Estimating left behind patterns in congested metro systems: a Bayesian model

Chao Yu and Haiying Li
Beijing Jiaotong University, Beijing, China

Xinyue Xu
State Key Lab of Rail Traffic Control and Safety, Beijing Jiaotong University, Beijing, China, and

Qi Sun
Beijing Metro Network Control Center, Beijing, China

Received 28 September 2020
Revised 22 January 2021
Accepted 22 January 2021

Abstract

Purpose – During rush hours, many passengers find it difficult to board the first train due to the insufficient capacity of metro vehicles, namely, left behind phenomenon. In this paper, a data-driven approach is presented to estimate left-behind patterns using automatic fare collection (AFC) data and train timetable data.

Design/methodology/approach – First, a data preprocessing method is introduced to obtain the waiting time of passengers at the target station. Second, a hierarchical Bayesian (HB) model is proposed to describe the left behind phenomenon, in which the waiting time is expressed as a Gaussian mixture model. Then a sampling algorithm based on Markov Chain Monte Carlo (MCMC) is developed to estimate the parameters in the model. Third, a case of Beijing metro system is taken as an application of the proposed method.

Findings – The comparison result shows that the proposed method performs better in estimating left behind patterns than the existing Maximum Likelihood Estimation. Finally, three main reasons for left behind phenomenon are summarized to make relevant strategies for metro managers.

Originality/value – First, an HB model is constructed to describe the left behind phenomenon in a target station and in the target direction on the basis of AFC data and train timetable data. Second, a MCMC-based sampling method Metropolis–Hasting algorithm is proposed to estimate the model parameters and obtain the quantitative results of left behind patterns. Third, a case of Beijing metro is presented as an application to test the applicability and accuracy of the proposed method.

Keywords MCMC, AFC data, Hierarchical Bayesian model, Left behind, Metro system

Paper type Research paper

This work was supported by the State Key Lab of Rail Traffic Control and Safety of China (No. RCS2020ZT003).

Smart and Resilient Transportation
Vol. 3 No. 2, 2021
pp. 149-161
Emerald Publishing Limited
e-ISSN: 2632-0495
p-ISSN: 2632-0487
DOI 10.1108/SRT-09-2020-0008

1. Introduction

In recent years, metro has been favored by more and more urban residents due to its advantages of fast speed, large volume, punctuality and low fares (Silva *et al.*, 2015; Noursalehi *et al.*, 2018). Taking several big cities in China as examples, in 2018, the average daily passenger volume of Beijing, Shanghai and Guangzhou metro has reached 10.172, 10.544 and 8.354 million, respectively. As a result, the construction speed of metro system is difficult to keep up with the growth of passengers' travel demand, which leads to many metro service problems (Xu *et al.*, 2016, 2018b). The most common issue is the phenomenon known as left behind, where passengers cannot get on the first train because of insufficient train capacity, especially during rush hours (Zhu *et al.*, 2018, 2017). Nowadays, the left behind phenomenon has attracted the attention of many metro operators because of its negative effects, such as potential safety risks due to crowded platforms, and an inaccurate estimate of the loading rate of trains (Sun and Xu, 2012; Delgado *et al.*, 2012; Mueller and Sgouridis, 2011). Therefore, it is of substantial importance in studying and quantifying this phenomenon.

Many researchers have investigated this phenomenon using ideal models. For example, Hamdouch *et al.* (2011) and Papola *et al.* (2009) described the boarding process of metro passengers using the First-Come-First-Serve (FCFS) principle. When the train vehicle reaches the maximum capacity, passengers who cannot get on the train are automatically identified as passengers left behind. However, the reality may differ from the above description, for instance, passengers may not strictly follow the FCFS principle when boarding the train and the maximum capacity of the train may be exceeded in rush hours (Xu *et al.*, 2018a).

Recently, automatic fare collection (AFC) that records a large amount of passengers' information provides a possibility for researchers to quantify the left behind phenomenon using data-driven methods. There are two main thoughts for AFC data to analyze the behavior characteristics: Frequentist and Bayesians. In view of Frequentist, a parameter in the model is regarded as an unknown but fixed value. For example, using AFC data, Zhao *et al.* (2017) proposed a method based on maximum likelihood estimation (MLE)to study the left behind patterns of Shenzhen metro system, in which the left behind patterns are seen as a vector of several fixed values. But for Bayesians, a parameter in the model follows a particular probability distribution, which is related to the prior information of this parameter and observed data. Bayesian inference framework has been got more and more attention by scholars in the field of passenger behavioral. Zhu *et al.* (2018) proposed both Bayesian inference methods and MLE method to learn the probabilistic mass function of the left behind phenomenon and showed that the Bayesian method makes better use of the advantages of observed data and prior information to obtain more accurate calculation results. Barry *et al.* (2007) presented a Bayesian model to estimate the distribution of origin-destination stations in New York City by using MetroCard information. Fu (2014) developed a Gaussian mixture model (GMM) combined with naive Bayesian framework to calculate the route choice patterns using the AFC data of London Underground. Sun *et al.* (2015) proposed an integrated Bayesian approach to study the passenger flow assignment in the metro network, and the case of Singapore metro proved the validity of proposed method. Li *et al.* (2018) proposed a Bayesian network model to capture the characteristics of departure time choice, whereas factors such as travel time saving, crowding, fare and departure time change are considered.

To take full use of the advantages of Bayesian inference framework, a more complex hierarchical Bayesian (HB) model is further presented to analyze the behavior characteristics of passengers. Lee and Sohn (2015) developed a HB model that incorporate several route-use patterns as unknown parameters into a Bayesian framework to research

route choice behaviour and illustrated the superiority of the method through Bayesian information criterion (BIC). Rahbar *et al.* (2019) presented a three-level HB model to estimate attributes of travel time components and to calibrate a transit assignment model. However, the HB model is so complex that it is difficult for traditional optimization methods to calculate the maximum likelihood function in the model (Robert, 2013). As a result, scholars calibrate the parameters of these models using computational methods, including the most widely used method MCMC (Pereyra *et al.*, 2016). Xu *et al.* (2018c) and Sun *et al.* (2015) applied Metropolis–Hasting (MH) sampling algorithm to calibrate the parameters based on Bayesian inference framework and examined that this method has great performance even if there are multi-dimensional parameters,. However, the above model and related estimating algorithm have been not applied in the field of left behind patterns. Thus, we will construct a HB model to describe the left behind phenomenon and determine the values of parameters of each layer model using MCMC method based on AFC data.

The main object of this paper is to study left behind patterns in metro system using a large number of passengers' travel records. First, we construct an HB model to describe the left behind phenomenon in a target station and in the target direction on the basis of AFC data and train timetable data. Second, we propose a MCMC-based sampling method MH algorithm to estimate the model parameters and obtain the quantitative results of left behind patterns. Third, a case of Beijing metro is presented as an application to test the applicability and accuracy of the proposed method. Finally, three main causes of left behind are summarized to deeply understand the mechanism of the above phenomenon and to make relevant passenger flow control strategies.

2. Methodology

In this section, a Bayesian inference framework is introduced to estimate left behind patterns in metro system and the processing flowchart is shown in Figure 1, more details are provided below.

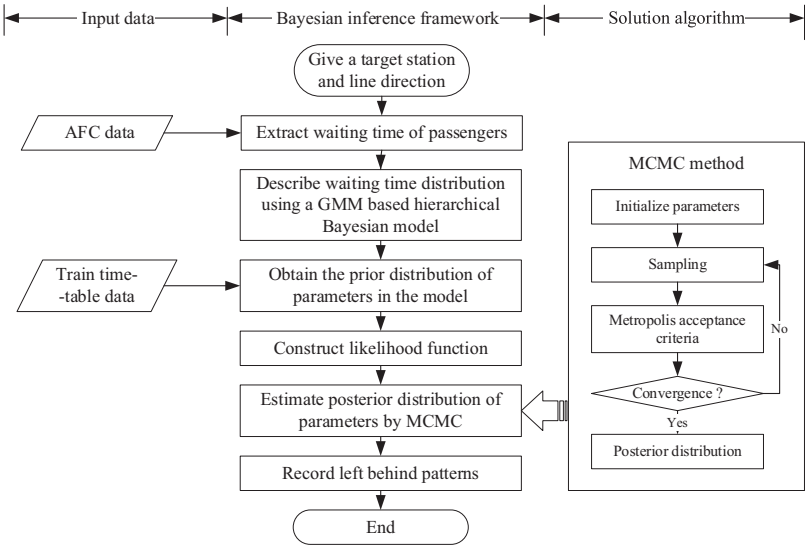

Figure 1.
Processing flowchart

2.1 Data preprocessing and notations

We define the metro network as (N, A), where N is the set of nodes, representing metro stations and A is the set of arcs, representing the links between two adjacent stations (Rahbar *et al.*, 2019). For a given OD pair (o, d), o, $d \in N$, there might be several routes in its feasible route set $\Omega_{o,d}$, which can be obtained by K-shortest algorithm. In this paper, only the AFC records on OD pair with $|\Omega_{o,d}|$=1 and no transfer process are used, so as to obtain the accurate waiting time of passengers at the origin station (Zhao *et al.*, 2017). Note that, for each OD pair, the boxplot method is proposed to mine the AFC data. The data with too long or too short travel time is considered as invalid data and will be removed. The notations and definitions used in this paper are listed in Table 1.

2.2 Extracting waiting time distribution from automatic fare collection data

According to the research by (Chakirov and Erath, 2011), the passengers, who have the shortest travel time and have a waiting time of zero (i.e. $wt^z_{o,d} = 0$), can be seen as the benchmark to calculate the waiting time of the other passengers given an OD pair. As there is no transfer process in the feasible route of OD pairs used in this paper, the difference of travel time between the target passenger and the benchmark passenger can be seen as his/her waiting time as follows:

$$wt^z_{o,d} = tt^z_{o,d} - t^{\min}_{o,d} \tag{1}$$

Notation	Definition
Sets	
L	Set of lines indexed by l, with two directions: up (u) and down (d) direction, $l \in L$
S	Set of stations indexed by s
W	Set of OD pairs with $\|\Omega_{o,d}\| = 1$ and no transfer process, indexed by$w$
D_l	Set of line directions of line l, $l \in L$, $D_l \in \{l^u, l^d\}$
H	Set of trains indexed by h
P	Set of passengers indexed by p
$P_{o,d}$	Set of passengers on the OD pair (o, d), (o, d) $\in W$
Parameters	
$l_{o,d}$	The line of feasible route on the OD pair (o, d), (o, d) $\in W$
$dl_{o,d}$	Direction of $l_{o,d}$, $dl_{o,d} \in D_{l_{o,d}}$
$hw_{o,d}$	Headway of line $l_{o,d}$ in direction $dl_{o,d}$
$t^z_{ac,o}$	Tap-in timestamp at origin station o of passenger z, $z \in P_{o,d}$
$t^z_{eg,d}$	Tap-out timestamp at destination station d of passenger z, $z \in P_{o,d}$
$tt^z_{o,d}$	Travel time of passenger z on OD pair (o, d), $z \in P_{o,d}$, (o, d) $\in W$, $tt^z_{o,d} = t^z_{eg,d} - t^z_{ac,o}$
$tt^{\min}_{o,d}$	The minimum travel time on OD pair (o, d), (o, d) $\in W$
$wt^z_{o,d}$	Waiting time of passenger z at origin station o in line direction $dl_{o,d}$, $z \in P_{o,d}$, $o \in S$
$WT_{o,d}$	The set of$wt^z_{o,d}$, $WT_{o,d} = \cup\left(wt^z_{o,d} \middle\| z \in P_{o,d}\right)$
Estimation variables	
$k_{o,d}$	The maximum number of trains that passengers need to wait for at origin station o in direction $dl_{o,d}$, $o \in S$
$\mu_{o,d}/\sigma_{o,d}/\omega_{o,d}$	Estimated mean/standard deviation/weight of waiting time of passengers who boarding ith train at origin station o in direction $dl_{o,d}$, $o \in S$

Table 1.
Notations and definitions

The waiting time of a single passenger can be calculated by the above method, which can to be generalized to the all passengers whose origin station is the target station in direction $dl_{o,d}$, so as to analyze the left behind patterns of the station in this direction.

2.3 Modelling passengers' waiting time using Bayesian theorem

In this section, a hierarchical Bayesian model is proposed to describe characteristics of left behind phenomenon in the target station using passengers' waiting time. Specifically, the waiting time distribution is represented as a GMM model. The details are shown as follows.

We use an illustrative example to show the modelling process (as shown in Figure 2). In a given line direction, trains arrive at the target station periodically according to the predefined train timetable [Figure 2(a)]. Due to left behind phenomenon, passengers waiting on the platform may successfully board the 1st train, the 2nd train, the 3rd train and the 4th train and have to suffer from their corresponding waiting time [Figure 2(b)]. Passengers' waiting time follows a multi-peaked distribution and can be expressed as GMM suggested by (Fu, 2014)(Chakirov and Erath, 2011). More specifically, each component in GMM (one-dimensional Gaussian function) can be regarded as the waiting time distribution of passengers boarding the ith train, and its corresponding mean ($\mu_{o,d}$) and weight ($\omega_{o,d}$) represent the average waiting time of such passengers and the proportion of such passengers in all waiting passengers, respectively [Figure 2(c)]. Furthermore, the difference between the mean of two adjacent components is approximately equal to the headway ($hw_{o,d}$) in direction $dl_{o,d}$. Thus, the number of the trains that passenger z from o to d have to wait for [i.e. $l(z)$] at station o can be obtained as follow:

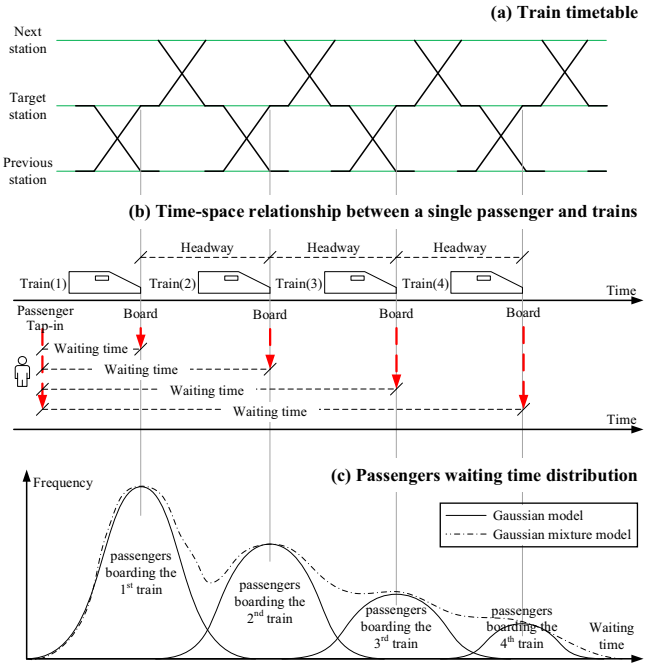

Figure 2.
Relationship between left behind phenomenon and passengers' waiting time distribution

$$l(z)=\begin{cases}1,\ 0\le wt^{z}_{o,d}<hw_{o,d}\\ 2,\ hw_{o,d}\le wt^{z}_{o,d}<2hw_{o,d}\\ \ldots..\\ k_{o,d},\ (k_{o,d}-1)\cdot hw_{o,d}\le wt^{z}_{o,d}<k_{o,d}\cdot hw_{o,d}\end{cases},z\in P_{o,d} \tag{2}$$

where $hw_{o,d}$ can be calculated by train timetable, that is, the total length of the target period divided by the total number of trains in direction $dl_{o,d}$ within the period; $k_{o,d}$ can be determined by the passengers with maximum waiting time. Therefore, the waiting time of passenger z can be calculated by fusion $k_{o,d}$ one-dimensional Gaussian functions as follow:

$$p\left(wt^{z}_{o,d}|\boldsymbol{\omega}_{o,d},\boldsymbol{\mu}_{o,d},\boldsymbol{\sigma}_{o,d}\right)=\sum_{i=1}^{k_{o,d}}\left(\omega^{i}_{o,d}\frac{1}{\sqrt{2\pi}\cdot\sigma^{i}_{o,d}}\exp\left(\frac{-\left(wt^{z}_{o,d}-\mu^{i}_{o,d}\right)^{2}}{2{\sigma^{i}_{o,d}}^{2}}\right)\right) \tag{3}$$

where $\boldsymbol{\omega}_{o,d}=\left(\omega^{1}_{o,d},\omega^{2}_{o,d},\ldots,\omega^{k_{o,d}}_{o,d}\right)$ is the vector of the weight of waiting time of passengers who board the ith train; $\boldsymbol{\mu}_{o,d}=\left(\mu^{1}_{o,d},\mu^{2}_{o,d},\ldots,\mu^{k_{o,d}}_{o,d}\right)$ and $\boldsymbol{\sigma}_{o,d}=\left(\sigma^{1}_{o,d},\sigma^{2}_{o,d},\ldots,\sigma^{k_{o,d}}_{o,d}\right)$ represent corresponding vector of the mean and standard deviation, respectively.

Further, the unknown parameters (i.e. $\boldsymbol{\omega}_{o,d}$, $\boldsymbol{\mu}_{o,d}$ and $\boldsymbol{\sigma}_{o,d}$) are estimated by Bayesian theorem, Note that the unknown parameters are regarded as random variables sampled from a probability distribution that is called hyper-prior constructed by several hyper-parameters (Lee and Sohn, 2015). The corresponding hyper-prior distributions are listed as equations (4)–(6): $\mu^{i}_{o,d}$ is regarded conforming to Gaussian distribution with hyper-parameters $\delta^{i}_{o,d},\nu^{i}_{o,d}$; $\sigma^{i}_{o,d}$ follows a Uniform distribution with hyper-parameters $\kappa^{i}_{o,d}$, $\gamma^{i}_{o,d}$; and $\boldsymbol{\omega}_{o,d}$ is regarded as conforming to Dirichlet distribution with hyper-parameters $\omega^{1}_{o,d},\omega^{2}_{o,d},\ldots,\omega^{k_{o,d}}_{o,d}$, as well as $\sum_{i=1}^{k_{o,d}}\omega^{i}_{o,d}=1$.

$$\mu^{i}_{o,d}\sim Gaussian\left(\delta^{i}_{o,d},\nu^{i}_{o,d}\right) \tag{4}$$

$$\sigma^{i}_{o,d}\sim Uniform\left(\kappa^{i}_{o,d},\gamma^{i}_{o,d}\right) \tag{5}$$

$$\boldsymbol{\omega}_{o,d}\sim Dirichlet\left(\omega^{1}_{o,d},\omega^{2}_{o,d},\ldots,\omega^{k_{o,d}}_{o,d}\right) \tag{6}$$

Refer to the time-space relationship between passengers waiting time and trains in Figure 2, we set $\delta^{i}_{o,d}=(i-0.5)\cdot hw_{o,d}$. According to Bayesian theorem, the joint posterior probability of unknown parameters can be formulated as follow:

$$\begin{aligned}p\left(\boldsymbol{\omega}_{o,d},\boldsymbol{\mu}_{o,d},\boldsymbol{\sigma}_{o,d}|WT_{o,d}\right)&=\frac{p\left(WT_{o,d}|\boldsymbol{\omega}_{o,d},\boldsymbol{\mu}_{o,d},\boldsymbol{\sigma}_{o,d}\right)p(\boldsymbol{\omega}_{o,d},\boldsymbol{\mu}_{o,d},\boldsymbol{\sigma}_{o,d})}{p\left(WT_{o,d}\right)}\\&\propto p\left(WT_{o,d}|\boldsymbol{\omega}_{o,d},\boldsymbol{\mu}_{o,d},\boldsymbol{\sigma}_{o,d}\right)p(\boldsymbol{\omega}_{o,d},\boldsymbol{\mu}_{o,d},\boldsymbol{\sigma}_{o,d})\end{aligned} \tag{7}$$

where $p(\boldsymbol{\omega}_{o,d}, \boldsymbol{\mu}_{o,d}, \boldsymbol{\sigma}_{o,d})$ is the joint prior probability density function of unknown parameters, and $p(WT_{o,d}|\boldsymbol{\omega}_{o,d}, \boldsymbol{\mu}_{o,d}, \boldsymbol{\sigma}_{o,d})$ is the likelihood function, $WT_{o,d}$ is the observed data and $p(WT_{o,d})$ represents the probability of observed data. Note that $p(WT_{o,d})$ is a fixed value and can be neglected during the process of Bayesian inference.

Each unknown parameter is treated as a random variable sampled from a certain distribution, so they are independent of each other, which means the prior distribution can be formulated combined with equations (4)–(6) as follows:

$$\begin{aligned} & p(\boldsymbol{\omega}_{o,d}, \boldsymbol{\mu}_{o,d}, \boldsymbol{\sigma}_{o,d}) \\ &= p(\boldsymbol{\omega}_{o,d}) p(\boldsymbol{\mu}_{o,d}) p(\boldsymbol{\sigma}_{o,d}) \\ &= p\left(\omega_{o,d}^1, \omega_{o,d}^2, \ldots, \omega_{o,d}^{k_{o,d}}\right) \left[\prod_{i=1}^{k_{o,d}} p\left(\mu_{o,d}^i | \delta_{o,d}^i, \nu_{o,d}^j\right) p\left(\delta_{o,d}^i, \nu_{o,d}^j\right)\right] \left[\prod_{i=1}^{k_{o,d}} p\left(\sigma_{o,d}^i | \kappa_{o,d}^i, \gamma_{o,d}^j\right) p\left(\kappa_{o,d}^i, \gamma_{o,d}^j\right)\right] \end{aligned} \tag{8}$$

If each passenger's travel process is independent of each other, the likelihood function of waiting time of passengers can be expressed as follows:

$$\begin{aligned} p\left(WT_{o,d} | \boldsymbol{\omega}_{o,d}, \boldsymbol{\mu}_{o,d}, \boldsymbol{\sigma}_{o,d}\right) &= \prod_{wt_{o,d}^z \in WT_{o,d}} p\left(wt_{r,o,d}^z | \boldsymbol{\omega}_{o,d}, \boldsymbol{\mu}_{o,d}, \boldsymbol{\sigma}_{o,d}\right) \\ &= \prod_{wt_{o,d}^z \in WT_{o,d}} \left[\sum_{i=1}^{k_{o,d}} p\left(wt_{o,d}^z | \omega_{o,d}^i, \mu_{o,d}^i, \sigma_{o,d}^i\right) p\left(\omega_{o,d}^i, \mu_{o,d}^i, \sigma_{o,d}^i\right)\right] \end{aligned} \tag{9}$$

Combine the prior distribution [equation (8)] and likelihood function [equation (9)] with the Bayesian theorem [equation (7)] and then get the final two-level hierarchical Bayesian inference formula as follows:

$$\begin{aligned} & p\left(\boldsymbol{\omega}_{o,d}, \boldsymbol{\mu}_{o,d}, \boldsymbol{\sigma}_{o,d} | WT_{o,d}\right) \propto \prod_{wt_{o,d}^z \in WT_{o,d}} \left[\sum_{i=1}^{k_{o,d}} p\left(wt_{o,d}^z | \omega_{o,d}^i, \mu_{o,d}^i, \sigma_{o,d}^i\right) p\left(\omega_{o,d}^i, \mu_{o,d}^i, \sigma_{o,d}^i\right)\right] \\ & \cdot p\left(\omega_{o,d}^1, \omega_{o,d}^2, \ldots, \omega_{o,d}^{k_{o,d}}\right) \left[\prod_{i=1}^{k_{o,d}} p\left(\mu_{o,d}^i | \delta_{o,d}^i, \nu_{o,d}^j\right) p\left(\delta_{o,d}^i, \nu_{o,d}^j\right)\right] \left[\prod_{i=1}^{k_{o,d}} p\left(\sigma_{o,d}^i | \kappa_{o,d}^i, \gamma_{o,d}^j\right) p\left(\kappa_{o,d}^i, \gamma_{o,d}^j\right)\right] \end{aligned} \tag{10}$$

Given the observed data $WT_{o,d}$, the posterior distribution of the $3k_{o,d}$ unknown parameters $(\mu_{o,d}^1, \sigma_{o,d}^1, \omega_{o,d}^1, \mu_{o,d}^2, \sigma_{o,d}^2, \omega_{o,d}^2, \ldots, \mu_{o,d}^{k_{o,d}}, \mu_{o,d}^{k_{o,d}}, \mu_{o,d}^{k_{o,d}})$ can be estimated using equation (10). Next, we will propose an MCMC method MH algorithm to estimate unknown parameters in the next subsection.

2.4 Estimating parameters using Markov Chain Monte Carlo

It is difficult to calculate the maximum likelihood by traditional optimization methods due to the complex formulation of parameters and the high-dimensional characteristics of parameter space. Therefore, researchers attempt to use the prior information of parameters and update the value of parameters through iteration until the value of likelihood function is

the maximum to address this problem, namely, MCMC method. Many kinds of MCMC algorithms have been developed and the most used algorithm (i.e. MH algorithm) is applied to estimate the posterior distribution of target parameters in this paper. All the $3k_{o,d}$ unknown parameters are synthesized into a vector $\boldsymbol{\psi} = (\psi_1, \psi_2, \psi_3, \ldots, \psi_{3k_{o,d}})$, and the estimation process using MH algorithm is as follows:

Step 1: Set iterative parameter $t = 1$, and set maximum number of iterations T, and set Burn-in value B, which is a fixed number to remove unstable sampling values of parameters.

Step 2: Initialize parameter $\boldsymbol{\psi}^{(t)} = \left({\psi_1}^{(t)}, {\psi_2}^{(t)}, {\psi_3}^{(t)}, \ldots, {\psi_{3k_{o,d}}}^{(t)}\right)$ according to the prior distribution shown in equations (4)–(6).

Step 3: Sampling each unknown parameters, set $i = 1$.

Step 3.1: Generate a candidate state ${\psi_i}^*$ from the presented probability distribution function $p\left({\psi_i}^* | {\psi_i}^{(t-1)}\right)$, that is, the original state vector is $\boldsymbol{\psi}^{(t-1)} = \left({\psi_1}^{(t-1)}, \ldots, {\psi_i}^{(t-1)}, \ldots, {\psi_{3k_{o,d}}}^{(t-1)}\right)$, and the candidate state vector is $\boldsymbol{\psi}^* = \left({\psi_1}^{(t-1)}, \ldots, {\psi_i}^*, \ldots, {\psi_{3k_{o,d}}}^{(t-1)}\right)$.

Step 3.2: Calculate the acceptance probability α according to MH criterion:

$$\alpha = \min\left\{1, \frac{p(\boldsymbol{\psi}^*)}{p\left(\boldsymbol{\psi}^{(t-1)}\right)} \cdot \frac{p\left({\psi_i}^{(t-1)} | {\psi_i}^*\right)}{p\left({\psi_i}^* | {\psi_i}^{(t-1)}\right)}\right\} \tag{10}$$

Step 3.3: Generate a random number u from a uniform distribution (0, 1).

Step 3.4: Determines whether the original state is updated to the candidate state: if $\alpha > u$, then set ${\psi_i}^{(t)} = {\psi_i}^*$; otherwise, set ${\psi_i}^{(t)} = {\psi_i}^{(t-1)}$.

Step 3.5: Determines whether all unknown parameters have been calculated: if $i < 3k_{o,d}$, set $i = i + 1$ and return to Step 3.1; otherwise, record the current state vector $\boldsymbol{\Psi}^{(t)}$.

Step 4: Determine whether to stop sampling: if $t < T$, set $t = t + 1$ and return to Step 3; otherwise, stop sampling and estimate each parameter as follows:

$$\psi_i = \sum_{t=B+1}^{T} {\psi_i}^{(t)} / (T - B) \tag{11}$$

3. Case study

In this section, a suburban metro line case of Beijing metro system is introduced to verify the performance of the proposed method in estimating the left behind phenomenon. The AFC data and train timetable from September 1, 2018 to October 1, 2018 are collected from Beijing Metro Network Control Center and about 176 million records are collected. All cases are implemented using Python 3.7.1 and Oracle databases in this section.

3.1 Case description and parameters settings

3.1.1 Introduction to Changping Line. The left behind patterns of stations in Changping Line is mainly studied. As shown in Figure 3, the Changping Line is a suburban line located northwest of Beijing metro system and consists of 12 stations. Around each station, there are large numbers of commuters who tend to take the metro to work in the urban central area during the morning rush hours (7 a.m.–9 a.m.). Therefore, the left behind patterns of

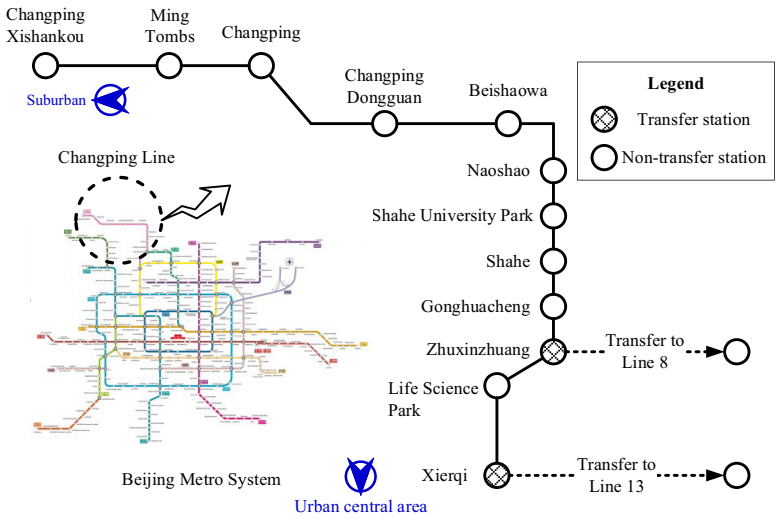

Figure 3.
Illustration of Changping Line

each station in the direction of "suburban-urban center area" during 7 a.m.–9 a.m. in workdays are studied in this section, which is of great significance to be aware of the experience of these metro passengers.

3.1.2 Parameter settings in the model. First, according to the train timetable, the average headway in Changping Line and in target direction can be obtained, that is $hw = 268$. And the maximum number of trains passengers have to wait for (k) can be calculated by using passengers' trip records and the headway according to equation (2). And then, we can determine the value of hyper-parameters in equations (4)–(6)based on the hw and k. In this case, we set $\delta^i = (i - 0.5)\,268$, $\nu^i = 10$, $\kappa^i = 0$, $\gamma^i = 268$, and $\omega^i = 1/k$, $i = 1,2,\ldots,k$. Finally, the parameters in MH algorithm [see equation (11)] is given as $T = 15000$ and $B = 9000$.

3.2 Estimation results and analysis

The estimation results are shown in Figure 4, where each picture represents the estimation results of the left behind patterns of corresponding station, in which the horizontal coordinate represents the waiting time of passengers and the vertical coordinate represents the frequency. And the estimated results of each parameter (μ, σ, ω) are listed at the bottom of each picture. To verify the accuracy of the results obtained by the proposed method, an MLE based method (see reference (Zhao *et al.*, 2017)) is also used to calculate the left behind patterns of target stations. And the comparison of results of MLE-based method and proposed method is shown in Table 2. The comparison results show that the results obtained by the two methods are similar, indicating that the proposed method performs well in estimating left behind patterns (Zhu *et al.*, 2018). However, the MLE-based method can only give the proportion of passengers boarding different trains, while the proposed method (HB + MCMC) can give more details, such as the average waiting time and standard deviation of passengers boarding different trains, which has a better explanatory characteristic.

Furthermore, according to the maximum number of trains passengers have to wait for (reflecting the severity of the left behind phenomenon), stations are divided into three grades: light congestion station, normal congestion) station and heavy congestion station,

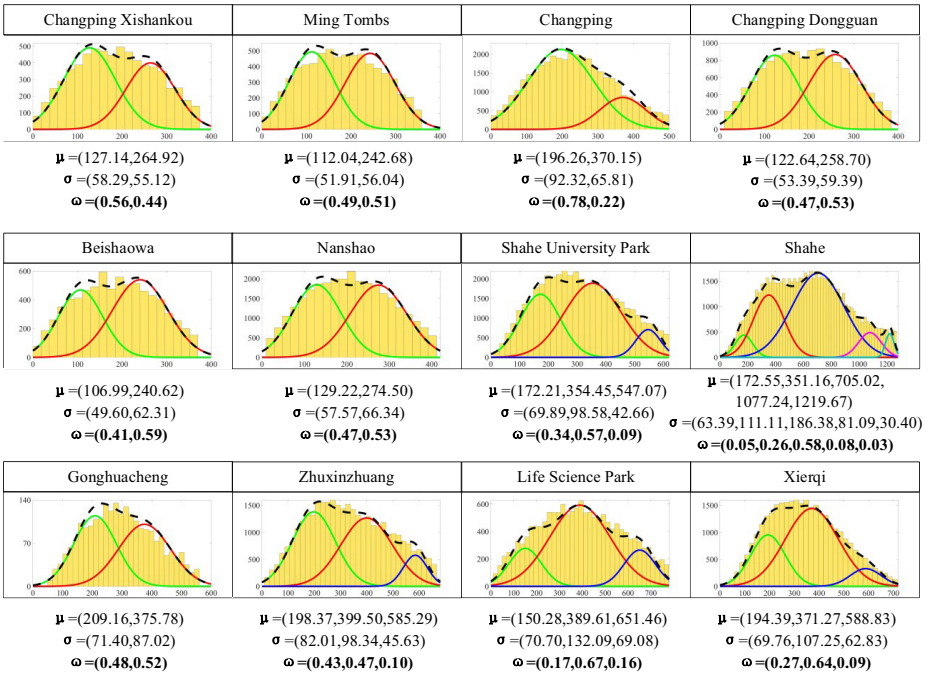

Figure 4.
Calculation results of left behind patterns of stations in Changping Line

Station	Method	1st train	2nd train	3rd train	4th train	5th train
Changping Xishankou	MLE	0.7919	0.2081	0.0000	0.0000	0.0000
	HB+MCMC	0.5649	0.4351	0.0000	0.0000	0.0000
Ming Tombs	MLE	0.8239	0.1761	0.0000	0.0000	0.0000
	HB+MCMC	0.4855	0.5145	0.0000	0.0000	0.0000
Changping	MLE	0.6277	0.3723	0.0000	0.0000	0.0000
	HB+MCMC	0.7822	0.2178	0.0000	0.0000	0.0000
Changping Dongguan	MLE	0.7683	0.2317	0.0000	0.0000	0.0000
	HB+MCMC	0.4710	0.5290	0.0000	0.0000	0.0000
Beishaowa	MLE	0.7985	0.2015	0.0000	0.0000	0.0000
	HB+MCMC	0.4089	0.5911	0.0000	0.0000	0.0000
Nanshao	MLE	0.7161	0.2839	0.0000	0.0000	0.0000
	HB+MCMC	0.4657	0.5343	0.0000	0.0000	0.0000
Shahe University Park	MLE	0.4218	0.5064	0.0718	0.0000	0.0000
	HB+MCMC	0.3416	0.5666	0.0917	0.0000	0.0000
Shahe	MLE	0.1142	0.2924	0.3127	0.1998	0.0809
	HB+MCMC	0.0528	0.2586	0.5863	0.0750	0.0273
Gonghuacheng	MLE	0.4325	0.5675	0.0000	0.0000	0.0000
	HB+MCMC	0.4826	0.5174	0.0000	0.0000	0.0000
Zhongxinzhuang	MLE	0.3842	0.4951	0.1207	0.0000	0.0000
	HB+MCMC	0.4284	0.4718	0.0998	0.0000	0.0000
Life Science Park	MLE	0.2844	0.4722	0.2434	0.0000	0.0000
	HB+MCMC	0.1688	0.6737	0.1575	0.0000	0.0000
Xierqi	MLE	0.3416	0.5530	0.1054	0.0000	0.0000
	HB+MCMC	0.2749	0.6397	0.0854	0.0000	0.0000

Table 2.
Comparison of estimation results of MLE and HB + MCMC

which are marked in yellow, orange and red, respectively, in Table 2. To explain the causes of different degrees of left behind phenomenon, we introduce two new kind of data: the average number of tap-in passengers during rush hours and the average train loading data during rush hours (provided by Beijing Metro Network Control Center). Figure 5 shows the relationship between the above three types of data in each station, through which we can summarize the three main reasons for the left behind phenomenon:

(1) *The limited train capacity*. For example, the loading rate of trains running in the section "Zhuxinzhuang - Life Science Park" has reached the upper limit (more than 100%), so when the train reaches Life Science Park, more than 80% of passengers waiting on the platform tend to have difficulty getting on the first train.

(2) *A large influx of passengers in a short period*. Take Shahe as an example, during the morning rush hours, an average of about 133 people entered the station every minute. As a result, the loading rate of trains soars from 69.96% to 107.81% after passing through the station, forcing passengers to wait for the next train.

(3) *Passengers' seat preference*. We observe that at Changping Xishankou and Ming Tombs, the loading rate of trains and number of passengers are small, but some passengers do not board the first train. This indicates that some passengers deliberately choose to wait for the next train under the condition of no crowding, and the most likely motivation for this behavior is to get a seat, which has also been found in cities such as Paris and Singapore (Kroes *et al.*, 2014) (Tirachini *et al.*, 2016).

4. Conclusions and discussions

This paper makes a quantitative analysis of the left behind phenomenon in the metro network based on a large number of passengers' travel records. The main contributions of this paper can be summarized as follows:

- After exploring the relationship between passengers' waiting time distribution and train timetable, a two-level HB model is proposed to describe the left behind patterns based on AFC data and train timetable data. Specifically, the waiting time of passengers is regarded as a GMM model, and each parameter of the GMM model are regarded as following Gaussian distribution, Uniform distribution and Dirichlet distribution, respectively.
- The HM algorithm is applied to estimate parameters in the proposed model. During the process of parameter estimation, we leverage the priori information of hyper-parameters

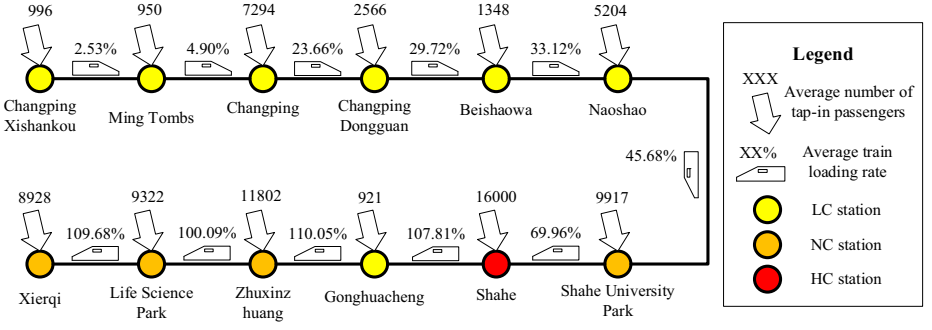

Figure 5.
Average number of tap-in passengers in each station and average train loading rate in each section in rush hours

based on train timetable data, which make full use of observed data to obtain more accurate results.

- A suburban metro line case of Beijing metro system is given as an application of proposed method and the estimating results illustrate that the method performs well in estimating left behind patterns.
- Three main reasons for the left behind phenomenon are summarized as follows: the limited train capacity, a large influx of passengers in a short period and passengers' seat preference. The above cause analysis can provide basis for metro operators to formulate relevant strategies, such as setting different headways in different sections in a metro line to meet the demand of passengers or adopting passenger flow control strategies in specific stations to avoid potential risks.

Future research will be carried out in the following two aspects:

(1) exploring more efficient iteration rules to improve the sampling efficiency to replace the traditional random walk process in MCMC methods; and

(2) reasoning the mechanism of mutual influence of left behind phenomenon between different stations.

References

Barry, J.J., Newhouser, R., Rahbee, A. and Sayeda, S. (2007), "Origin and destination estimation in New York city with automated fare system data", *Transp. Res. Rec. J. Transp. Res. Board*, doi: 10.3141/1817-24.

Chakirov, A. and Erath, A. (2011), "Use of public transport smart card fare payment data for travel behaviour analysis in Singapore", *16th Int. Conf. Hong Kong Soc. Transp. Stud.* 10.3929/ethz-a-006742061.

Delgado, F., Munoz, J.C. and Giesen, R. (2012), "How much can holding and/or limiting boarding improve transit performance?", *Transportation Research Part B: Methodological*, Vol. 46 No. 9, pp. 1202-1217, doi: 10.1016/j.trb.2012.04.005.

Fu, Q. (2014), *Modelling Route Choice Behaviour with Incomplete Data: An Application to the London Underground*, University of Leeds.

Hamdouch, Y., Ho, H.W., Sumalee, A. and Wang, G. (2011), "Schedule-based transit assignment model with vehicle capacity and seat availability", *Transportation Research Part B: Methodological*, Vol. 45 No. 10, pp. 1805-1830, doi: 10.1016/j.trb.2011.07.010.

Kroes, E., Kouwenhoven, M., Debrincat, L. and Pauget, N. (2014), "On the value of crowding in public transport for ile-de-France", *Transp. Res. Rec*, doi: 10.3141/2417-05.

Lee, M. and Sohn, K. (2015), "Inferring the route-use patterns of metro passengers based only on travel-time data within a Bayesian framework using a reversible-jump Markov chain Monte Carlo (MCMC) simulation", *Transportation Research Part B: Methodological*, Vol. 81, pp. 1-17, doi: 10.1016/j.trb.2015.08.008.

Li, X., Li, H. and Xu, X. (2018), "A Bayesian network modeling for departure time choice: a case study of Beijing subway", *PROMET - Traffic&Transportation*, Vol. 30 No. 5, pp. 579-587, doi: 10.7307/ptt.v30i5.2644.

Mueller, K. and Sgouridis, S.P. (2011), "Simulation-based analysis of personal rapid transit systems: service and energy performance assessment of the Masdar city PRT case", *J. Adv. Transp*, Vol. 47, pp. 512-525, doi: 10.1002/atr.

Noursalehi, P., Koutsopoulos, H.N. and Zhao, J. (2018), "Real time transit demand prediction capturing station interactions and impact of special events", *Transportation Research Part C: Emerging Technologies*, Vol. 97, pp. 277-300, doi: 10.1016/j.trc.2018.10.023.

Papola, N., Filippi, F., Gentile, G. and Meschini, L. (2009), "Schedule-based transit assignment: new dynamic equilibrium model with vehicle capacity constraints", in: *Schedule-Based Modeling of Transportation Networks*, Springer, Boston, MA, pp. 1-26.

Pereyra, M., Schniter, P., Chouzenoux, É., Pesquet, J.C., Tourneret, J.Y., Hero, A.O. and McLaughlin, S. (2016), "A survey of stochastic simulation and optimization methods in signal processing", *IEEE Journal of Selected Topics in Signal Processing*, Vol. 10 No. 2, pp. 224-241, doi: 10.1109/JSTSP.2015.2496908.

Rahbar, M., Hickman, M., Mesbah, M. and Tavassoli, A. (2019), "Calibrating a Bayesian transit assignment model using smart card data", *IEEE Transactions on Intelligent Transportation Systems*, Vol. 20 No. 4, pp. 1574-1583, doi: 10.1109/TITS.2018.2852726.

Robert, C.P. (2013), "Bayesian computational tools", *Annu. Rev. Stat. Its Appl*, doi: 10.1146/annurev-statistics-022513-115543.

Silva, R., Kang, S.M. and Airoldi, E.M. (2015), "Predicting traffic volumes and estimating the effects of shocks in massive transportation systems", *Proceedings of the National Academy of Sciences*, Vol. 112 No. 18, pp. 5643-5648, doi: 10.1073/pnas.1412908112.

Sun, Y. and Xu, R. (2012), "Rail transit travel time reliability and estimation of passenger route choice behavior", *Transp. Res. Rec. J. Transp. Res. Board*, doi: 10.3141/2275-07.

Sun, L., Lu, Y., Jin, J.G., Lee, D.H. and Axhausen, K.W. (2015), "An integrated Bayesian approach for passenger flow assignment in metro networks", *Transportation Research Part C: Emerging Technologies*, Vol. 52, pp. 116-131, doi: 10.1016/j.trc.2015.01.001.

Tirachini, A., Sun, L., Erath, A. and Chakirov, A. (2016), "Valuation of sitting and standing in metro trains using revealed preferences", *Transport Policy*, Vol. 47, pp. 94-104, doi: 10.1016/j.tranpol.2015.12.004.

Xu, X., Liu, J., Li, H. and Jiang, M. (2016), "Capacity-oriented passenger flow control under uncertain demand: algorithm development and real-world case study", *Transportation Research Part E: Logistics and Transportation Review*, Vol. 87, pp. 130-148, doi: 10.1016/j.tre.2016.01.004.

Xu, R., Li, Y., Zhu, W. and Li, S. (2018a), "Empirical analysis of traveling backwards and passenger flows reassignment on a metro network with automatic fare collection (AFC) data and train diagram", *Transp. Res. Rec*, doi: 10.1177/0361198118781395.

Xu, X., Xie, L., Li, H. and Qin, L. (2018c), "Learning the route choice behavior of subway passengers from AFC data", *Expert Systems with Applications*, Vol. 95, pp. 324-332, doi: 10.1016/j.eswa.2017.11.043.

Xu, X., Li, H., Liu, J., Ran, B. and Qin, L. (2018b), "Passenger flow control with multi-station coordination in subway networks: algorithm development and real-world case study", *Transp. B*, Vol. 0, pp. 1-27, doi: 10.1080/21680566.2018.1434020.

Zhao, J., Zhang, F., Tu, L., Xu, C., Shen, D., Tian, C., Li, X.Y. and Li, Z. (2017), "Estimation of passenger route choice pattern using smart card data for complex metro systems", *IEEE Transactions on Intelligent Transportation Systems*, Vol. 18 No. 4, pp. 790-801, doi: 10.1109/TITS.2016.2587864.

Zhu, Y., Koutsopoulos, H.N. and Wilson, N.H.M. (2017), "A probabilistic passenger-to-train assignment model based on automated data", *Transportation Research Part B: Methodological*, Vol. 104, pp. 522-542, doi: 10.1016/j.trb.2017.04.012.

Zhu, Y., Koutsopoulos, H.N. and Wilson, N.H.M. (2018), "Inferring left behind passengers in congested metro systems from automated data", *Transportation Research Part C: Emerging Technologies*, Vol. 94, pp. 323-337, doi: 10.1016/j.trc.2017.10.002.

Corresponding author
Chao Yu can be contacted at: 18114053@bjtu.edu.cn