trepanier, and Bruno Agard. Analysing the variability of transit users behavior with smart card data. In Intelligent Transportation Systems Conference, 2006. ITSC'06. IEEE, pages 44-49.

[11] Ticiana L. Coelho Da Silva, De Mac, Jos Do, F A., and Marco A Casanova. Discovering frequent mobility patterns on moving object data. In ACM Sigspatial International Workshop on Mobile Geographic Information Systems, pages 60-67, 2014.

[12] Huang Lian, Tang Xiaolin, Lin Yulong, et al. Spatio-temporal analysis of bus passenger flow based on card data——Taking Shenzhen as an example [J]. China High-tech Enterprises, 2016 (17): 87-89.