EasyChair Preprint

№ 3147

An Spatial and Time Analysis Method of Big Bata of Public Transport Card Base on Classification Statistics and Visualization

Cheng Ding, Cheng Wang, Yu Cao and Jianwei Chen

April 9, 2020

An Spatial and Time Analysis Method of Big Bata of Public Transport Card Base on Classification Statistics and Visualization

Cheng Wang *
School of Computer Science and Technology Huaqiao University,No. 668,Jimei Avenue Xiamen
wangcheng@hqu.edu.cn

Cheng Ding
School of Computer Science and Technology Huaqiao University,No. 668,Jimei Avenue Xiamen
dingcheng@stu.hqu.edu.cn

Yu Cao
School of Computer Science and Technology Huaqiao University,No. 668,Jimei Avenue Xiamen
caoyu0701@foxmail.com

Jianwei Chen
Department of Mathematics and Statistics
San Diego State University
San Diego
jchen@sdsu.edu

***Abstract*— This paper presents a spatial and time analysis method of big data of public transport card based on classification statistics and visualization , including the use of classification statistics to obtain different types of IC card type distribution, swiping card number distribution, passenger flow time distribution, to determine the peak passenger flow and to Visualization method presented, and to do an effect analysis about reduce the number of bus swiping card in peak period , so as to provide a basis for public policy development. Secondly, taking Xiamen city as an example, using the above methods, it is concluded that the student cards' swiping card peak period share more time with all peak period. Here, two measures are proposed to increase the amount of student card swiping in peak hours and to set up student special line, so as to alleviate congestion and estimate the implementation effect.**

Keywords—Urban traffic, Visualization, Classification statistics; Bus swiping card data, Peak period, Discount card

I. INTRODUCTION

Bus is an essential bridge throughout people's daily life. In recent years, it is more and more difficult to get on the bus in rush hour. There are also a series of problems on bus, such as the debate about students' occupation of seats and giving up seats to the elderly. These problems have a great impact on the operation of buses. Students and the elderly generally use preferential cards when they take buses. They also enjoy more care when they use preferential cards. They crowd into buses with office workers during peak hours, which causes dissatisfaction of the general public. Thus, a topic arises: can people with preferential cards avoid the rush hours of going to and from work? To solve these problems is the focus of today's society. Use the data of bus swiping card to analyze the status quo of various types of cards, and put forward some targeted suggestions and estimate the implementation results according to the analysis and processing results. This not only provides a direction for the reform of public transport operation management, improves the utilization efficiency of resources, but also contributes to the harmonious development of society.

Tanaka Mikio and Sato Norio of Japan used IC card data to analyze passengers' travel behavior, and combined with data mining technology to study the rules and characteristics of subway passenger flow in Japan. The passenger travel time information obtained from the mining is used in the optimization research of metro traffic planning [1]. The University of Westminster in the UK uses public transport data for the analysis of the public transport market, including the number of card swiping per capita, travel time of passengers and regional distribution [2].

Yang Zhiwei, Zhao Qian, etc. combined the bus card passenger flow data with the bus passenger flow survey data, used the ordered clustering algorithm to divide the bus peak area, established the regression equation for predicting the passenger flow under different peak conditions, and used the card swiping volume in different peak areas to realize the prediction of the total passenger flow in different peak areas [3].

The difference of this paper is not to analyze bus IC card as a class, but to subdivide bus card into elderly card, student card and ordinary card.

II. AN SPATIAL AND TIME ANALYSIS METHOD OF BUS INTELLIGENT CARD DATA BASED ON CLASSIFICATION STATISTICS AND VISUALIZATION

A. Data preprocessing of IC card

Due to the huge amount of data and information in IC, there are still some gaps and wrong data in the data. In order to improve the quality of data, we need to preprocess the data [4].

- Data elimination：

1) Excluding other public transport card swiping data, such as subway, BRT, taxi, etc.

2) Excluding data outside the research time range.

3) Excluding records with missing field data.

Fund Project: The National Natural Science Foundation of China Youth Fund (51608209), the National Social Science Foundation(18BTJ031), the NSF(DMS-0907710), the Fujian Province Science and Technology Plan(2019H0017), the Project of Quanzhou Science and Technology Program of China(2018Z008)

Author brief introduction: Cheng Wang, born in 1984, Ph. D. associate professor. His research interests include traffic big data; **Cheng Ding**, born in 1995, M. S. candidate. His research interests include traffic big data; Yu Cao, born in 1995, M.S. candidate. Her research interests include traffic big data; **Jianwei Chen** is a professor at Department of Mathematics and Statistics, San Diego State University, USA. His research interests include data mining.

4) Excluding records with abnormal field data.

- Attribute elimination:

Some attributes are useless for the analysis and prediction of the passenger flow by swiping the card. The existence of these data can not improve the accuracy of the prediction results, but occupy the storage space of the database and reduce the efficiency of data mining [5].

B. The classification and statistics method of public transport passenger flow and its visual presentation method based on card type

There are many different types of bus cards. Generally speaking, there are three types of public transport cards: ordinary card, discount card and free card. The number of cards and the number of times of swiping are classified and counted to determine the peak period of swiping and present it in a visual way.

1) Type distribution and the distribution of swiping times

There are many different types of IC cards, let the total IC cards be denoted by K， The total IC cards K are divided into A_1、A_2、A_3……A_k.where, The ordinary card(A_{oc}),discont card(A_{dc}) and free card(A_{fc}) are as follows.

$$A_{oc} = \{A_1、A_2 \cdots\cdots A_i\} \tag{1}$$

$$A_{dc} = \{A_{i+1}、A_{i+2} \cdots\cdots A_j\} \tag{2}$$

$$A_{fc} = \{A_{j+1}、A_{j+2} \ldots\ldots A_k\} \tag{3}$$

and

$$A_{oc} \cup A_{dc} \cup A_{fc} = K \tag{4}$$

$$A_{oc} \cap A_{dc} = \emptyset \tag{5}$$

$$A_{oc} \cap A_{fc} = \emptyset \tag{6}$$

$$A_{dc} \cap A_{fc} = \emptyset \tag{7}$$

The IC card number corresponds to a cardholder, and each IC card number is unique. Count the number of active cards in each card category according to the card number, it will be denoted by M, and：

$$M_{oc} = count(A_1) + count(A_2) + \cdots\cdots + count(A_i) \tag{8}$$

Or

$$M_{oc} = M - M_{dc} - M_{fc} \tag{9}$$

$$M_{dc} = count(A_{i+1}) + count(A_{i+2}) + \cdots\cdots + count(A_j) \tag{10}$$

The percentage of all kinds of cards in active cards is：

$$\eta_1(A_x) = \frac{count(A_x)}{\sum_{y=1}^{k} count(A_y)} \times 100\% \tag{11}$$

and

$$\eta_1(A_{oc}) = \frac{M_{oc}}{M} \times 100\% \tag{12}$$

$$\eta_1(A_{dc}) = \frac{M_{dc}}{M} \times 100\% \tag{13}$$

$$\eta_1(A_{fc}) = \frac{M_{fc}}{M} \times 100\% \tag{14}$$

Count the swiping times of each type of card, and record it as N，then

$$N_{oc} = count(A_1) + count(A_2) + .. + count(A_i) \tag{15}$$

or:

$$N_{oc} = N - N_{dc} - N_{fc} \tag{16}$$

$$N_{dc} = count(A_{i+1}) + count(A_{i+2}) + \cdots\cdots + count(A_j) \tag{17}$$

$$N_{fc} = count(A_{j+1}) + count(A_{j+2}) + \cdots\cdots + count(A_k) \tag{18}$$

The proportion of swiping times of all kinds of cards is

$$\eta_2(A_x) = \frac{count(A_x)}{\sum_{y=1}^{k} count(A_y)} \times 100\% \tag{19}$$

and

$$\eta_2(A_{oc}) = \frac{N_{oc}}{N} \times 100\% \tag{20}$$

$$\eta_2(A_{dc}) = \frac{N_{dc}}{N} \times 100\% \tag{21}$$

$$\eta_2(A_{fc}) = \frac{N_{fc}}{N} \times 100\% \tag{22}$$

2) Time distribution of passenger flow

For the time distribution analysis of passenger flow, the equal interval division method is used. During the survey period, a day is divided into several periods with the same interval time, and the number of swipes per day in that period is counted. The survey period is R days, and the average number of daily card swiping in the n-th period is

$$Q_n = \frac{\sum_{i=1}^{R} N_i}{R} \tag{23}$$

3) Spatial distribution of passenger flow

For the spatial distribution analysis of passenger flow, the statistical method of the total number of people getting on and off the train according to the station in a certain period of time is adopted. In a certain period of time, count the number of passenger flow of each station in that period. The survey period is days, and the number of station passengers in n-th periods is

$$N_{sum} = N_{up} + N_{dowm} \tag{24}$$

4) Selection method of peak passenger flow

The equal interval division method is used to divide every possible peak section in a more detailed way. In order to see the difference of data more clearly, a bar chart is used. After that, the specific peak hours of all kinds of cards are determined by the sliding window method. The concept of the algorithm is

shown in Figure 1, where the time period with the largest sum of values is the peak hours. That is to say, the possible peak areas are divided into time periods at every minute interval, in which the peak hours account for time periods.

$$t = \frac{60}{x} \tag{25}$$

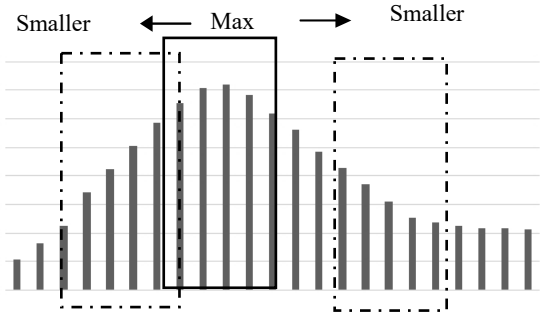

Fig. 1. Using sliding window method to calculate peak hour

C. prediction and analysis method of reducing the number of preferential card swiping in peak period

The discount card includes a discount card and a free card. According to the above method, if there is no overlap or little overlap between the peak hours of each card type and the peak hours of the overall card swiping situation, it means that the main card swiping time of this card type is not within the overall peak hours, and is not the main factor, then no measures need to be taken; if the peak hours of this card type and the peak hours of the overall card swiping overlap or mostly overlap, then this card type is Focus on the analysis object.

There are only two methods in this paper

1) Change the IC card amount in peak hours

According to "urban public transport operation, planning and economy", the implementation effect of the scheme is estimated according to the percentage of change of the number of traffic users per one percent change of fare in the cost elasticity of different passenger groups in the hypothetical system [6].

2) Special line for discount card

The method is to directly remove the people with preferential cards from the peak bus flow.

The above two methods have their own advantages and disadvantages. This paper only discusses and analyzes the mitigation of public transport passenger flow after implementation, and compares the card swiping before and after adjustment with the line chart.

III. DATA ANALYSIS OF IC CARD IN XIAMEN

A. Data description of E-card

E-card is public transportation IC card in Xiamen, and the original data of IC card is provided by Xiamen E-card company. In this paper, the data of IC card used is the data of the whole month of May 2016 in Xiamen, which has 54 164 739 records. All of which are preprocessed by Oracle database. It is found that some data have abnormal and wrong data, such as transaction amount is too large, transaction time is not in May, and data of IC card appears in non-operation hours. After data preprocessing, these error data and useless fields are removed, and IC card records of BRT are removed. The final data is the total IC card data of general public transportation in May 2016, which has 54 160 438 records. The valid fields used are shown in Table 1-the description of valid fields of E-card data in Xiamen.

TABLE I. XIAMEN E-CARD DATA VALID FIELD DESCRIPTION

Name	Notes	Remarks
CARD_PAN	Card Number	
CARD_TYPE	Card type	
CORP_ID	Merchant number	Ordinary bus：50241310002-50241310007 BRT：50241310008，50241310009
TRAD_AMT	Transaction amount	
BUSI_TIME	Transaction time	

The number of each IC card is unique, and the card type corresponds to the type of IC card held by the passenger, such as: senior citizen's preferential card, student's preferential card, general card, etc.; the merchant number is the vehicle type number, which is divided into general public transport and BRT (bus rapid transit) in large categories; the transaction amount of the card shows the amount of one-time consumption of the card held by the passenger; the transaction time includes the card date and time. The above data fields are important prerequisites for IC card data analysis. Figure 2 example of partial card swiping data

	CARD_PAN	CARD_TYPE	CORP_ID	TRAD_AMT	BUSI_TIME
1	0196228293	299013	50241310002	0.80	20160507010835
2	0938134671	401087	50241310002	0.80	20160507011136
3	5519428961	013001	50241310002	0.80	20160507010831
4	6059335400	403084	50241310002	0.80	20160507012354
5	6381395915	403084	50241310002	0.80	20160507012736
6	6636139070	299007	50241310002	0.80	20160507010941
7	8012270020376987	270001	50241310002	0.80	20160507012355
8	8012270021265088	270001	50241310002	0.80	20160507010427
9	8012270027874177	270001	50241310002	0.80	20160507010308
10	7564437050	299009	50241310003	0.80	20160507010513
11	4394264302	404087	50241310002	0.80	20160507010559
12	4534147150	299007	50241310002	0.80	20160507012353
13	8012013031844436	013003	50241310002	0.80	20160507011140
14	8012124029589916	124000	50241310002	0.50	20160507010313

Fig. 2. Part of the bus card data

B. Travel time analysis and visualization results of different types of bus cards

1) Type distribution, swiping frequency distribution and visual presentation results

According to the card transaction amount and card type query card data, we know that the card transaction amount is 0.80 yuan, 0.50 yuan and 0.00 yuan, and according to the card_type, find out the corresponding card_type name in the card type description provided by e-card company. The query results are shown in Table 3 card swiping data.

TABLE II. TABLE 2 CARD AMOUNT

CARD_TYPE	CARD_TYPE_NAME	TRAD_AMT
125 000	Elderly card	0.00
126 000	labor model preferential card	0.00
127 000	Martyr family card	0.00
132 000	Veteran cadre card	0.00
133 000	Preferential treatment card	0.00
124 000	Student card	0.50
319 000	Campus card	0.50

TABLE III. TABLE 3 SWIPING CARD DATA

CARD_TYPE_NAME	CARD_TYPE	Number of active cards	Count
Elderly card	125 000	97 905	3 553 813
labor model preferential card	126 000	1 360	56 807
Martyr family card	127 000	32	1 908
Veteran cadre card	132 000	1 123	35 630
Preferential treatment card	133 000	1 256	39 422
Student card	124 000	176 224	5 484 167
Campus card	319 000	60 388	1 419 343
Total		1 701 729	54 160 438

From table 2 and table 3, it can be seen that there is less data of the card with the amount of 0.00 yuan, including the labor model preferential card, Inferior family card, veteran cadre card and special care card. In this paper, these five kinds of 0.00 yuan discount cards are collectively referred to as elderly cards; the student preferential card and campus card are similar in the bus IC card system. In this paper, these two kinds of preferential cards are collectively referred to as the student card. According to formula(9)~(11) and (16)~(18),It can be concluded that：M_{sc}=236 612、M_{ec}=101 676、M_{oc}=1 363 441、N_{sc}=6 903 510、N_{ec}=3 687 580、N_{oc}=43 569 348.

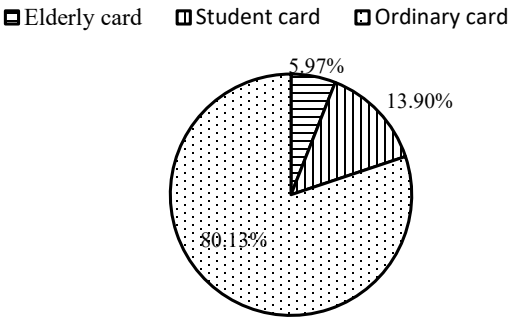

Fig. 3. Percentage of active cards

2) Time distribution and visualization results of passenger flow

This paper mainly studies the time distribution of working day passenger flow. The working day in May 2016 is shown in Figure 4.

Fig. 4. May 2016 calendar map

3) Analysis result of bus passenger flow in peak hour

The peak period of normal public transport on working days is between 6:00-10:00 and 16:00-20:00 in general, between 7:00-11:00 and 14:00-18:00 in elderly, between 6:00-9:00 and 16:00-20:00 in student. Divide peak hours into t=12 periods at x=5 minute intervals.

It can be seen from Figure 4 that there are 21 working days in May 2016. According to the statistics of business card swiping on ordinary public transport working days in one hour interval, take the data between 6:00 and 24:00, calculate the average daily business card swiping times in each time period during the survey period according to formula (23), and show them with a line chart.

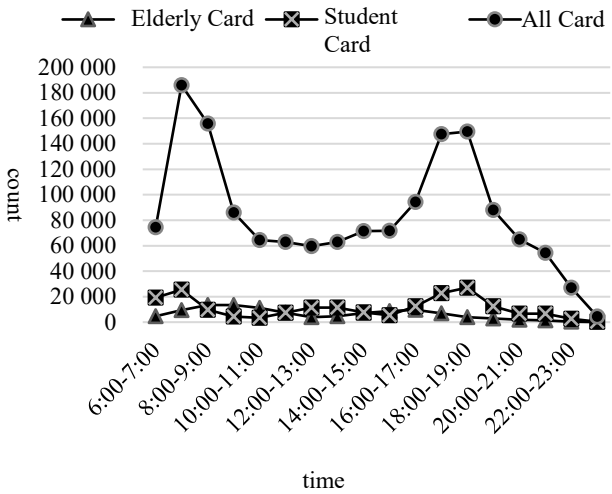

Fig. 5. May 2016 Xiamen City, the general bus work day average swiping card situation line chart

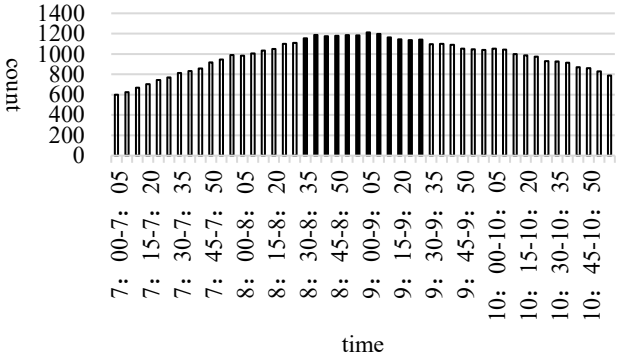

Fig. 6. The Elderly card Morning peak map

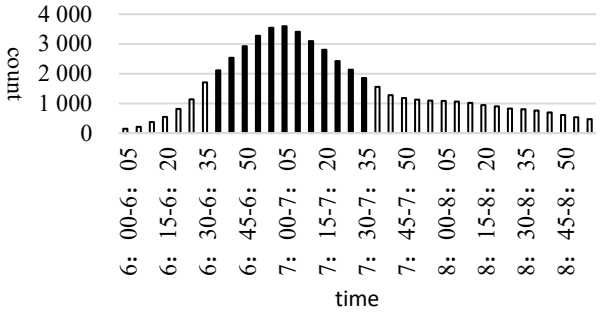

Fig. 7. The student card Morning peak map

From the point of view of the coincidence degree of peak hours, the peak of early and late card swiping of old people's card is staggered with the peak of overall card swiping, and the early peak of student's IC card swiping overlaps with the overall early peak swiping by 20 minutes, and the late peak overlaps by 45 minutes.

TABLE IV. CARD CLASS SWIPING CARD PEAK HOURS

Card type	Morning peak	Evening peak
count	7:15-8:15	17:35-18:35
Elderly Card	8:30-9:30	15:45-16:45
Student Card	6:35-7:35	17:50-18:50

C. The predict result of reducing the number of preferential card swiping in peak period

As the preferential card for competing with ordinary people for public transportation resources in the peak period of Xiamen city is mainly student card, only student card is analyzed, and the following measures are considered for student card:

1) The amount of the student's IC card in the rush hours of the working day (7:15-8:15 and 17:35-18:35) becomes 0.80 yuan as the same as the ordinary card.

2) Set up student specific line in peak hours, during which students can only take student specific line.

The first method is to equate student's IC card with ordinary card in peak hours, so there is no fairness problem. At the same time, due to the increase of ticket price, the use of student's IC card in peak hours is reduced. The second method is to directly remove the student population from the peak bus passenger flow.

.

TABLE V. PEAK HOUR STUDENT CARD ADJUSTMENT SITUATION

Morning peak				**Evening peak**			
Before adjustment		***After adjustment***		***Before adjustment***		***After adjustment***	
Count	Proportion	Count	Proportion	Count	Proportion	Count	Proportion
18 604	9.84%	16 371	8.76%	27 078	16.34 %	23 829	14.67 %

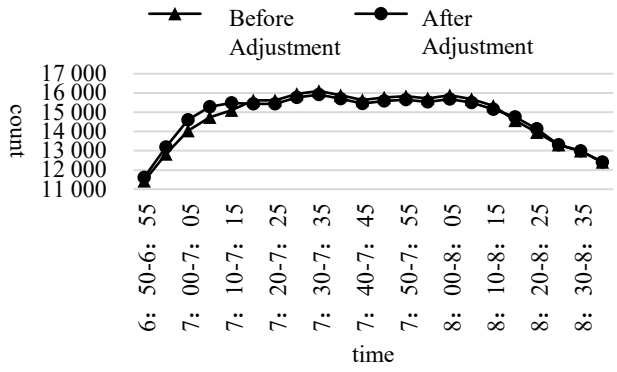

Fig. 8. After increase student card amount in peak hours compared to the morning peak situation

In the rush hours, the number of IC card swiping of the students in the student specific line will be reduced to 0 by default.

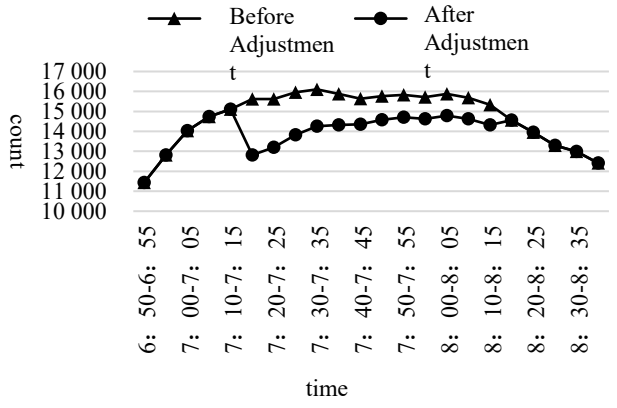

Fig. 9. Peak hours to open a student line morning peak adjustment before

and after the card comparison chart

D. Analysis conclusion

1) It can be seen from Figure 3 that the percentage of frequent use of discount cards in the elderly accounts for 5.97% of total cards, and the number of swiping cards accounts for 6.81% of the total, only up 0.84%. The number of student cards accounts for 13.90% of total cards, the number of swiping cards accounts for 12.75% of the total, and down 1.15%. Thus it may be known, most of the people who hold discount card will not use the card inordinately due to the ticket price is halved (0.50 yuan) or free (0.00 yuan), leading to crowded situation on the bus. There is no increase in travel and competition for public transport resources in rush hours due to preferential policies.

2) According to the data of public transport card swiping, only a small number of the old people who hold discount card in Xiamen travel by public transport at the peak of the working day, most of them choose to travel earlier or later to avoid the peak, and do not compete with ordinary people such as working person for public transport resources. Students who hold discount card can hardly avoid the peak of public transport travel on the working day.

3) It can be seen from the result of two measures proposed to reduce the number of student's IC card swiping in peak hours of working days that the measures of setting up student specific line in rush hours are more effective, releasing more public transport resources, and better relieving the pressure of take a bus in rush hours.

IV. Summary and Prospect

This paper obtains the passenger flow information of various concession cards from a large number of public transport IC card data, and analyze the situation of discount card used in rush hours. We find that there is no situation that the people who hold discount card in Xiamen city increase their travel and compete for public transport resources in rush hours because of preferential policies. Most of the old people who hold discount card will choose to travel earlier or later to avoid peak hours when they take public transport on weekdays. During the peak period, student's IC card is the main type to compete with ordinary people for public transport resources. In order to reduce the number of card swiping in rush hours, this paper proposes a better measure to set up student specific line in rush hours.

References

[1] Tanaka Mikio, Sato Norio, Sakuma Yasushi, et al. A Study on the Applicantion of Data Mining Technology to the Analysis of Passenger Flow Data [J]. Railway Technical Research Institute, 2002,16(11):37-42

[2] Bagchi,M,White,ER,003.Use of public transports smart card data for understanding travel behavior[C]. Proceedings of the European Transport Conference, Strasbourg 8-10 October.

[3] YANG Z W.ZHAO Y.ZHAO S C.JIN L.MAO Y. Passenger Flow Volume Forecasting Method Based on Public Transit Intelligent Card(IC)Survey Data[J]. Traffic standardization,2009,(09):115-119.

[4] LIAO Z R. Based on the analysis of bus IC data bus traffic[D]. Yunnan University,2010.

[5] LIU X Q. Bus passenger flow analysis and prediction based on transport IC card big data[D]. Guangdong University of Technology,2016.

[6] Vukan R.Vuchic. Urban Transit Operations,Planning,and Economics[M]. Beijing: China Railway Publishing,2012.

[7] Hazalton M L. Statistical inference for time varying origin-destination matrices[J]. Transportation Research Part B Methodological, 2008,42(6):542-552.

[8] Alsger A, Tavassoli A, Mesbah M, et al. Evaluation of effects from sample-size origin-destination estimation using smart card fare data[J]. Journal of Transportation Engineering Part A-Systems, 2017,143(4):1-10.

[9] Jingfeng Yang,Jian Gang Jin,Jianjun Wu, et al.Optimizing Passenger Flow Control and Bus-Bridging Service for Commuting Metro Lines[J].Computer-Aided Civil and Infrastructure Engineering,2017,32(6):458-473. DOI:10.1111/mice.12265.

[10] Qun Chen.Global Optimization for Bus Line Timetable Setting Problem[J].Discrete dynamics in nature and society,2014,2014(Pt.1):636937-1-636937-9.

[11] Dan.Zheng,Yao.Wang,Peng Zhi.Tang, et al.Application of Data Mining in the Forecasting of Railway Passenger Flow[J].Advanced Materials Research,2014,2817(1670):958-961. DOI:10.4028/www.scientific.net/AMR.834-836.958.

[12] Yiru Wang,Yiru Wang,Jinhua Zhang, et al.A Network Flow Approach for Optimizing the Passenger Throughput at an Airport Security Checkpoint[J].IOP Conference Series: Materials Science and Engineering,2019,490(4):042047 (6pp). DOI:10.1088/1757-899X/490/4/042047.

[13] TANG Hai-yan,QI Wei-gui,DING Bao.Prediction of elevator traffic flow based on SVM and phasespace reconstruction[J]. Journal of Harbin University of Technology(English version),2011,18(3):111-114.

[14] GAO ZiYou,ZHOU HuaLiang,LI KePing, et al.Analyzing and evaluating the three-line rail traffic[J]. Chinese Science Series E(English version),2008,51(7):949-956. DOI:10.1007/s 11431-008-0095-8.

[15] Miriam F. Bongo,Lanndon A. Ocampo.Exploring critical attributes during air traffic congestion with a fuzzy DEMATEL–ANP technique:a case study in Ninoy Aquino International Airport[J]. Journal of modern communications(English version),2018,26(2):147-161.