ISSN 1330-3651 (Print), ISSN 1848-6339 (Online)

https://doi.org/10.17559/TV-20220903113056
Original scientific paper

Identifying Left Behind Passengers at Subway Stations from Auto Fare Collection Data

Lianghui XIE, Zhenji ZHANG, Daqing GONG*

Abstract: With the rapid growth in transport demand, it has become a frequent occurrence that passengers are left behind especially during peak hours in subway, which has led to a significant reduction in the level of service. In this paper, we propose a left behind passengers identifying method based on Automatic Fare Collection (AFC) and Automated Vehicle Location (AVL) data. Firstly, we choose the passengers with the limited deterministic information as the research objects; secondly, we propose a classification-based method for identifying left behind passengers by the probabilistic model; next, the accuracy and effectiveness of the proposed method is verified by the simulation experiment and the case of Beijing Subway. Ultimately, the proposed method will support research related to the operation, management and future development of subways.

Keywords: Automatic Fare Collection (AFC); left behind; probabilistic model; subway; temporal distribution

1 INTRODUCTION

Subway systems have been well recognized as an effective, efficient, and sustainable transportation infrastructure in cities. In some megacities, subway systems can provide daily travel services for millions of passengers every day. There is a tidal wave phenomenon in passenger travel demand, with large concentrations of passengers during the morning and evening peak hours, which leads to overloading of the subway system. Even with minimum headway intervals, there are still passengers who actively or passively choose the next train (i.e. left behind) due to train capacity constraints or because they cannot tolerate the level of crowding in the carriages. It leads to higher journey times and reduces journey time reliability [1], which can have a significant impact on passenger satisfaction. Left behinds due to overcrowding has become a major concern for many subway operators [2-3]. To better deal with the increasing demand, operators are interested in better understanding the performance of the system and developing strategies to improve capacity utilization [4], such as accelerating frequency of departures, operating express trains and so on.

Identifying left behinds through manual surveys is time consuming and labor-intensive. As a result, it is difficult for operators to understand the performance of the system in real time. The availability of Automated Fare Collection (AFC) and Automated Vehicle Location (AVL) data affords the opportunity to monitor system operations and facilitate the development of relevant metrics to measure passenger experience. In recent years, it has been quickly adopted in the research of subway networks. For example, route choices can be approximately predicted [5], operation schedules can be optimized [6], the relationship between urban spatial layouts and passenger flows has been analyzed [7], and Close Contact passengers have been identified [8]. In general, AFC data can be fully leveraged in research as the data reflects the time and space information of every passenger entering and exiting a rapid transit system.

Zhao et al. [5] calculated the number of trains that a passenger waited for at his/her original station using maximum likelihood estimation based on AFC data. Zhu et al. [4] proposed a probability distribution model for the number of left behind passengers under congested conditions. Maximum likelihood estimation (MLE) and Bayesian estimation are used to estimate the probability function of stay-over based on a particular time period and station. Chen et al. [9] proposed an estimation method of the delayed-boarding probability distribution which was constructed by analyzing the relationship among passenger tap-in and tap-out time, access and egress time and the number of delayed-boarding times. Maximum likelihood estimation was used in the estimation of the distribution. They all assume that the probability of the number of trains needed to be waited for is stable in same time slot, but it does not correspond to the reality of the situation when there are inter-zone trains or express trains. Miller et al. [10] introduced the concept of cumulative capacity shortage (CCS) and proposed a bi-level regression model that transit agencies can use to estimate the number of passengers left behind on a platform by high-frequency trains operating at capacity. Sipetas et al. [11] developed the model to use emerging technologies for estimating the number of left behind passengers by processing video feeds from surveillance cameras in transit stations. They required an important data source, and that was passenger flow. These studies used the origin–destination–transfer inference model, which has some known drawbacks given existing limitations. It means that the above methods have certain limitations in application.

The objective of this paper is to propose an efficient method to identify left behind passengers from AFC and AVL data. The output of the model supports various applications:
1. Monitoring system operations and developing strategies to improve capacity utilization.
2. Developing performance metrics from the passenger's point of view and improving passenger satisfaction indicators.
3. Providing important input to some path choice estimation models, such as the left behind probabilities. It can also be used to analyze train-loading and crowding levels [12].

The remaining paper sections are organized as follows. Section 2 describes the problems studied in this paper. Section 3 proposes a probabilistic model to identify the left behind passengers. Section 4 verifies the accuracy of the method and discusses its practical application. Lastly, Section 5 concludes the paper.

2 PROBLEM DESCRIPTION

The AFC system usually records the transaction information of every trip across the rail network, such as the date, time, and stations where passengers enter and exit. AFC data also include ticket information, such as ticket numbers, ticket types, and fares associated with every trip information. Taking Beijing Subway as an example, the types of information in its AFC data include card ID, boarding time, boarding station, alighting time, alighting station, card type, reconciliation date, settlement date, automatic fare gate ID, payment value, and operating company. A sample of the key information of Beijing Subway is shown in Tab. 1.

The AFC system does not directly show the number of trains passengers actually took, so we do not know if passengers are left behind. We assume that a passenger enters at time t^{in} and exits at time t^{out} as Fig. 1 shows. Access time is defined as the time for the passenger to walk from the entry gate to the platform, waiting time is the time waiting on the platform, and egress time is the time to walk to the exit fare gate after alighting. In Fig. 1, train 1 could not meet the passenger's entry time, and train 4 could not meet the passenger's exit time. The passenger may board train 2 or train 3.

Table 1 AFC data key information of the Beijing subway

Name	Description	Sample
Encrypted Card ID	Encrypted card serial number	70***
Boarding time	Boarding time (year/month/day/hour/minute/second)	20180326 18:14:00
Boarding station ID	Identification details of the boarding station	150***
Alighting time	Card tag time of alighting	20180326 18:54:27
Alighting station ID	Identification details of alighting station	150***
Type of card	Ordinary card, senior card, student card, one-way ticket, etc.	Ordinary card
Reconciliation date	Time for reconciliation of accounts	20180326 00:00:00
Settlement date	Time for clearing currency receipts and payments	20180326 00:00:00
Automatic fare gate ID	The number assigned to every automatic fare gate	52***
Payment value	Deduction amount	**
Operating company	The number assigned to every operating company	**

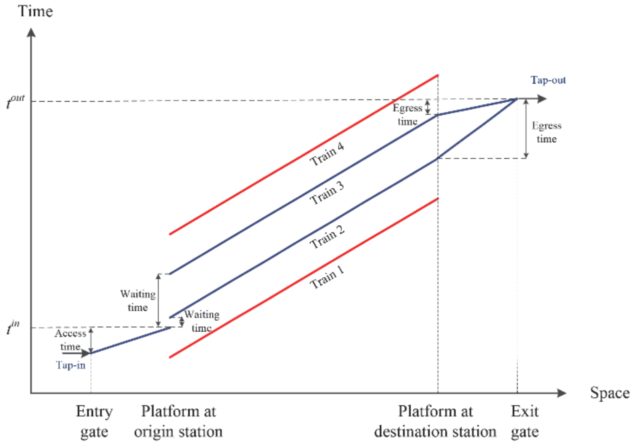

Figure 1 Time-space diagram for a journey without transfer

In this paper, we need to find out the train number that passengers took actually, so as to infer whether they were left behind.

3 METHOD

According to the statistics from existing research, there are a large number of passengers without transfer who have only one feasible path [13]. These passengers are defined as reference passengers, as a basis for estimating the distribution of access time and egress time [14], or inferring other passengers' travel paths [15]. Clearly, those trips are representative enough. We can take these passengers as a sample to infer the proportion of left behind passengers.

3.1 Definitions and Notations

In this section, we give some definitions and notations used in this paper, as listed in Tab. 2.

Table 2 Notations in this paper

Symbols	Description
O	The origin station.
D	The destination station.
t_i^{in}	The tap-in time of passenger *i*.
t_i^{out}	The tap-out time of passenger *i*.
$d_i^{(0)}$	The departure time of the previous train of train1.
$d_i^{(1)}$	The departure time of train1.
$a_i^{(1)}$	The arrival time of train1.
$a_i^{(2)}$	The arrival time of the next train of train1.
T_i^{a}	The access time of passenger *i*.
T_i^{w}	The waiting time of passenger *i*.
T_i^{e}	The egress time of passenger *i*.
A_O	The set of the access time plus waiting time at station *O*.
E_D	The set of the egress time at station *D*.
$T_O^{\min}$	The minimum access time at station *O*.
$T_D^{\min}$	The minimum egress time at station *D*.
F_i	The departure interval when passenger *i* tapped-in.

3.2 Minimum Access and Egress Time

When there is only one train that satisfies the passenger's tap-in and tap-out time, the passenger must take the train. Suppose passenger i entered station O at time t_i^{in} and exited station D at time t_i^{out}. If we can find train1 that satisfies Eq. (1), then train1 is the only feasible train taken by passenger i.

$$\begin{cases} d_i^{(0)} < t_i^{\text{in}} \\ a_i^{(2)} > t_i^{\text{out}} \end{cases} \tag{1}$$

where $d_i^{(0)}$ is the departure time of the previous train of train1, and $a_i^{(2)}$ is the arrival time of the next train of train1.

Then, we can get the access time, waiting time, and egress time of passenger i from Eq. (2).

$$\begin{cases} T_i^{\text{a}} + T_i^{\text{w}} = d_i^{(1)} - t_i^{\text{in}} \\ T_i^{\text{e}} = t_i^{\text{out}} - a_i^{(1)} \end{cases} \tag{2}$$

where T_i^{a} is the access time of passenger i, T_i^{w} is the waiting time of passenger i, $d_i^{(1)}$ is the departure time of train1, T_i^{e} is the egress time of passenger i, and $a_i^{(1)}$ is the arrival time of train1.

Suppose A_O is the set of the access time plus waiting time at station O from the passengers who have the only one feasible train, E_D is the set of the egress time at station D from these passengers. When the sum of T_i^{a} and T_i^{w} is minimized, T_i^{w} can be considered equal to zero. The minimum access time $T_O^{\min}$ at the station O is equal to $\min\left(A_O\right)$, and the minimum egress time $T_D^{\min}$ at the station D is equal to $\min\left(E_D\right)$.

The access and egress time distributions can be estimated from manual surveys conducted at stations or using AFC and AVL data [16].

In general, the access time follows a normal distribution [9]. The probability density function of the access time is

$$f_{\mathrm{a}}(t)=\frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-\frac{(t-\mu)^2}{2\sigma^2}} \tag{3}$$

In previous studies, the egress time follows a normal distribution [9] or Frechet distribution [15]. After investigating the actual data, we fit it with normal distribution, Frechet distribution, Extreme maximum distribution and Extreme minimum distribution respectively. The result shows that there is a significant difference in the egress behavior of passengers during peak hours and off-peak hours. In peak hours, the egress time follows an Extreme maximum distribution, while it follows a minimum value distribution in off-peak hours, as shown in Fig. 2 and Fig. 3. The probability density function of the egress time distribution is

$$f_{\mathrm{e}}(t)=\begin{cases}\dfrac{1}{\beta_1}\mathrm{e}^{\left(-\frac{t-\alpha_1}{\beta_1}-\mathrm{e}^{-\frac{t-\alpha_1}{\beta_1}}\right)}, t\in \text{ peak hours}\\ \dfrac{1}{\beta_2}\mathrm{e}^{\left(\frac{t-\alpha_2}{\beta_2}-\mathrm{e}^{\frac{t-\alpha_2}{\beta_2}}\right)}, t\in \text{ off-peak hours}\end{cases} \tag{4}$$

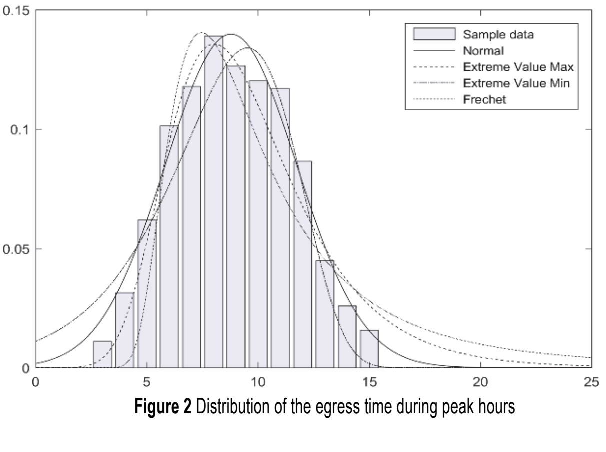

Figure 2 Distribution of the egress time during peak hours

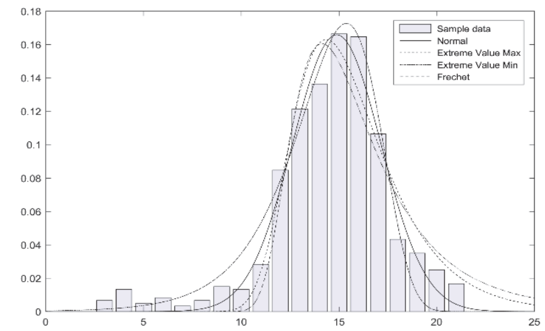

Figure 3 Distribution of the egress time during off-peak hours

3.3 Left Behind Probability Estimation

There are three kinds of different passengers, including passengers without left behind definitely, passengers with a unique itinerary, and passengers with indefinite itineraries. Hence different left behind probability models are established.

3.3.1 Passengers Without Left Behind Definitely

When there is only one train that satisfies the passenger's tap-in/out time and the minimum access/egress time, the passenger must take the train without left behind (i.e., if we can find train1 that satisfies Eq. (5), then train1 is the only feasible train taken by passenger i without left behind).

$$\begin{cases}d_i^{(0)}<t_i^{\mathrm{in}}+T_O^{\min}\\ a_i^{(2)}>t_i^{\mathrm{out}}-T_D^{\min}\end{cases} \tag{5}$$

3.3.2 Passengers with a Unique Itinerary

We assume that passengers will exit the station as soon as possible after getting off the train. In actual situations, some passengers may wait for their friends or go to the bathroom, they will not leave the station right away. According to statistics, the proportion of these passengers is less than 1.78%. Therefore, we believe that the error caused by this assumption is acceptable.

Passengers exit the station like a tide, as shown in Fig. 4. When the departure interval is greater than the confidence interval of the departure time distribution, the tap-out time of passengers from different trains does not overlap, and we can easily infer the train that a passenger got off (flagged as train1).

The previous train of train1 is train0. If $d_i^{(0)}\geq t_i^{\mathrm{in}}+T_O^{\min}$, we cannot confirm whether train0 had already departed when passengers arrived at the platform of original station. Let c_1 mean that train0 had already departed and c_2 that train0 had not departed. Then, the number of trains a passenger was left behind is

$$N_i=\begin{cases}\mathrm{INT}\left[\left(d_i^{(0)}-t_i^{\mathrm{in}}-T_O^{\min}\right)/F_i\right], \text{if } c_1 \text{ is ture}\\ \mathrm{INT}\left[\left(d_i^{(0)}-t_i^{\mathrm{in}}-T_O^{\min}\right)/F_i\right]+1, \text{if } c_2 \text{ is ture}\end{cases} \tag{6}$$

where N_i is the number of trains passenger i was left behind.

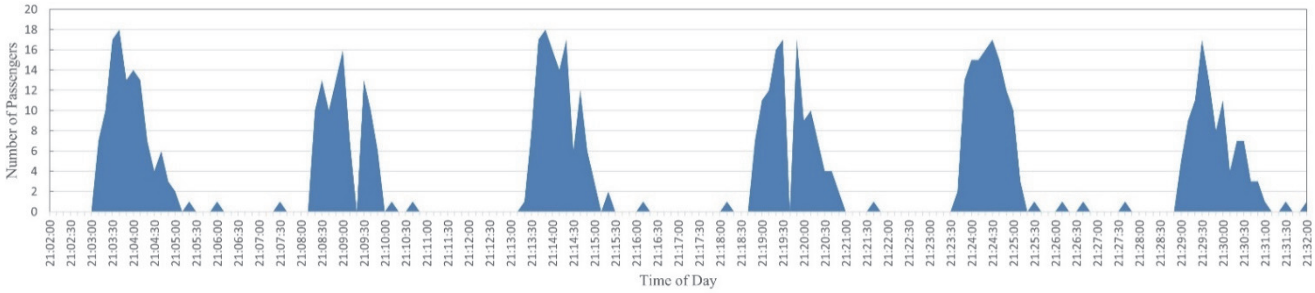

Figure 4 Tap-out time distribution of passengers at Haidian Wuluju Station during off-peak hours (departure interval is 6 minutes)

The probability of passenger i arriving at the origin station platform after the departures of train0 (i.e. c_1) is

$$P_i\left(c_1\right)=P_i\left(t_i^{\text{in}}+T_i^{\text{a}}>d_i^{(0)}\right)=1-\int_{T_O^{\min}}^{d_i^{(0)}-t_i^{\text{in}}} f_{\text{a}}(t) \tag{7}$$

The probability of passenger i arriving at the origin station platform before the departures of train0 (i.e. c_2) is

$$P_i\left(c_2\right)=P_i\left(t_i^{\text{in}}+T_i^{\text{a}}\leq d_i^{(0)}\right)=\int_{T_O^{\min}}^{d_i^{(0)}-t_i^{\text{in}}} f_{\text{a}}(t) \tag{8}$$

3.3.3 Passengers with Indefinite Itineraries

When the departure interval is less than the confidence interval of the departure time distribution, the tap-out time of passengers from different trains overlaps, as shown in Fig. 5. In the overlap, we cannot confirm whether the passengers in the overlap were from train1 with a longer egress time or from train2 with a shorter egress time. If $d_i^{(0)}\geq t_i^{\text{in}}+T_O^{\min}$, we also cannot confirm whether train0 had already departed when passengers arrived at the platform of original station.

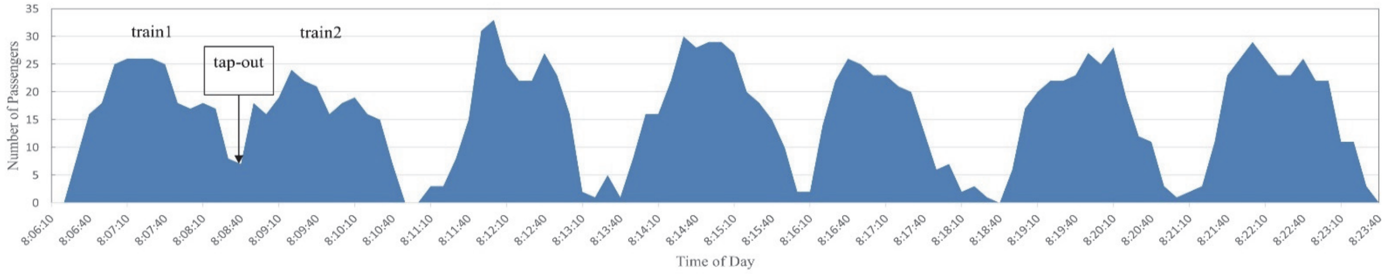

Figure 5 Tap-out time distribution of passengers at Haidian Wuluju Station during peak hours (departure interval is 2 minutes and 30 seconds)

There are four classes of passengers with indefinite itinerary:

C_1 - train0 had already departed when the passenger arrived at the platform of original station and the passenger took train1.

C_2 - train0 had not departed when the passenger arrived at the platform of original station and the passenger took train1.

C_3 - train0 had already departed when the passenger arrived at the platform of original station and the passenger took train2.

C_4 - train0 had not departed when the passenger arrived at the platform of original station and the passnger took train2.

Then, the number of trains a passenger was left behind is.

$$N_i=\begin{cases}\text{INT}\left[\left(d_i^{(0)}-t_i^{\text{in}}-T_O^{\min}\right)/F_i\right], & \text{if } C_1 \text{ is ture}\\ \text{INT}\left[\left(d_i^{(0)}-t_i^{\text{in}}-T_O^{\min}\right)/F_i\right]+1, & \text{if } C_2 \text{ is ture}\\ \text{INT}\left[\left(d_i^{(0)}-t_i^{\text{in}}-T_O^{\min}\right)/F_i\right]+1, & \text{if } C_3 \text{ is ture}\\ \text{INT}\left[\left(d_i^{(0)}-t_i^{\text{in}}-T_O^{\min}\right)/F_i\right]+2, & \text{if } C_4 \text{ is ture}\end{cases} \tag{9}$$

Suppose $P_i(\text{train}\,1)$ is the probability of passenger i taking train1 and $P_i(\text{train}\,2)$ is the probability of passenger i taking train2.

The probability of C_1 for passenger i is:

$$\begin{aligned}P_i\left(C_1\right)&=P_i\left(t_i^{\text{in}}+T_i^{\text{a}}>d_i^{(0)}\right)\cdot P_i(\text{train}\,1)=\\&=\left(1-\int_{T_0^{\min}}^{d_i^{(0)}-t_i^{\text{in}}} f_{\text{a}}(t)\right)\cdot\frac{f_{\text{e}}\left(t_i^{\text{out}}-a_i^{(1)}\right)}{f_{\text{e}}\left(t_i^{\text{out}}-a_i^{(1)}\right)+f_{\text{e}}\left(t_i^{\text{out}}-a_i^{(2)}\right)}\end{aligned} \tag{10}$$

The probability of C_2 for passenger i is:

$$\begin{aligned}P_i\left(C_2\right)&=P_i\left(t_i^{\text{in}}+T_i^{\text{a}}\leq d_i^{(0)}\right)\cdot P_i(\text{train}\,1)\\&=\int_{T_0^{\min}}^{d_i^{(0)}-t_i^{\text{in}}} f_{\text{a}}(t)\cdot\frac{f_{\text{e}}\left(t_i^{\text{out}}-a_i^{(1)}\right)}{f_{\text{e}}\left(t_i^{\text{out}}-a_i^{(1)}\right)+f_{\text{e}}\left(t_i^{\text{out}}-a_i^{(2)}\right)}\end{aligned} \tag{11}$$

The probability of C_3 for passenger i is:

$$\begin{aligned}P_i\left(C_3\right)&=P_i\left(t_i^{\text{in}}+T_i^{\text{a}}>d_i^{(0)}\right)\cdot P_i(\text{train}\,2)\\&=\left(1-\int_{T_0^{\min}}^{d_i^{(0)}-t_i^{\text{in}}} f_{\text{a}}(t)\right)\cdot\frac{f_{\text{e}}\left(t_i^{\text{out}}-a_i^{(2)}\right)}{f_{\text{e}}\left(t_i^{\text{out}}-a_i^{(1)}\right)+f_{\text{e}}\left(t_i^{\text{out}}-a_i^{(2)}\right)}\end{aligned} \tag{12}$$

The probability of C_4 for passenger i is

$$P_i\left(C_4\right)=P_i\left(t_i^{\text{in}}+T_i^{\text{a}}\le d_i^{(0)}\right)\cdot P_i(\text{train 2})$$
$$=\int_{T_0^{\min}}^{d_i^{(0)}-t_i^{\text{in}}} f_{\text{a}}(t)\cdot\frac{f_{\text{e}}\left(t_i^{\text{out}}-a_i^{(2)}\right)}{f_{\text{e}}\left(t_i^{\text{out}}-a_i^{(1)}\right)+f_{\text{e}}\left(t_i^{\text{out}}-a_i^{(2)}\right)} \tag{13}$$

4 CASE STUDY

4.1 Model Verification

A simulation experiment is used to verify the accuracy of the proposed method. Suppose there is a line in a ongested subway system (as shown in Fig. 6), 1000 passengers entered Station 2 in 30 minutes. Their tap-in time follows a uniform distribution while their access time and egress time follow a normal distribution and an extreme maximum distribution respectively.

Since Station 2 is not an origin station, the train already has passengers on board when it arrives. Suppose each train has an available capacity of 30-120 passengers randomly and the departure interval is 2 minutes. A first come first served discipline is used to load passengers onto trains with available capacity.

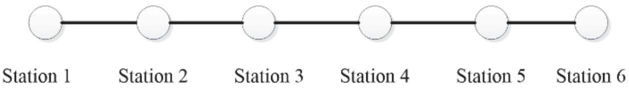

Figure 6 Network for simulation experiment

According to the simulated results (see Tab. 3), there were 428 left behind passengers (including 6 passengers being left behind 2 times) and 578 passengers without left behind.

Table 3 Simulated results

The train number	The available capacity of the train	The number of passengers arriving at the platform	The number of passengers without left behind	The number of left behind passengers
1	81	70	70	0
2	39	74	39	35
3	51	66	16	50
4	103	65	53	12
5	49	66	37	29
6	88	67	59	8
7	31	57	23	34
8	91	71	57	14
9	68	77	54	23
10	47	56	24	32
11	47	77	15	62
12	56	55	0	61
13	88	53	27	26
14	116	66	66	0
15	38	80	38	42
Total	993	1000	578	428

The proposed method was used to identify the left behind passengers based on the synthetic tap-in and tap-out time. The results show that 573 passengers were without left behind while 433 passengers were left behind (including 6 passengers being left behind 2 times). The Pearson correlation coefficient (r) between the estimated and simulated values was 0.998, the mean absolute error (MAE) was 0.733, the mean squared error (MSE) was 1.000, and the root mean squared error ($RMSE$) was 1.000 (see Fig. 7). The estimated values closely approximate the simulated values, which reflected the accuracy and effectiveness of the method.

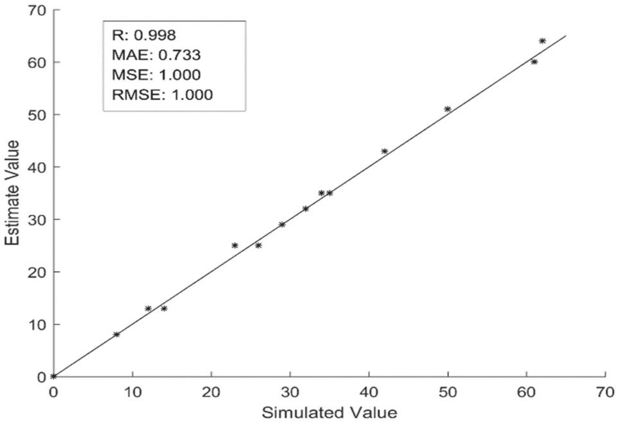

Figure 7 Comparison between simulated and estimated values

4.2 Application Case

Beijing Subway is used as an application case study in this research. At the end of 2021, Beijing Subway network consisted of 459 stations including 72 transfer stations and 27 lines with a total rail length of 783 km, becoming a veritable complex network system. The passengers who entered Chegongz Huang West Station on March 26 - 29. 2018 were analyzed.

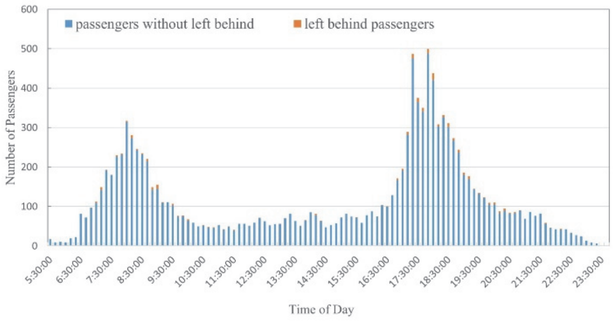

Figure 8 Number of left behind passengers at Chegongzhuang West Station (upward direction)

After calculation, in the upward direction of Chegongzhuang West Station, 261 of the 12287 passengers (2.1%) were left behind on March 26 - 29. 2018. This event occurred mainly in the morning and evening peak hours. There were 74 left behind passengers in the morning peak hours (7 am - 10 am) with a percentage of 2.3%, and 136 left behind passengers in the evening peak hours (5 pm - 8 pm) with a percentage of 2.7% (see Fig. 8). The higher percentage of passengers who were left behind during the evening peak hours may be because morning passengers do not want to be late and thus try to get on crowded trains, while they may choose less crowded trains after a busy day.

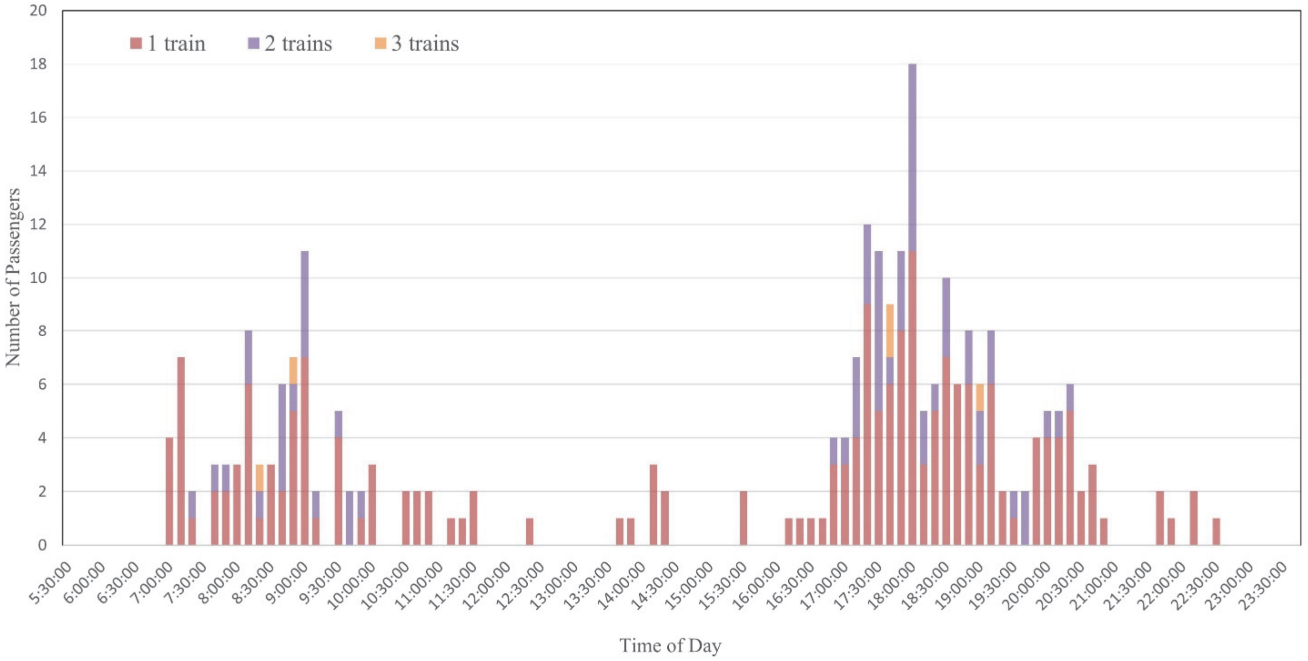

Figure 9 Number of consecutive trains a passenger was left behind at Chegongz huang West Station (upward direction)

Among the left behind passengers, 63 passengers were left behind on 2 consecutive trains, and 5 passengers were left behind on 3 consecutive trains (see Fig. 9). During the peak hours, 60 passengers were left behind on 2 consecutive trains with a percentage of 95.2% (31.7% were during the morning peak hours and 63.5% were during the evening peak hours). All cases of a passenger being left behind on 3 trains occurred during the peak hours. The results are consistent with the general cognition.

5 CONCLUSIONS

This paper proposed an efficient method to identify left behind passengers at subway stations using AFC and AVL data. Specifically, we developed such a method to infer the left behind probability of passengers without transfer. We divided these passengers into three categories: (1) passengers without left behind definitely; (2) passengers with a unique itinerary; (3) passengers with indefinite itineraries. The probabilistic model was used to infer the probability of passengers being left behind. The simulation experiment demonstrated that the estimated values closely approximate the simulated values.

In the coming research, we will apply the method to large-scale networks in different cities to further test its applicability.

Acknowledgements

This work was supported by Beijing Natural Science Foundation under Grant 9222025, Beijing Social Science Foundation under Grant 19JDGLA002, 18JDGLA018, 21JCB093, 21JCC092, MOE (Ministry of Education in China) Project of Humanities and Social Sciences under Grant 19YJC630043, the National Natural Science Foundation of China under Grant J1824031, the National Social Science Fund of China under Grant 21FGLB059 and was partially supported by Beijing Logistics Informatics Research Base. We appreciate their support very much. The work presented in this study remains the sole responsibility of the authors.

6 REFERENCES

[1] Zhang, T. (2019). *Research on route passenger flow allocation based on reliability of urban rail transit*. Beijing Jiaotong University.

[2] Delgado, F., Munoz, J. C., & Giesen, R. (2012). How much can holding and/or limiting boarding improve transit performance? *Transportation Research Part B: Methodological*, *46*(9), 1202-1217. https://doi.org/10.1016/j.trb.2012.04.005

[3] Sun, Y. & Xu, R. (2012). Rail Transit Travel Time Reliability and Estimation of Passenger Route Choice Behavior: Analysis Using Automatic Fare Collection Data. *Transportation Research Record*, *2275*(1), 58-67. https://doi.org/10.3141/2275-07

[4] Zhu, Y., Koutsopoulos, H. N., & Wilson, N. H. M. (2017). Inferring Left Behind Passengers in Congested Metro Systems from Automated Data. *Transportation Research Procedia, 23*, 362-379. https://doi.org/10.1016/j.trpro.2017.05.021

[5] Zhao, J., Zhang, F., Tu, L., Xu, C., Shen, D., Tian, C., Li, X. Y., & Li, Z. (2017). Estimation of Passenger Route Choice Pattern Using Smart Card Data for Complex Metro Systems. *IEEE Transactions on Intelligent Transportation Systems, 18*(4), 790-801. https://doi.org/10.1109/TITS.2016.2587864

[6] Sun, L., Jin, J. G., Lee, D. H., Axhausen, K. W., & Erath, A. (2014). Demand-driven timetable design for metro services. *Transportation Research Part C: Emerging Technologies, 46*, 284-299. https://doi.org/10.1016/j.trc.2014.06.003

[7] Huang, J., Levinson, D., Wang, J., Zhou, J., & Wang, Z. J. (2018). Tracking job and housing dynamics with smartcard data. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(50), 12710-12715. https://doi.org/10.1073/pnas.1815928115

[8] Xie, L. H., Zhang, Z. J., & Gong, D. Q. (2022). A Heuristic Close Contact Tracing Method for Urban Rail Transit. *Journal of Transportation Systems Engineering and Information Technology, 22*(4), 218-227.

[9] Chen, X., Luo, X., Zhu, Y., & Yuan, M. S. (2022). Delayed-Boarding Probability Distribution for Metro Stations Using Auto Fare Collection Data. *Journal of Southwest Jiaotong University*, *57*(2), 418-424.

[10] Miller, E., Sánchez-Martínez, G. E., & Nassir, N. (2018). Estimation of Passengers Left Behind by Trains in High-Frequency Transit Service Operating Near Capacity. *Transportation Research Record*, *2672*(8), 497-504. https://doi:10.1177/0361198118794291

[11] Sipetas, C., Keklikoglou, A., & Gonzales, E. J. (2020). Estimation of left behind subway passengers through archived data and video image processing. *Transportation Research Part C: Emerging Technologies*, *118*, 102727. https://doi.org/10.1016/j.trc.2020.102727

[12] Zhu, Y., Koutsopoulos, H. N., & Wilson, N. H. M. (2017). A probabilistic Passenger-to-Train Assignment Model based on automated data. *Transportation Research Part B: Methodological*, *104*, 522-542. https://doi.org/10.1016/j.trb.2017.04.012

[13] Hörcher, D., Graham, D. J., & Anderson, R. J. (2017). Crowding cost estimation with large scale smart card and vehicle location data. *Transportation Research Part B: Methodological*, *95*, 105-125. https://doi.org/10.1016/j.trb.2016.10.015

[14] Shi, J., Zhou, F., Zhu, W., & Xu, R. (2015). Estimation method of passenger route choice proportion in urban rail transit based on AFC data. *Journal of Southeast UniversityNatural Science Edition*, *45*(1), 184-188. https://doi.org/10.1155/2015/350397

[15] Hong, S. P., Min, Y. H., Park, M. J., Kim, K. M., & Oh, S. M. (2016). Precise estimation of connections of metro passengers from Smart Card data. *Transportation*, *43*(5), 749-769. https://doi.org/10.1007/s11116-015-9617-y

[16] Zhu, Y. (2014). Passenger-to-train assignment model based on automated data, *Master Thesis*, Massachusetts Institute of Technology.

Contact information:

Lianghui XIE
School of Economics and Management,
Beijing Jiaotong University,
No. 3 Shangyuancun, Haidian District, Beijing 100044, P. R. China
E-mail: 19113031@bjtu.edu.cn

Zhenji ZHANG
School of Economics and Management,
Beijing Jiaotong University,
No. 3 Shangyuancun, Haidian District, Beijing 100044, P. R. China
E-mail: zhjzhang@bjtu.edu.cn

Daqing GONG
(Corresponding author)
School of Economics and Management,
Beijing Jiaotong University,
No. 3 Shangyuancun, Haidian District, Beijing 100044, P. R. China
E-mail: dqgong@bjtu.edu.cn