Research Article

Bayesian Analysis for Metro Passenger Flows Using Automated Data

Chunya Li,1 Shifeng Xiong,2 Xuan Sun,3 and Yong Qin4

1School of Mathematics Physics and Statistics Shanghai University of Engineering Science, Shanghai 201620, China
2NCMIS, KLSC Academy of Mathematics and Systems Science Chinese Academy of Sciences, Beijing 100190, China
3School of Traffic and Transportation Beijing Jiaotong University, Beijing 100044, China
4State Key Lab of Rail Traffic Control and Safety Beijing Jiaotong University, Beijing 100044, China

Correspondence should be addressed to Shifeng Xiong; xiong@amss.ac.cn

Received 2 April 2021; Revised 7 January 2022; Accepted 11 February 2022; Published 15 March 2022

Copyright © 2022 Chunya Li et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

With the fast development of metro systems in many big cities, it is important to study the characteristics of passenger flows based on metro data for the management to guarantee service quality and safety. In this article, we build statistical models for the data of passengers’ tap-in and tap-out times in both no-transfer and one-transfer cases, and propose a Bayesian approach to estimate parameters in the models. These estimators can be used to evaluate a number of measures, which describe degrees of congestion and comfort, and to quantify their uncertainties. Application of our approach to Beijing metro shows different passengers follow different patterns between different routes and between off-peak and peak hours.

1. Introduction

As urban rail transit systems become more complex, there has grown up a pressing need for the safety and operation efficiency of urban rail transit in big cities. To alleviate the problem of traffic congestion and to meet passenger travel comfort requirements, operation management departments have formulated many programs, such as adding a current-limiting fence and adjusting the train operation status. Therefore, it is very important to analyse the metro passenger flows, especially in peak hours. Such a study can provide technical support for operators to implement reasonable management and efficient guidance, which are crucial to the quality assurance of the metro service and safety.

With continuous improvement of rail transit, automatic fare collection (AFC) systems can collect a wealth of passenger travel information and provide an important data support for metro operation departments to study passenger travel characteristics, formulate passenger flow control measures, and adjust driving plans. In recent years, mining behavior characteristics of passengers from AFC data has become an active branch in the field of passenger flow research (see, e.g., Paul [1]; Kusakabe et al. [2]; Sun and Xu [3]; Zhao and Yao [4]; Zhou et al. [5]; Sun and Schonfeld [6]; Chen et al. [7]; and Wu et al. [8]). For a fixed pair of origin-destination stations, studying behaviors of passengers based on only AFC data is actually a black box problem due to the lack of boarding information, which brings great challenges in accurately describing passenger flows. Probabilistic models are appropriate tools to study this problem. Fu et al. [9]; Lee and Sohn [10]; and Hörcher et al. [11] estimated the posterior probabilities of choosing different routes. Zhu et al. [12] proposed a probabilistic passenger-to-train assignment model to infer the probability of the passenger boarding each feasible train. Zhu et al. [13] inferred left-behind passengers in congested metro systems based on the probabilistic passenger-to-train assignment model. Besides AFC data, more information such as the distribution of access time is needed in these probabilistic models. Zhao et al. [14] developed a method to estimate passengers’ route choice pattern using smart card data. However, their method is based on the assumption that the time that passengers spend to walk between the platform and the origin-destinations...
entrance/exit is less than the departure interval between two adjacent trains. In practice, this assumption may not hold, especially during peak hours. Leurent and Xie [15] proposed a stochastic model to reposition the distances along platforms during train waiting. Their method requires both smartphone and AFC data, while smartphone data are not easy to obtain in many research studies. With only AFC data, few studies presented accurate probability distribution models. Recently, Xiong et al. [16] have proposed a new passenger-to-train assignment model. By deriving the probability distributions of AFC data, they provided maximum-likelihood estimates of unknown parameters in this model in no-transfer situations.

This article presents a Bayesian analysis of the statistical models of AFC data in both no-transfer and one-transfer situations. Our work extends Xiong et al. [16] in the following two directions. First, Xiong et al. [16] provided only point estimation for the parameters that lacks uncertainty quantification. Instead, we propose a Bayesian inferential method to quantify the uncertainty of parameter estimation. Interval estimation for the parameters and related passengers’ indices can be derived. Second, we show that Xiong et al. [16]’s estimation method cannot be straightforwardly extended to one-transfer situations due to an identifiability problem. With some prior information, our Bayesian method avoids this problem and yields reasonable estimators. Examples including real data applications and simulation results show the effectiveness of the proposed method.

The article is organized as follows. Section 2 presents the Bayesian inferential method for no-transfer situations. Section 3 describes our method for one-transfer situations. Section 4 gives applications of Bayesian estimators in deriving passengers’ characteristic indices. Section 5 conducts simulation studies. Section 6 provides real data examples in the Beijing metro system. Section 7 draws conclusions.

2. Bayesian Inference for No-Transfer Situations

In this section, we focus on a typical metro route, in which the passenger begins to tap in at the origin station and ends in tapping out at the destination station by taking only a train; that is, there is no any transfer station. We make the Bayesian inference for the metro passenger flow including estimating the probabilities of taking trains and the probability density functions of egress time, where egress time is the time it takes to walk to the tap-out fare gate after alighting from the train.

Since millions of people take metro trains everyday in big cities like Beijing, it is not difficult to collect sufficient data from passengers whose tap-in times are the same in large stations [16]. Suppose we get the AFC data of n passengers whose tap-in times are \(t_i\). Let \( T = \{ t_i, i = 1, \ldots, n \} \) denote the set of their tap-out times. In addition, we assume that the egress times of the \( n \) passengers are independently and identically distributed from the log-normal distribution \( LN(\mu, \sigma^2) \) with unknown parameters \( (\mu, \sigma^2) \), where \(-\infty < \mu < \infty \) and \( 0 < \sigma < \infty \). A random variable \( X \) is distributed from \( LN(\mu, \sigma^2) \) if \( \log(X) \) is distributed from the normal distribution \( N(\mu, \sigma^2) \) with mean \( \mu \) and variance \( \sigma^2 \). The log-normal distribution is commonly used to describe passengers’ walking times at metro stations [17]. Define

\[
\begin{align*}
DT_j & : \text{departure time of train } j \text{ from the origin station,} \\
AT_j & : \text{arrival time of train } j \text{ at the destination station,} \\
m_i & : \text{the number of feasible itineraries for passenger } i, \text{ i.e., } m_i = \# \{ j : DT_j \ldots t_0, AT_j \ldots t_i \}, \\
\pi_i & : \text{the probability that the whole passengers take the } jth \text{ train.}
\end{align*}
\]

Here, \( DT_j \) and \( AT_j \) can be obtained from the automatic vehicle location (AVL) system. Let \( \phi = (\pi', \theta')', \theta = (\mu, \sigma)' \) and \( \pi = (\pi_1, \ldots, \pi_m)' \). We need to estimate the unknown parameters \( \phi \). Based on the joint distribution of \( T \), the likelihood function can be given by

\[
P(T|\phi) = \prod_{i=1}^{n} \left( \sum_{j=1}^{m_i} \pi_j f(t_i - AT_j|\theta) \right).
\]

And the posterior distribution is

\[
P(\phi|T) \propto P(T|\phi)P_\pi(\pi)P_\theta(\theta),
\]

where the prior distributions of \( \pi \) and \( \theta \) denoted by \( P_\pi(\cdot) \) and \( P_\theta(\cdot) \), respectively, are independent.

Here, we choose the prior distribution of \( \theta \) as the normal distribution \( \mathcal{N}(0, \sigma^2) \) [18]. For the probability parameters \( \pi \), the Dirichlet distribution [19] is a natural prior. With hyperparameter vector \( \alpha = (\alpha_1, \ldots, \alpha_m)' \), let \( \pi = (\pi_1, \ldots, \pi_m) \sim \text{Dirichlet}(\alpha) \), that is,

\[
P_\pi(\pi) = \frac{1}{B(\alpha)} \prod_{j=1}^{m} \pi_j^{\alpha_j - 1},
\]

where \( B(\alpha) = \prod_{j=1}^{m} \Gamma(\alpha_j)/\Gamma(\sum_{j=1}^{m} \alpha_j) \) and \( \Gamma(t) = \int_{0}^{\infty} x^{t-1} e^{-x} dx \). In our problem, it seems not very reasonable to use \( \alpha_1 = \ldots = \alpha_m = 1 \), which gives uniform priors for \( \pi_1, \ldots, \pi_m \). We adopt to use a data-driven method to specify the hyperparameters. Let \( \overline{\pi} = (\overline{\pi}_1, \ldots, \overline{\pi}_m) \) be the maximum-likelihood estimators of \( \pi \) (see Xiong et al. [16]). We use \( \alpha = \overline{\pi} \) in (4). It should be pointed out that the selection of \( \alpha \) does not heavily influence the posterior when the sample size is sufficiently large.

We apply the Metropolis—Hastings [20, 21] algorithm to draw random samples from the posterior in (3), and the
corresponding Bayesian point and interval estimation for the parameters can be given.

3. Bayesian Inference for One-Transfer Situations

This section discusses one-transfer situations. We first show that the parameters in the likelihood function are not identifiable based on only automated data. Then, we give the Bayesian estimation method with some data-driven priors.

A one-transfer metro trip with two segments can be described by the following four steps: (1) tapping in and walking to the platform at the origin station, (2) waiting on the platform for a train until boarding a wanted train, (3) transiting between platforms and waiting until taking another train, and (4) getting off the train and tapping out at the exit gate (see Figure 1).

Similar to the previous section, suppose the tap-in times of the $n$ passengers are $t_i$. Let $T = \{t_i, i = 1, \ldots, n\}$ denote the set of their tap-out times, and assume that the egress times of the $n$ passengers are independently and identically distributed from $LN(\mu, \sigma^2)$. Define

$$\phi = (\rho', \theta'), \rho = (\rho_1, \ldots, \rho_{m_2}, \ldots, \rho_{m_k}, \ldots, \rho_{m_k}'), \text{ and } \theta = (\mu, \sigma').$$

In the maximum-likelihood estimation method that maximizes (4), we can only obtain the estimators of $\sum_{j=1}^{s_1} \rho_{jk}, k = 1, \ldots, K$, while the individual $\rho_{jk}$ are inestimable. For example, let $K = 2$ and $m_1 = m_2 = 2$. The parameter vector $\rho = (\rho_{11}, \rho_{12}, \rho_{21}, \rho_{22})'$ contains four components. By maximizing (4), we can estimate only two quantities, $\rho_{11} + \rho_{21}$ and $\rho_{12} + \rho_{22}$. The estimators of individual $\rho_{11}, \rho_{12}, \rho_{21}, \rho_{22}$ cannot be determined. Therefore, the maximum-likelihood estimation method in Xiong et al. [16] for no-transfer situations cannot be extended to the one-transfer situations.

To overcome the above identifiability problem, we use prior information of the parameters and present a Bayesian inferential approach. Similar to the previous section, let the prior distributions of $\theta$ and $\rho$ be independent. We choose the prior distribution of $\theta$ as the noninformative prior $P_{\theta}(\theta) \equiv 1/\sigma^2$ and assign the Dirichlet prior to $\rho$, that is,

$$DT_j^{(1)}: \text{departure time of train } j \text{ from the origin station of segment 1},$$

$$AT_j^{(1)}: \text{arrival time of train } j \text{ at the transfer station of segment 1}$$

$$DT_k^{(2)}: \text{departure time of train } k \text{ from the transfer station of segment 2},$$

$$AT_k^{(2)}: \text{arrival time of train } k \text{ at the destination station of segment 2},$$

$m_i$: the number of feasible itineraries for passenger $i$, i.e., $m_i = \#S_i$, where $S_i = \{(j, k): DT_j^{(1)} \geq t_i, AT_k^{(2)} < DT_k^{(2)}\}$.

$K_i$: the number of feasible itineraries on segment 2 for passenger $i$, i.e., $K_i = \#\{k: \text{there exists } j \text{ s.t. } (j, k) \in S_i\}$.

$m_i^{(k)}$: the number of feasible itineraries on segment 1 for passenger $i$ who takes the $k$th itinerary in segment 2, i.e., $m_i^{(k)} = \#\{(j, k) \in S_i\}, k = 1, \ldots, K_i$.
\[
P_\rho(\rho) = \frac{1}{B(\alpha)} \prod_{k=1}^{K} \prod_{j=1}^{m(k)} \rho_\delta^{\alpha_{jk}-1}, \tag{7}
\]
where \( \alpha = (\alpha_{11}, \ldots, \alpha_{m(1)1}, \ldots, \alpha_{1K}, \ldots, \alpha_{m(1)K})' \) are hyperparameters.

Consequently, the posterior distribution of \( \phi \) is
\[
P(\phi|T) \propto P(T|\phi)P_\rho(\rho)P_\theta(\theta), \tag{8}
\]
where \( P(T|\phi) \) is given in (6). It can be seen from (8) that the individual \( \rho_{jk} \) is not completely involved in \( \sum_{j=1}^{m(k)} \rho_{jk}, k = 1, \ldots, K \). The Metropolis–Hastings algorithm can be used to draw random samples from the posterior distribution of each component in the parameter vector \( \phi \).

We now provide a method to specify the hyperparameters \( \alpha \) in (7). First, we can compute the maximum-likelihood estimators \( \hat{\tau}_k \) of \( \tau_k = \sum_{j=1}^{m(k)} \rho_{jk} \) in (6) for \( k = 1, \ldots, K \). Next, consider the segment from the transfer station to the destination station, which is a no-transfer route. Let \( \delta_{jk} \) denote the probability that the passenger, whose tap-in time is \( AT_j^{(1)} \), takes his/her \( k \)th feasible itinerary in this no-transfer route for \( j = 1, \ldots, m(k) \), \( k = 1, \ldots, \bar{K} \), where \( \bar{K} \) can be different from \( K \). Based on the automated data on the segment, we can use the maximum-likelihood estimation method in Xiong et al. [16] to give estimators \( \hat{\delta}_{jk} \) of \( \delta_{jk} \). For \( \bar{K} > K \), we modify \( \hat{\delta}_{jk} \) as \( \sum_{k=K+1}^{\bar{K}} \delta_{jk} \). Consequently, we select
\[
\alpha_{jk} = \tau_k \delta_{jk}, j = 1, \ldots, m(k), k = 1, \ldots, K. \tag{9}
\]

### 4. Applications

This section presents some applications of the Bayesian estimators derived in the previous sections to the analysis of metro passenger flows. First of all, we should point out that the parameters \( \pi \) and \( f(·|\theta) \) can be viewed as measures that describe some aspects of characteristics of metro passenger flows. Therefore, our estimators can be used to infer these measures. Furthermore, let \( T1 \) denote the boarding time, that is, the time from tap-in to boarding a train; and let \( T2 \) denote the transfer time, that is, the time from alighting from the first train to boarding the second train. In the following, we infer \( T1 \) and \( T2 \) from both population and individual perspectives based on the estimators of \( \pi \) and \( \theta \).

#### 4.1. Population Inference

In no-transfer situations, for a passenger whose top-in time is \( t_0 \), his/her boarding time \( T1 \) is a random variable, which only takes the finite values \( \{DT_j^{(1)} - t_0 \} \). It is not difficult to derive the probability distribution, mean, and standard deviation of boarding time \( T1 \) as

\[
P(T1 = DT_j^{(1)} - t_0) = \pi_j, j = 1, \ldots, m,
\]

\[
E(T1) = m \sum_{j=1}^{m} \pi_j (DT_j - t_0),
\]

\[
\text{sd}(T1) = \left( \sum_{j=1}^{m} \pi_j (DT_j - t_0)^2 - \left( \sum_{j=1}^{m} \pi_j DT_j - t_0 \right)^2 \right)^{1/2},
\]

respectively. With the estimators of \( \pi \) in Section 2, we can estimate these quantities.

For one-transfer situations, the probability distribution, mean, and standard deviation of \( T1 \) of passengers who tap in at time \( t_0 \) are

\[
P(T1 = DT_j^{(1)} - t_0) = \sum_{k=1}^{K} \rho_{jk},
\]

\[
j = 1, \ldots, m(k),
\]

\[
E(T1) = \sum_{k=1}^{K} \sum_{j=1}^{m(k)} \rho_{jk} (DT_j^{(1)} - t_0),
\]

\[
\text{sd}(T1) = \left( \sum_{k=1}^{K} \sum_{j=1}^{m(k)} \rho_{jk} (DT_j^{(1)} - t_0)^2 - \left( \sum_{k=1}^{K} \sum_{j=1}^{m(k)} \rho_{jk} DT_j^{(1)} - t_0 \right)^2 \right)^{1/2},
\]

\[
\]
respectively. With the estimators of \( \rho \) in Section 3, we can also give estimators of the above quantities. Similar to \( T_1 \), the transfer time \( T_2 \) is also a random variable that takes the finite values \( \{DT_k^{(2)} - AT_j^{(1)}, k = 1, \ldots, K, j = 1, \ldots, m^{(k)}\} \). We can derive the probability distribution, mean, and standard deviation of \( T_2 \) as

\[
P(T_2 = DT_k^{(2)} - AT_j^{(1)}) = \rho_{jk},
\]

\[
k = 1, \ldots, K,
\]

\[
j = 1, \ldots, m^{(k)},
\]

\[
E(T_2) = \sum_{k=1}^{K} \sum_{j=1}^{m^{(k)}} \rho_{jk} (DT_k^{(2)} - AT_j^{(1)}),
\]

\[
\text{sd}(T_2) = \left( \sum_{k=1}^{K} \sum_{j=1}^{m^{(k)}} \rho_{jk} (DT_k^{(2)} - AT_j^{(1)})^2 - \left( \sum_{k=1}^{K} \sum_{j=1}^{m^{(k)}} \rho_{jk} (DT_k^{(2)} - AT_j^{(1)}) \right)^2 \right)^{1/2},
\]

\[
\text{respectively, and the estimators of } \rho \text{ can also be plugged in the above equations.}
\]

4.2. Individual Inference. For no-transfer situations, Zhu et al. [12] presented a probabilistic passenger-to-train assignment model (PTAM), which can be used to infer the most likely train that a certain passenger took as long as the parameters in it are estimated. Xiong et al. [16] proposed a modified PTAM (MPTAM) for both no-transfer and one-transfer situations. The parameters in MPTAM are actually those in the likelihood functions (1) and (4), which can be estimated by the proposed Bayesian method.

\[
P_i(T_1 = DT_j - t_0) = p_{ij},
\]

\[
i = 1, \ldots, n,
\]

\[
j = 1, \ldots, m_i,
\]

\[
E_i(T_1) = \sum_{j=1}^{m_i} p_{ij} (DT_j - t_0),
\]

\[
\text{sd}_i(T_1) = \left( \sum_{j=1}^{m_i} p_{ij} (DT_j - t_0)^2 - \left( \sum_{j=1}^{m_i} p_{ij} (DT_j - t_0) \right)^2 \right)^{1/2},
\]

\[
\text{respectively, and the estimators of } p_{ij} \text{ in (13) can be plugged in the above equations to give estimated values.}
\]

For one-transfer trips, the MPTAM formula is

\[
\rho_{ijk} = P(\text{the passenger took the } j \text{ th feasible itinerary on segment 1 and the } k \text{ th feasible itinerary on segment 2}|t_0, t_i)
\]

\[
= \frac{f(t_i - AT_j^{(2)}) \rho_{jk}(t_0)}{\sum_{k=1}^{K} \sum_{j=1}^{m^{(k)}} f(t_i - AT_k^{(2)}) \rho_{jk}(t_0)}
\]

\[
(15)
\]
The posterior distributions, means, and standard deviations of boarding time $T_1$ and transfer time $T_2$ of this passenger are

$$P_i(T_1 = DT_j^{(1)} - t_0) = \sum_{k=1}^{K} \sum_{j=1}^{m(k)} P_{ijk},$$

$$E_i(T_1) = \sum_{k=1}^{K} \sum_{j=1}^{m(k)} P_{ijk}(DT_j^{(1)} - t_0),$$

$$sd_i(T_1) = \left(\sum_{k=1}^{K} \sum_{j=1}^{m(k)} P_{ijk}(DT_j^{(1)} - t_0)^2 - \left(\sum_{k=1}^{K} \sum_{j=1}^{m(k)} P_{ijk}(DT_j^{(1)} - t_0)\right)^2\right)^{1/2},$$

$$P_i(T_2 = DT_k^{(2)} - AT_j^{(1)}) = P_{ijk},$$

$$E_i(T2) = \sum_{k=1}^{K} \sum_{j=1}^{m(k)} P_{ijk}(DT_k^{(2)} - AT_j^{(1)}),$$

$$sd_i(T2) = \left(\sum_{k=1}^{K} \sum_{j=1}^{m(k)} P_{ijk}(DT_k^{(2)} - AT_j^{(1)})^2 - \left(\sum_{k=1}^{K} \sum_{j=1}^{m(k)} P_{ijk}(DT_k^{(2)} - AT_j^{(1)})\right)^2\right)^{1/2},$$

respectively. Estimators of the above quantities can be computed based on the estimators of $\rho$ and $f$ in Section 3.

5. Simulation Study

5.1. No-Transfer Cases. In this section, we conduct a simulation study to evaluate the proposed Bayesian method for no-transfer situations. Consider the combinations of parameters as

$$(a). (\mu, \sigma) = (0, 0.5), (b). (\mu, \sigma) = (0.7, 0.3),$$

$$1^* (\pi_1, \pi_2) = (0.5, 0.5), 2^* (\pi_1, \pi_2) = (0.7, 0.3),$$

$$3^* (\pi_1, \pi_2, \pi_3) = (0.1, 0.3, 0.6), 4^* (\pi_1, \pi_2, \pi_3) = (0.8, 0.1, 0.1).$$

We assume that all passengers tap in at time 0, and let $AT^{(1)}_i = 2, AT^{(1)}_2 = 4, AT^{(1)}_3 = 6, \ldots, DT^{(1)}_1 = 3, DT^{(1)}_2 = 5, DT^{(1)}_3 = 7, \ldots$, (unit: minute). According to each parameter combination, we generate $n$ tap-out times $t_1, \ldots, t_n$ with $n = 100$ and 300, and then use the proposed Bayesian method in Section 2 to estimate the parameters. The $R$ package MCMC is adopted to implement the Bayesian inference. Root mean square errors (RMSEs) over 100 replicates are displayed in Table 1. It can be seen that our method can yield accurate estimation as $n$ increases.

5.2. One-Transfer Cases. We next conduct a simulation study for one-transfer cases. The following parameter combinations are considered:

$$(a). (\mu, \sigma) = (0, 0.5), (b). (\mu, \sigma) = (0.7, 0.3),$$

$$1^* (\rho_{11}, \rho_{21}) = (0.5, 0.5), 2^* (\rho_{11}, \rho_{21}) = (0.7, 0.3),$$

$$3^* (\rho_{11}, \rho_{21}, \rho_{12}, \rho_{22}) = (0.4, 0.1, 0.4, 0.1), 4^* (\rho_{11}, \rho_{21}, \rho_{12}, \rho_{22}) = (0.4, 0.3, 0.2, 0.1).$$

We assume that all passengers tap in at time 0, and let $AT^{(1)}_1 = 2, AT^{(1)}_2 = 4, DT^{(2)}_1 = 5, DT^{(2)}_2 = 7, DT^{(2)}_3 = 9, \ldots$ (unit: minute). Observations of tap-out times are generated according to each parameter combination with $n = 100$ and 200. In the proposed Bayesian method, we use the method in the last paragraph of Section 3 to determine the hyperparameters $\alpha$ in (7). We first compute the estimators $\hat{\tau}_k, k = 1, \ldots, K$. Since there are no data to estimate $\delta_{jk}, j = 1, \ldots, m(k), k = 1, \ldots, K$, we randomly generate $\delta_{jk}$ in $[\rho_{jk} - \epsilon, \rho_{jk} + \epsilon]$ with $\epsilon > 0$. The hyperparameters are set as
the proposed estimators perform well with small $\varepsilon$. For the more informative the prior.

describes the distance between $\delta_{jk}$ and $p_{jk}$. The lesser the $\varepsilon$, the more informative the prior.

We also use the R package MCMC to implement the Bayesian inference proposed in Section 3. For $\varepsilon = 0.05$ and $\varepsilon = 0.025$, root mean square errors (RMSEs) of our estimators over 100 replicates are shown in Table 2. As expected, the proposed estimators perform well with small $\varepsilon$.

6. Case Study

6.1. Two No-Transfer Routes. In this section, we apply our method to analyse several datasets from Beijing metro. The Beijing metro system is complex with 311 stations and 23 lines [22]. We first consider two no-transfer routes in Line 6 of Beijing metro as follows:

Route 1 from Qingnianlu to Dongdaqiao (4 stops)
Route 2 from Huangguo to Dongdaqiao (6 stops)

For Route 1, there are 25 and 649 trips collected during off-peak and peak hours, respectively. For Route 2, we have the corresponding data of sample sizes 15 and 525, respectively. All the data were extracted from weekdays in one month of 2019 at tap-in time 7:00 am (peak hour) and 1:00 pm (off-peak hour). We conduct the Bayesian inferential method in Section 2 via the R package MCMC. Table 3 presents the results of estimated parameters. The point estimators (posterior means) and interval limits of 95% credible intervals are reported. From Table 3, we can see that $\pi_3$ is larger than $\pi_1$ in Route 1 during off-peak hours, and this is most likely because the passengers’ tap-in times are very close to the departure time of the first train. Besides, during peak hours, the values of probabilities are not highly concentrated on one, which reflects the presence of crowded stations at that time. The origin station of Route 1, Qingnianlu, lies in Route 2, and the passengers in Route 1 are more difficult to get on the train than Route 2 during peak hours, which is reflected by the fact that $\pi_2$ and $\pi_4$ of Route 1 are larger than those of Route 2.

Figure 2 depicts the density curves of the egress times, and Table 4 presents some characteristics including means, standard deviations (SDs), and modes of the estimated densities. These results show that there is no big difference in egress time between the peak and off-peak hours for both Route 1 and Route 2. This is because Dongdamen, the destination station of the two routes, is a relatively unpopular station in Line 6. Thus, even during peak hours, this station is not so crowded. The passengers on the crowded trains are more likely to get off at the next stations in Line 6.

Besides parameter estimation, as mentioned in Section 4, we can infer boarding times of the passengers. Table 5 presents the means and limits of 95% credible interval of boarding times. We can see that the boarding times of peak and off-peak hours are close in Route 1, and this is because the headway between every two trains during off-peak hours is longer than that during peak hours. On the other hand, the boarding times in Route 1 are longer than those in Route 2, which is consistent to the fact that $\pi_1$ in Route 1 is much less than that in Route 2.

We next present an application of our method in individual inference with MPTAM. We take Route 1 during peak hours for example and let

$$ p_j(t) = P(\text{the passenger took itinerary } j | t_0 = 7:00 \text{am}, t_i = t). $$

(19)

for $j = 1, \ldots, 5$. Using (13), we compute the values of $p_j$ for given $t$ and show their curves in Figure 3. It can be seen that $p_1, p_2, p_3, p_4$, and $p_5$ are very close to 1 within a certain period of time. Besides, we find that the uncertainties of them are very small since the uncertainties of the estimators of $\theta$ have little effect on $f(\cdot|\theta)$. We also compute the mean boarding times of the passenger whose tap-out time is $t$, $E(t) = \sum_{j=1}^{5} p_j(t)(DT_j - t_0)$, and show its curve in Figure 4. As expected, the curve increases with a ladder-like shape as $t$ increases.

6.2. A One-Transfer Route. The AFC and AVL data used here are from a one-transfer route in Beijing metro in 15 weekdays of 2018. The first and second segments of this one-transfer route both contain six stops. We have $n = 359$ passengers’ AFC data, and their tap-in times are all 8:00 am. Due to a confidentiality agreement with the organization that provides the data, the original data including the station and train information cannot be open.

The proposed Bayesian method is used to analyse the passengers’ tap-out times, $t_1, \ldots, t_n$. The hyperparameters are specified by the method in Section 3 based on 52 passengers’ data of the second segment. Table 6 shows the parameter estimation results. We can see that the passengers’ boarding probabilities, $\rho_{11}, \ldots, \rho_{13}$, are dispersed. This indicates that there were many left-behind passengers during peak hours because of crowded platforms and train carriages. Table 7 gives the means and limits of 95% credible intervals of boarding times and transfer times. Compared with the train headway (about two minutes), the boarding and transfer times seem too long, and this may reflect uncomfortable travel experience of the passengers.

<table>
<thead>
<tr>
<th>$\alpha_{jk}$</th>
<th>$\tau_k$</th>
<th>$\delta_{jk}$</th>
<th>$\rho_{jk}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n = 100$</td>
<td>0.076 (0.081)</td>
<td>0.089 (0.082)</td>
<td>0.214 (0.212)</td>
</tr>
<tr>
<td>$n = 300$</td>
<td>0.054 (0.047)</td>
<td>0.048 (0.050)</td>
<td>0.148 (0.148)</td>
</tr>
</tbody>
</table>

Table 1: RMSEs in Section 5.1 (root standard deviations in parentheses).
Table 2: RMSEs in Section 5.2 (root standard deviations in parentheses).

<table>
<thead>
<tr>
<th>n</th>
<th>$\epsilon$</th>
<th>$\mu$</th>
<th>$\sigma$</th>
<th>$\pi_1$</th>
<th>$\pi_2$</th>
<th>$\pi_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>0.05</td>
<td>0.111 (0.113)</td>
<td>0.134 (0.152)</td>
<td>0.267 (0.192)</td>
<td>0.205 (0.164)</td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>0.043 (0.027)</td>
<td>0.060 (0.031)</td>
<td>0.198 (0.155)</td>
<td>0.135 (0.072)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>0.025</td>
<td>0.085 (0.053)</td>
<td>0.065 (0.057)</td>
<td>0.109 (0.058)</td>
<td>0.141 (0.152)</td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>0.032 (0.017)</td>
<td>0.037 (0.029)</td>
<td>0.083 (0.059)</td>
<td>0.118 (0.077)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>0.05</td>
<td>0.144 (0.137)</td>
<td>0.141 (0.131)</td>
<td>0.179 (0.113)</td>
<td>0.224 (0.176)</td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>0.064 (0.069)</td>
<td>0.033 (0.033)</td>
<td>0.060 (0.037)</td>
<td>0.176 (0.075)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>0.025</td>
<td>0.091 (0.081)</td>
<td>0.076 (0.020)</td>
<td>0.054 (0.047)</td>
<td>0.161 (0.089)</td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>0.034 (0.027)</td>
<td>0.017 (0.015)</td>
<td>0.026 (0.020)</td>
<td>0.126 (0.055)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3: Estimated parameters in Section 6.1

<table>
<thead>
<tr>
<th>Route 1</th>
<th>Off-peak hour</th>
<th>Peak hour</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu$</td>
<td>$\sigma$</td>
<td>$\pi_1$</td>
</tr>
<tr>
<td>Lower limit</td>
<td>-0.160</td>
<td>0.246</td>
</tr>
<tr>
<td>Point estimator</td>
<td>-0.010</td>
<td>0.304</td>
</tr>
<tr>
<td>Upper limit</td>
<td>-0.051</td>
<td>0.360</td>
</tr>
<tr>
<td>Lower limit</td>
<td>-0.147</td>
<td>0.451</td>
</tr>
<tr>
<td>Point estimator</td>
<td>-0.046</td>
<td>0.510</td>
</tr>
<tr>
<td>Upper limit</td>
<td>0.017</td>
<td>0.545</td>
</tr>
<tr>
<td>Lower limit</td>
<td>-0.100</td>
<td>0.311</td>
</tr>
<tr>
<td>Point estimator</td>
<td>-0.072</td>
<td>0.323</td>
</tr>
<tr>
<td>Upper limit</td>
<td>-0.050</td>
<td>0.335</td>
</tr>
<tr>
<td>Lower limit</td>
<td>-0.028</td>
<td>0.303</td>
</tr>
<tr>
<td>Point estimator</td>
<td>-0.007</td>
<td>0.321</td>
</tr>
<tr>
<td>Upper limit</td>
<td>0.016</td>
<td>0.344</td>
</tr>
</tbody>
</table>

Figure 2: Density functions of egress times in Section 6.1 (unit: minute). (a) Route 1. (b) Route 2.
Table 4: Characteristics of density functions in Section 6.1 (unit: minute).

<table>
<thead>
<tr>
<th></th>
<th>Peak hour</th>
<th>Off-peak hour</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Sd</td>
</tr>
<tr>
<td>Route 1</td>
<td>0.980</td>
<td>0.325</td>
</tr>
<tr>
<td>Route 2</td>
<td>1.045</td>
<td>0.344</td>
</tr>
</tbody>
</table>

Table 5: Characteristics of boarding times in Section 6.1 (unit: minute).

<table>
<thead>
<tr>
<th></th>
<th>Peak hour</th>
<th>Off-peak hour</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Sd</td>
</tr>
<tr>
<td>Route 1</td>
<td>5.460</td>
<td>2.450</td>
</tr>
<tr>
<td></td>
<td>5.574</td>
<td>2.450</td>
</tr>
<tr>
<td></td>
<td>5.682</td>
<td>2.636</td>
</tr>
<tr>
<td>Route 2</td>
<td>4.599</td>
<td>2.459</td>
</tr>
<tr>
<td></td>
<td>4.735</td>
<td>2.609</td>
</tr>
<tr>
<td></td>
<td>4.901</td>
<td>2.788</td>
</tr>
</tbody>
</table>

Figure 3: Curves of $p_1, \ldots, p_5$ in Section 6.1.

Figure 4: A curve of $E(T_1|t_i - t)$ in Section 6.1.

Table 6: Results of estimated parameters for Beijing metro of one-transfer cases.

<table>
<thead>
<tr>
<th></th>
<th>$\mu$</th>
<th>$\sigma$</th>
<th>$\rho_{11}$</th>
<th>$\rho_{21}$</th>
<th>$\rho_{31}$</th>
<th>$\rho_{12}$</th>
<th>$\rho_{22}$</th>
<th>$\rho_{32}$</th>
<th>$\rho_{13}$</th>
<th>$\rho_{23}$</th>
<th>$\rho_{33}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lower limit</td>
<td>0.010</td>
<td>0.513</td>
<td>0.000</td>
<td>0.207</td>
<td>0.287</td>
<td>0.065</td>
<td>0.213</td>
<td>0.196</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Point estimator</td>
<td>0.015</td>
<td>0.519</td>
<td>0.004</td>
<td>0.211</td>
<td>0.289</td>
<td>0.067</td>
<td>0.216</td>
<td>0.206</td>
<td>0.001</td>
<td>0.000</td>
<td>0.007</td>
</tr>
<tr>
<td>Upper limit</td>
<td>0.019</td>
<td>0.534</td>
<td>0.016</td>
<td>0.217</td>
<td>0.293</td>
<td>0.072</td>
<td>0.218</td>
<td>0.209</td>
<td>0.002</td>
<td>0.001</td>
<td>0.012</td>
</tr>
</tbody>
</table>
Similar to the no-transfer situation, we compute

\[ p_{jk}(t) = P(\text{the passenger took the } j\text{th feasible itinerary in segment 1 and the } k\text{th feasible itinerary in segment 2}| t_0 = 0, t_i = t) \],

for \( j = 1, 2, 3, k = 1, 2, 3 \) based on our estimators and MPTAM (15). Figure 5 shows their trajectories. For an individual passenger, we can use his/her tap-out time to infer which combination of itineraries he/she may take via the trajectories.

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Sd</th>
<th>Mean</th>
<th>Sd</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lower limit</td>
<td>4.395</td>
<td>1.704</td>
<td>3.541</td>
<td>1.909</td>
</tr>
<tr>
<td>Point estimator</td>
<td>4.446</td>
<td>1.709</td>
<td>3.561</td>
<td>1.916</td>
</tr>
<tr>
<td>Upper limit</td>
<td>4.471</td>
<td>1.715</td>
<td>3.581</td>
<td>1.937</td>
</tr>
</tbody>
</table>

Table 7: Characteristics of the boarding and transfer times in Section 6.2 (unit: minute).

Figure 5 shows their trajectories. For an individual passenger, we can use his/her tap-out time to infer which combination of itineraries he/she may take via the trajectories.

Figure 6 shows the curves of \( E(T1|t_i = t) \) and \( E(T2|t_i = t) \) computed by (9) and (10). They can be used to infer the mean boarding and transfer times of an individual passenger given his/her tap-out time.
7. Conclusion

In this article, we build statistical models for AFC data in both no-transfer and one-transfer cases, and then propose a Bayesian approach to estimate the parameters in the models. These estimators can be used in many measures for describing degrees of congestion and comfort in metro passenger flows. Applications to simulated and real datasets indicate that our approach is effective.

Compared with the maximum-likelihood estimation, the Bayesian method possesses some appealing features including flexibility in small-sample cases, improvement in model identifiability, and quantification of uncertainty. With appropriate prior information, it is a very useful tool for analysing metro data. Based on automated data, we will extend our approach to more complex situations in metro systems such as those with more than one route choices in the future. In addition, the log-normal assumption for the egresstime may be violated if the station has more than one exits [16]. A future topic is to apply Bayesian nonparametric density estimation methods to such situations.

Data Availability

The data used in Section 6, which contain the automatic fare collection data and automatic vehicle location data of several routes in Beijing metro, are from income accounting in several subway companies. The authors can use them in scientific research, but are not authorized to make them publicly available.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

Xiong’s work was partially supported by the National Key R&D Program of China (Grant nos. 2021YFA1000300, 2021YFA1000301, and 2021YFA1000303) and the National Natural Science Foundation of China (Grant nos. 12171462 and 11871033). Sun’s work was partially supported by the Foundation Sciences of Beijing Jiaotong University (Grant no. 2020YJS206). Qin’s work was partially supported by the High-level Talents Training Program of Ministry of Transportation of China (Grant no. 118100010).

References


