Citation: ZHENG Haixing, ZHU Haiming, JIANG Yin, CHEN Feng, GE Han. Estimating Passengers' on-and-off Volumes at Transit Station [J], Urban Transport of China, 2018 (06): 36–42, 17.

Estimating Passengers' on-and-off Volumes at Transit Station

ZHENG Haixing[1], ZHU Haiming[1], JIANG Yin[2], CHEN Feng[3], GE Han[4]

1. Tianjin Urban Planning & Design Institute, Tianjin 300201, China;
2. Tianjin Municipal Engineering Design & Research Institute, Tianjin 300201, China;
3. China Metro Engineering Consulting Corporation, Beijing 100037, China;
4. Tianjin Municipal Public Security Traffic Management Bureau, Tianjin 300201, China

Abstract: The characteristics of passenger flow are important indications for public transit network planning and operation optimization. To make up the missing passenger volumes data from the IC card readings, this paper proposes a methodology to estimate the passengers' on-and-off volumes at transit station. The method can effectively solve the problem of time difference between IC card charging system and bus GPS system, and the lagging in the IC card swiping. Taking Tianjin and Shenzhen as examples, the paper presents the application of the method from macroscopic to microscopic views, such as bus line investigation and passenger flow characteristics. The integration between bus and rail transit is also discussed. The results show the high accuracy and strong applicability of the methodology in estimating passengers' on and-off volumes at transit station. **DOI:** 10.13813/j.cn11-5141/u.2018.0605-en

Keywords: transportation planning; big data; IC card data; on-and-off volume at transit station; data fusion

0 Introduction

The characteristics of passenger flow are important indications for public transit network planning and operation optimization. Comprehensive and systematic mining and analysis of passenger flow characteristics are of great significance to improve the level of industry decision-making and management, to promote the establishment and improvement of industry access, supervision, subsidies and other systems, and to promote the sustainable and healthy development of urban economy [1]. The traditional technical means to obtain passenger flow characteristics are mainly manual surveys, including manual count and questionnaire survey. Limited by technical means and cost, there are many problems in the manual survey data, such as insufficient samples, randomness and even sample deviations, which makes it difficult to make a comprehensive and in-depth objective evaluation of passenger flow characteristics. The industry planning and management of public transportation are in need of the support of scientific and information-based technology means and methods.

In recent years, public charging system of transit IC card (hereinafter referred to as "IC card") has been widely used in various cities. The transaction data of IC card record in detail the relevant about the use of public transportation system by passengers, such as the card number, bus or rail transit line, vehicle, transaction time, and transaction cost. It provides valuable data resources for analyzing passenger flow of urban public transportation system. However, since the IC card charging system failed to fully consider the needs of transportation analysis when it was built, the data lack some key travel information. For example, the IC card data of one-ticket bus line lack such key information as passenger on-and-off volumes and time, which brings inconvenience to passenger flow analysis and makes it difficult to carry out deep data mining [2]. To make up for the above deficiencies, researchers in the world have carried out a series of studies on the estimation of passenger on-and-off volumes at transit station through the fusion of IC card data and vehicle GPS data.

For the estimation of passengers' on-volumes at transit station, the IC card data were clustered temporally in Reference [3], and the time labels of each cluster were matched with the estimated time of the vehicle passing through the station. References [4–5] made use of the GPS and IC card data fusion to achieve the estimation of passengers' on volumes at transit station through matching the vehicle arrival time with the IC card swiping time. The estimation of on-volumes was carried out in Reference [6] according to the time matching of GPS data and IC card data and the similarity of GPS arrival interval between adjacent stations and clustering interval between adjacent IC cards. In Reference [7], based on the analysis of GPS and IC card data fusion, single user was taken as the basic unit to estimate the on-volumes, and the clock difference between GPS system

Received: 2018-06-06
Supported by: The Third Batch of Loan-Assisted Technical Project of the World Bank in Tianjin (P148129)
First author: ZHENG Haixing (1990–), female, from Shaoyang, Hunan Province, Master's degree, engineer, is mainly engaged in the research of transportation big data and model, as well as transportation planing. E-mail: 574854245@qq.com

and IC card system was further considered.

For the estimation of passenger's off-volumes at transit station, Reference [8] took a single passenger as the basic analysis unit, and determined the alighting station through considering the spatial relationship between the passenger's boarding station and alighting station. Reference [4] made use of GPS data and IC card data fusion to estimate the alighting station based on the characteristics of commuters travelling back and forth between the starting station and the terminal. Reference [9] relied only on the passenger boarding time data to determine his alighting station through transfer analysis and cluster analysis in the absence of GPS data. In Reference [10], the probability model of trip chain was used to determine the random stop. In Reference [7], it was proposed to estimate the alighting station based on the first and last rides (the same line), continuous transfer, historical trip probability, etc.

Among the above estimation methods, some do not consider the clock difference and the lagging in the IC card swiping; some use estimation rules which are not perfect and prone to misjudgment; and some are on the basis of a single user which affects the estimation efficiency. In view of this, the paper proposes a method of estimating on-and-off volumes at transit station based on clustering similarity and data fusion. After the IC card data are clustered temporally, a cluster is used as the basic analysis unit, and then the lagging of card swiping is processed according to the relationship between the median and the corresponding arrival time interval vehicle GPS. Through the similarity between the label vector of IC card clustering time and the vehicle GPS estimated opening time vector under different shifting values, the problem of clock difference in IC card data is analyzed. Effective estimation of alighting station is achieved through comprehensive consideration of the transfer relationship, the relationship between the first and last rides (including the first and last rides of the day, the last ride of the day and the first ride of the next day), the commuting relationship and the diversity of bus line choices, etc. Taking Tianjin and Shenzhen as examples, the paper presents the application of the method from macroscopic to microscopic views and discusses the integration between bus and rail transit, providing support for upgrading the level of public transportation services in Tianjin Binhai New Area.

1 Data preparation

1.1 Data overview

The data mainly included dynamic operation data, and basic data such as bus/rail transit line stations, etc.

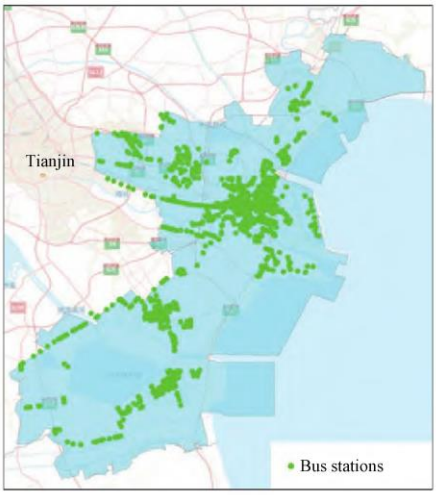
1) GPS arrival and departure data: By preprocessing the original GPS data, the vehicle GPS arrival and departure time table is generated, which mainly includes line ID, line name, license plate number, trip ID, driving direction, station sequence, station ID, arrival time, and departure time.

2) IC card data: It includes bus and rail transit card data. By preprocessing the detailed IC card data, the IC card information table is generated, which mainly includes card number, card swiping time, card-swiping terminal equipment, license plate number (or rail transit station), and card-swiping fees.

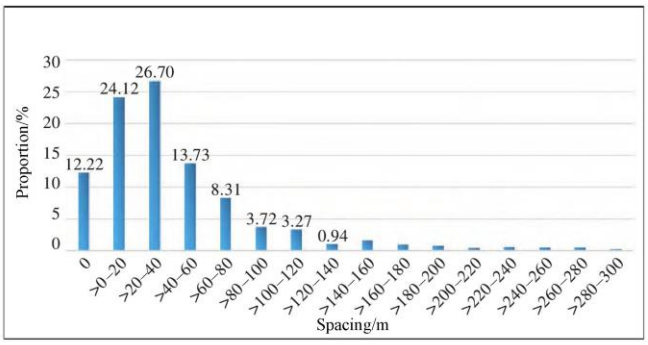
3) Daily passenger transport volume of the line: Statistics of passenger transport volume by line, direction and date include the card-swiping volume and cash payment volume, which is used for the sample expansion analysis of on-and-off volumes at transit station.

4) Transit station basic data: It mainly includes the line ID, line name, direction, station sequence, station ID, station name, and station latitude and longitude.

To improve the estimation efficiency, the transit station data were preprocessed. Spatial clustering was performed for different platforms at the same station to form station groups. The clustering conditions are as follows: bus stations have the same or similar names, with distance less than 150 m (as shown in Figure 1). Stations with an average platform spacing of less than 100 m account for approximately 89%, 92% less than 120 m, and approximately 95% less than 150 m.



a Distribution of bus stations in Tianjin Binhai New Area



b Average spacing between different platforms at the same station

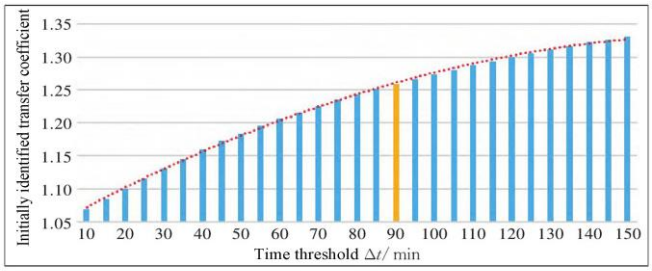
Figure 1 Station clustering analysis

1.2 Sensitivity analysis of transfer identification parameters

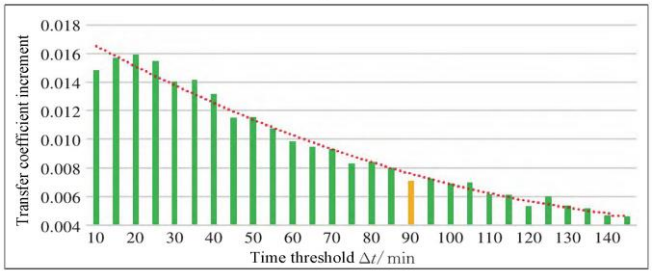
The main influence parameters for the estimation method include the transfer time reachability threshold Δt (the interval between two adjacent card-swipings for boarding) and the transfer space reachability threshold Δd (the distance between the last stop and this boarding station). In order to further determine the values of parameters reasonably, the estimation result of passengers' on-and-off volumes at transit stations in Tianjin Binhai New Area wa taken as an example to analyze the sensitivity of different values of parameters to the estimated results.

1) Transfer time reachability

As shown in Figure 2, as the transfer time reachability threshold Δt increased, the initially identified transfer coefficient (only satisfying the time reachability condition) increased continuously; and when Δt was greater than 90 min, the change of the initially identified transfer coefficient tended to be stable.



a Analysis of the change of transfer coefficient



b Analysis of incremental change of transfer coefficient

Figure 2 Threshold of time accessibility (transfer between buses)

2) Transfer space reachability

As shown in Figure 3, with the increase in transfer space reachability threshold Δd, the off-volume estimation rate increased gradually; and when Δd was greater than 1 km, the off-volume estimation rate tended to be stable.

2 Model establishment

2.1 Estimation of on-volumes based on clustering similarity

1) Select the GPS arrival and departure data and IC card data in one day for one bus, and sort them by time.

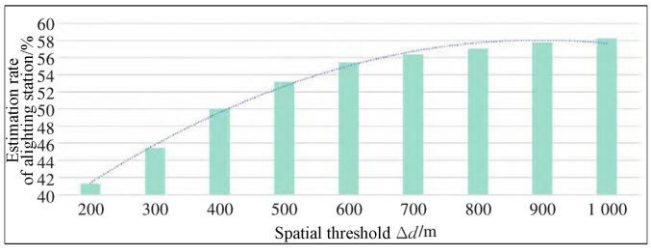


Figure 3 Threshold of space accessibility (transfer between buses)

2) Cluster the IC card data by time. The clustering condition is that the adjacent time interval is less than Δt (in general, the interval time of card-swipings for boarding at the same station is less than 72 s [9]). Multiple clusters are presented if the mentioned interval time exceeds 72 s due to the lagging in the card swiping or other reasons. In addition, the minimum value of the corresponding swiping time of the cluster is regarded as its time label.

3) Calculate the average GPS arrival and departure time. Use it as the estimated vehicle opening time, and take this time as the basis for matching with the IC card swiping time.

4) Determine the optimal shifting value of the IC card swiping time considering the clock difference. Select the GPS estimated opening time vector M of the subject vehicle and time label vector N corresponding to IC card cluster (the first recorded swiping time of the cluster). Perform the similarity tests between the vector M and the time label vector N of the IC card cluster after the k^{th} shifting respectively. Then select the shifting value when the similarity is the best (that is, when the value of the similarity evaluation index F is the minimum) as the optimal shifting value (see Figure 4).

$$F=\sum_{j=1}^{n}\left|M_j-N_{k,j}\right|/n, \tag{1}$$

where M_j is the estimated opening time of the vehicle at the j^{th} station; $N_{k,j}$ is the minimum time label for the IC card cluster of the vehicle to be matched with the station j after the k^{th} shifting; n is the number of stations. Combined with the actual data, the maximum shifting is 15 min.

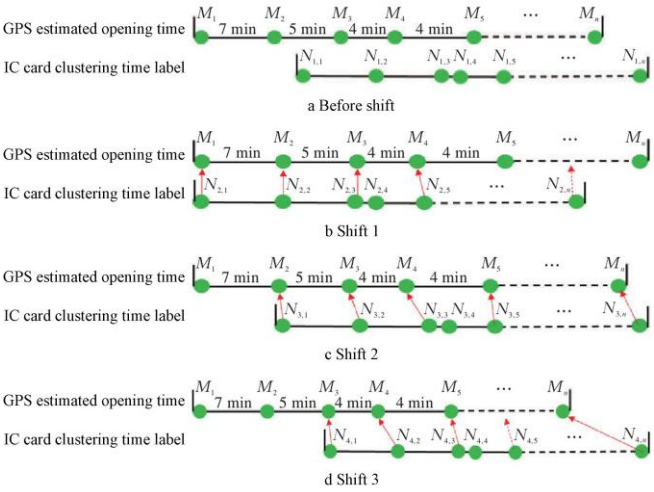


Figure 4 Procedures of optimal shifting value

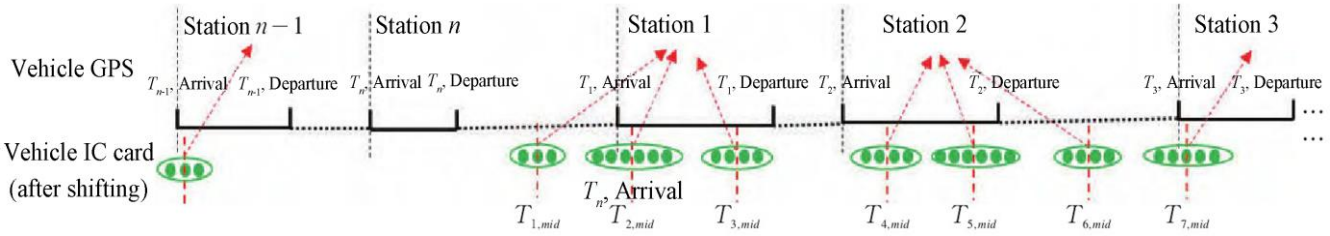


Figure 5 Estimating the boarding passengers at stations

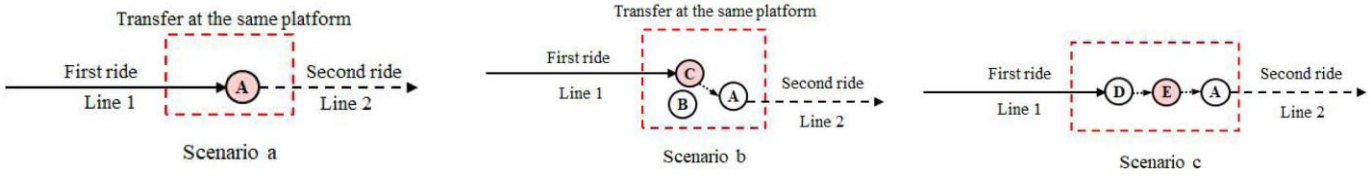


Figure 6 Estimating the off-volumes at stations based on transfer relationship

5) Carry out the interval division with the GPS arrival time of each station as the dividing point. Shift the selected single-bus single-day IC card data as a whole according to the optimal shifting time and take the IC card cluster as a unit. Then determine the corresponding boarding station based on the fact that the middle value falls into the GPS interval (see Figure 5).

2.2 Estimating the off-volumes at transit stations based on transfer, commuting and round-trip relationships

1) Estimating off-volumes at transit stations based on transfer relationship

The identification of transfer relationship should satisfy two conditions: temporal proximity and spatial proximity.

① The card-swiping time difference between a passenger's two rides is $\leq \Delta t$. According to surveys, the average bus passenger travel time for a single trip is about 30 min. Therefore, when the mode of travel is transfer from bus to subway/bus and the two routes are different, Δt takes 90 min (see Table 1 for analysis); when the mode is transferred from subway to bus, Δt takes 20 min considering the time for walk.

Tab. 1 Results of on-and-off volumes at transit stations

Category	Shenzhen		Tianjin Binhai New Area	
	Daily card-swiping /times	Proportion/%	Daily card-swiping /times	Proportion /%
Successful estimation of on-volumes	3 078 880	98	196 792	93
Successful estimation of off-volumes	1 906 313	61	123 321	58
Successful estimation of on-and-off volumes	1 843 175	59	120 614	57
Total card-swiping	3 142 018		211 604	

Note: The IC card-swiping in Tianjin Binhai New Area is only the card swiping data corresponding to the bus routes operated by Binhai New Area Public Transport Group Co., Ltd.

② The alighting station of the first ride and the boarding station of the second ride are located within the same group of stations or within 1 km apart, and the routes for the two rides are different.

As shown in Figure 6, in Scenario a, the alighting station of the first ride is the boarding station A of the second ride; in Scenario b, the alighting station of the first ride is Station C, which is the closest to the boarding station A of the second ride and within the same station group as the latter; and in Scenario c, the alighting station of the first ride is Station E, which is closest to the boarding station A of the second ride and within 1 km apart.

2) Estimation of off-volumes at transit stations based on the first and last rides and commuting relationship

According to commuter's round-trip characteristics between morning peak and evening peak, the alighting station for the first ride in the morning peak and the boarding station for the last ride in the evening peak should generally be located in the same station group (or within a distance of 1 km), and the alighting station for the last ride in the evening peak and the boarding station for the first ride in the morning peak should also be located in the same station group. According to the general travel characteristics of passengers, in general, if the boarding stations for the first ride on the day and the next day are located in the same station group, then the alighting station for the first ride and the boarding station for the last ride on the day should also be located in the same station group, and the alighting station for the last ride and the boarding station for the first ride on the day should also be located in the same station group.

As shown in Figure 7, in Scenarios d and f, the alighting station for the first ride is Station B1, which is the closest to the boarding station B2 for the last ride and within a distance of less than 1 km, and the alighting station for the last ride is Station A2, which is closest to the boarding station A1 for the first ride and within a distance of less than 1 km. In Scenario e, the alighting station for the last ride is Station A2, which is closest to the boarding station A1 for the first ride and within a distance of less than 1 km.

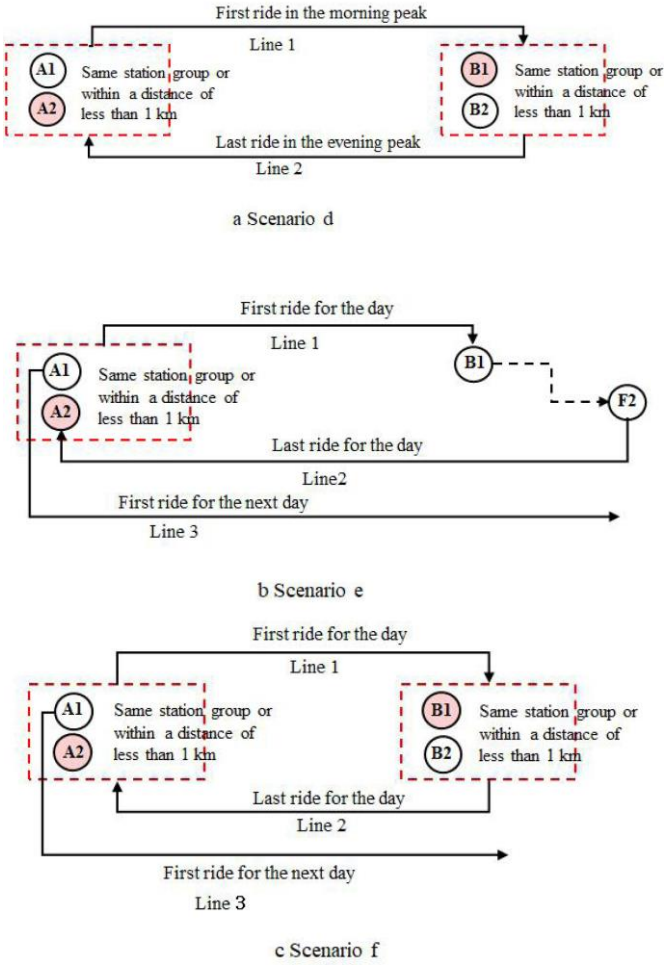


Figure 7 Estimating off-volumes based on the first and last trips and commuting relationship

For other travels without obvious regularity, considering their randomness, estimation of the alighting station is no longer based on the estimated probability of the alighting station based on history records, but based on the actual analysis demand, and sample expansion is performed in terms of line or line network.

3) Sample expansion

Sample expansion is performed for the estimation of on-and-off passenger flows separately according to the estimation rate of on-and-off passenger flows and the card-swiping rate of each line.

3 Results and verification

3.1 Results

According to the sensitivity analysis of transfer identification, it is suggested that transfer time reachability threshold Δt takes 90 min and transfer space reachability threshold Δd takes 1 km. Taking Shenzhen City and Tianjin Binhai New Area as examples, the passengers' on-and-off volumes at transit station are estimated. As shown in Table 1, the successful rate of on-volumes estimation in Shenzhen is about 98%, and that of off-volumes is about 61%; due to the absence of subway data in the IC card data of Tianjin Binhai New Area, the proportions of successful estimations of on-and-off volumes are slightly lower, with about 93% for on-volumes and about 58% for off-volumes.

3.2 Estimation method verification

3.2.1 From microscopic views: test route survey

Taking bus M352 route (Xinbaili–Shenzhen North) in Shenzhen as an example, we compared the survey result of passenger flow (15 trips in total) with the estimated result of IC card data on the same day (sample expansion was performed according to the successful route estimation ratio and card-swiping rate) (see Figures 8 and 9), and found that the trend of passenger flow distribution between the two results was the same, and there was a good correlation between them.

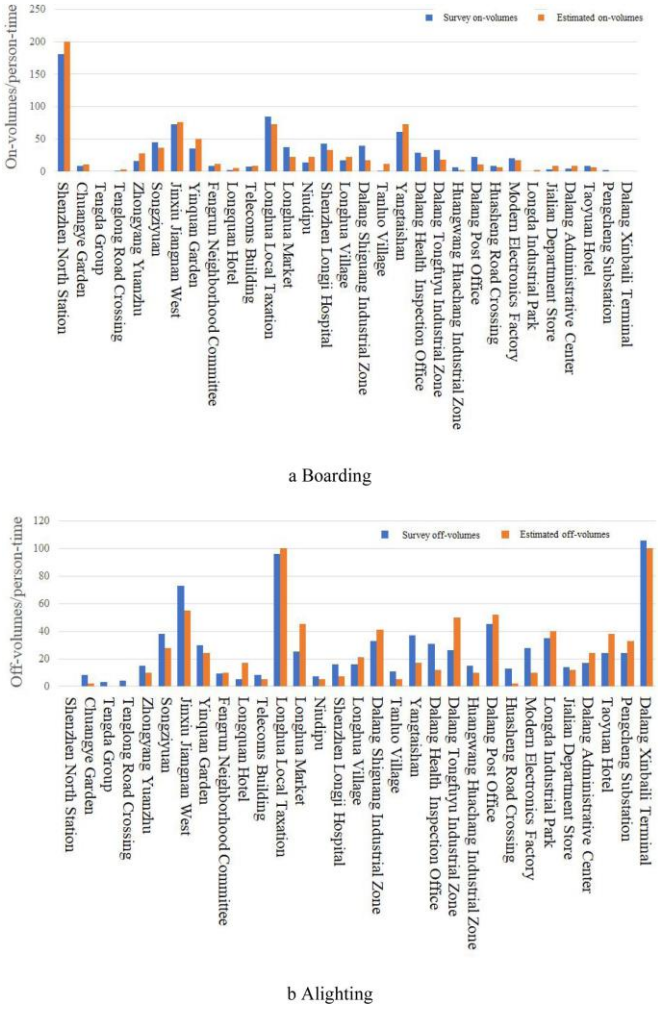
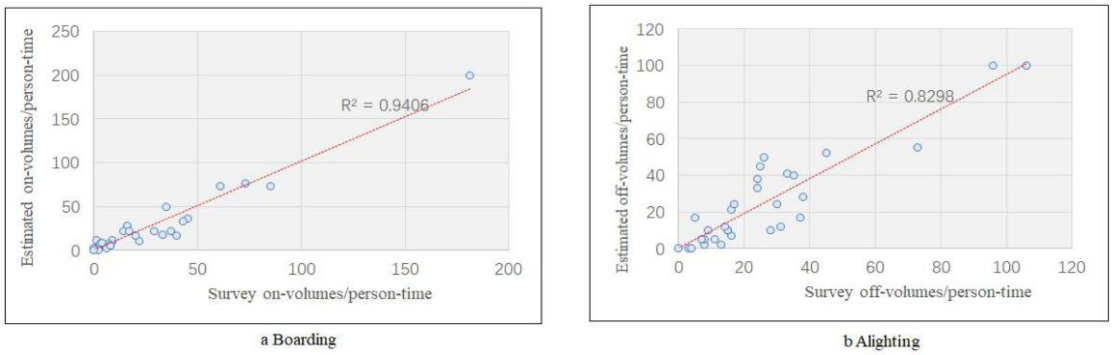
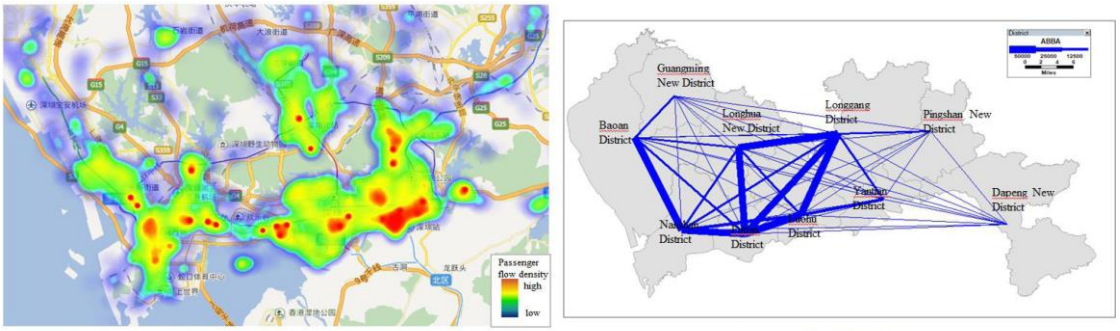


Figure 8 Verification of the estimated on-and-off volumes at stops along bus M352 route



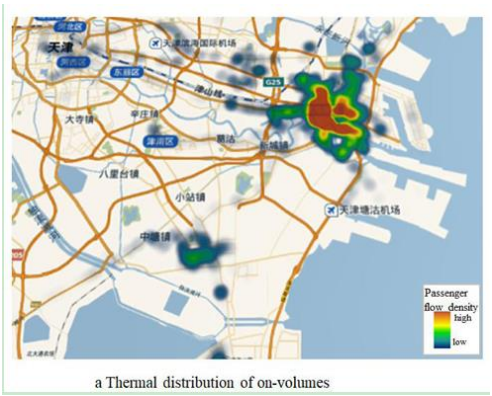
a Boarding　　b Alighting

Figure 9　Calculation errors in the estimated on-and-off volumes at stops along bus M352 route

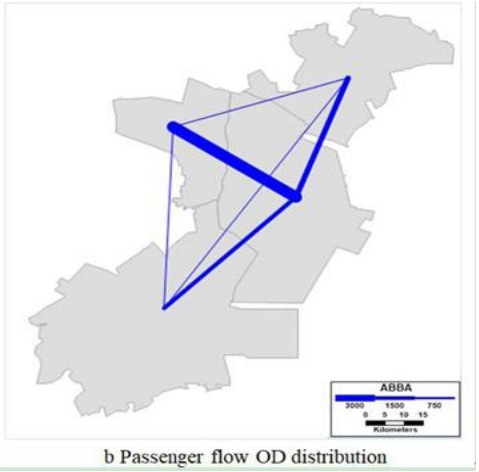


a Thermal distribution of on-volumes　　b Passenger flow OD distribution

Figure 10　Spatial distribution of bus passenger flow in Shenzhen



a Thermal distribution of on-volumes



b Passenger flow OD distribution

Figure 11　Spatial distribution of bus passenger flow in Binhai New District of Tianjin

To further evaluate the accuracy of the estimation method, *GEH* was selected for error testing.

$$GEH = \sqrt{2\times(C-V)^2/(C+V,)} \quad (2)$$

where C is the calculated value, while V is the actual value. It is generally accepted that there is no significant difference between the calculated value sequence and the actual value sequence when *GEH* is less than 5.0. Compared with the relative error, *GEH* can better evaluate the error; and its value is only related to the degree of deviation from the true value, which has nothing to do with the positive or negative deviation. In addition, its error sensitivity to smaller values is rather low.

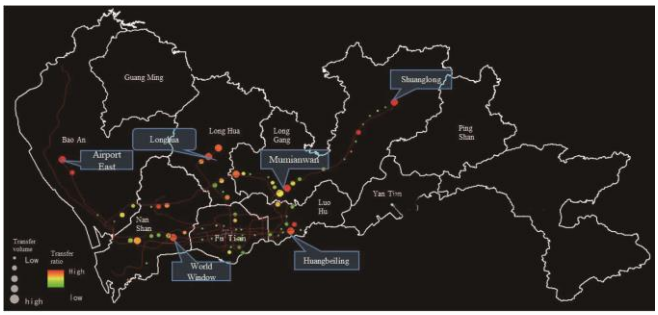
The average *GEH* values of estimated on-and-off volumes at stops along bus M352 route based on Formula (2) are 1.65 and 1.99, respectively, which are acceptable. Therefore, there is no obvious difference between the analysis result of IC card data and the manual survey result, and the estimation method accuracy is high.

3.2.2 From macroscopic views: comparative analysis of passenger flow characteristics

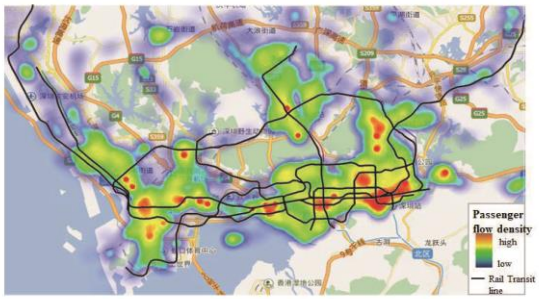
Tianjin Binhai New area is similar to Shenzhen in geographical space, which is a long and narrow zone along the coast. The spatial distribution of bus passenger flow in Shenzhen echoes with its geographical spatial distribution, showing obvious banded characteristics along the east–west direction to the main urban area (Nanshan District–Futian District–Luohu District), and Baoan District, Longhua District and Longgang District on the periphery of the main

urban area are closely related to the passenger flow in the main urban area (see Figure 10). In Tianjin Binhai New Area, the high passenger density is concentrated in the core area, and the connection between the core area and the west section is relatively close (see Figure 11).

In the aspect of integrated development of bus and rail transit, the main bus passenger flow concentration stations in Shenzhen are basically covered by rail transit, and the degree of integration of rail transit and bus is relatively high, especially the station at the end of the rail transit line (about 47.5% of the passengers entering and leaving the subway at the Baoan Airport East Station are from the bus transfer, as shown in Figure 12.). In Tianjin Binhai New Area, only the Jinbin Light Rail Line 9 is opened, and the distance between the rail station and the bus passenger concentration stations is relatively long (see Figure 13).



a Transfer volume and transfer ratio of passenger flow from bus to rail transit station

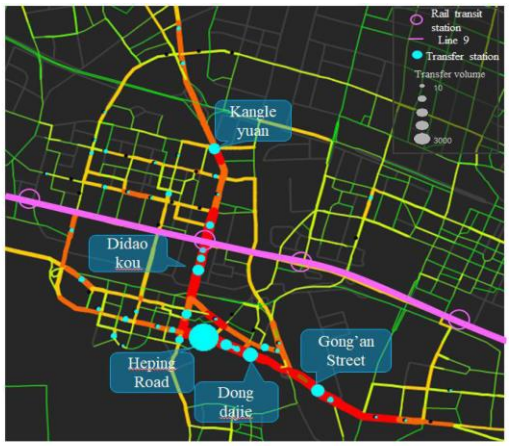


b Coupling relationship between bus passenger flow and rail transit line

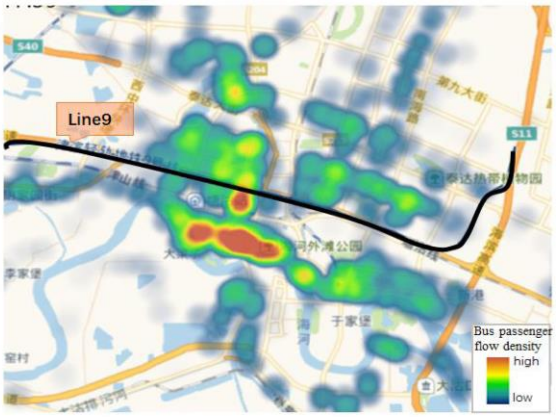
Figure 12 Integration of bus and rail transit in Shenzhen

4 Conclusion

On the basis of summarizing previous studies on the estimation of passengers' on-and-off volumes at transit station, we proposed a method to estimate passenger's on-and-off volumes at transit station based on big data fusion of public transport by using bus GPS data and IC card data, and presented its application in two cities of Tianjin and Shenzhen. The results showed that the estimation method has the characteristics of high success ratio, high accuracy and good generality. On this basis, overall perception of the operation of public transport from several spatial dimensions such as stations, lines and areas could be realized. Besides, the method can be applied to urban public transport planning, policy evaluation, line network planning and other fields to



a Main bus transfer stations and their coupling relationship with rail transit station



b Coupling relationship between bus passenger flow and rail transit line

Figure 13 Integration of bus and rail transit in Binhai New District of Tianjin

provide scientific and quantitative decision-making basis for related work.

References

[1] Chen Feng, Liu Jianfeng. Characteristics of Bus Passenger Flow Based on IC Card Data: A Case Study in Beijing [J]. Urban Transport of China, 2016, 14(1): 51–58+64 (in Chinese).

[2] Zhou Rui. Passenger Flow Calculation for Bus Stations Based on IC Card Data [D]. Beijing: Beijing Jiaotong University, 2012 (in Chinese).

[3] Yin Changyong, Chen Yanyan, Chen Shaohui. Bus Station Passenger Matching Method Based on Cluster Analysis Method [J]. Journal of Transport Information and Safety, 2010, 28 (3): 21–24 (in Chinese).

[4] Luo Lei. 基于 IC 卡信息的公交客流空间分布特征分析方法研究 [D]. Nanjing: Southeast University, 2010 (in Chinese).

[5] Ma Xiaolei, Liu Congcong, Liu Jianfeng, et al. Boarding Stop Inference Based on Transit IC Card Data [J]. Journal of Transportation Systems Engineering and Information Technology, 2015, 15(4): 78–84 (in Chinese).

[6] Chen Shaohui, Chen Yanyan, Lai Jianhui. An Approach on Station ID and Trade Record Match Based on GPS and IC Card Data [J]. Journal of Highway and Transportation Research and Development, 2012, 29(5): 102–108 (in Chinese).

[7] Hou Xianyao, Chen Xuewu, Chen Zhengrong, et al. 基于 IC 卡和 AVL 系统数据的公交乘客上下车站点判别方法 [C]//第七届中国智能交通年会学术委员会. 第七届中国智能交通年会优秀论文集. Beijing: Publishing House of Electronics Industry, 2012: 25–32 (in Chinese).

[8] Chen Xuewu, Dai Xiao. 公交 IC 卡持卡乘客下车站点确定方法研究

[C]//第三届中国智能交通年会学术委员会. 2007 第三届中国智能交通年会论文集. Nanjing: Southeast University Press, 2007: 94–98 (in Chinese).
[9] Zhao Peng. Researching the Origin- Destination Site Estimation of Passengers Based on Bus Smart Card Records of Chengdu [D]. Chengdu: Southwest Jiaotong University, 2015 (in Chinese).
[10] Hu Jihua, Deng Jun, Huang Ze, et al. Trip-Chain Based Probability Model for Identifying Alighting Stations of Smart Card Passengers [J]. Journal of Transportation Systems Engineering and Information Technology, 2014, 14(2): 62–67+86 (in Chinese).

