Hindawi
Journal of Advanced Transportation
Volume 2017, Article ID 6374858, 10 pages
https://doi.org/10.1155/2017/6374858

WILEY | Hindawi

Research Article

Estimating Bus Loads and OD Flows Using Location-Stamped Farebox and Wi-Fi Signal Data

Yuxiong Ji,[1] Jizhou Zhao,[1] Zhiming Zhang,[2] and Yuchuan Du[1]

[1]*Key Laboratory of Road and Traffic Engineering of the Ministry of Education, Tongji University, Shanghai 201804, China*
[2]*School of Automotive Studies, Clean Energy of Automotive Engineering Research Center, Tongji University, Shanghai 201804, China*

Correspondence should be addressed to Yuchuan Du; ycdu@tongji.edu.cn

Received 2 February 2017; Revised 7 April 2017; Accepted 23 April 2017; Published 23 May 2017

Academic Editor: Wai Yuen Szeto

Electronic fareboxes integrated with Automatic Vehicle Location (AVL) systems can provide location-stamped records to infer passenger boarding at individual stops. However, bus loads and Origin-Destination (OD) flows, which are useful for route planning, design, and real-time controls, cannot be derived directly from farebox data. Recently, Wi-Fi sensors have been used to collect passenger OD flow information. But the data are insufficient to capture the variation of passenger demand across bus trips. In this study, we propose a hierarchical Bayesian model to estimate trip-level OD flow matrices and a period-level OD flow matrix using sampled OD flow data collected by Wi-Fi sensors and boarding data provided by fareboxes. Bus loads on each bus trip are derived directly from the estimated trip-level OD flow matrices. The proposed method is evaluated empirically on an operational bus route and the results demonstrate that it provides good and detailed transit route-level passenger demand information by combining farebox and Wi-Fi signal data.

1. Introduction

Bus loads and OD flow matrices are commonly used to represent transit passenger demand. Bus loads, which depict the number of passengers on a bus, reflect bus crowding along a bus route. They are useful for schedule optimization and performance analyses [1, 2]. Origin-Destination (OD) flow matrices, which provide the number of passengers travelling between two specific stops, are important inputs in the design of multiple service patterns (e.g., short-turning, limited-stop) [3, 4] and real-time controls [5].

Transit agencies have increasingly deployed automated technologies to collect passenger demand information. For example, Automatic Passenger Counter (APC) systems collect boarding and alighting instances at each bus stop automatically. APC data are mainly used to estimate total ridership, average journey length, and bus loads. They could also be used to estimate OD flow matrices [6–10].

Automatic Fare Collection (AFC) systems have been deployed to reduce the costs for fare collection. The fare media used in AFC systems includes token, paper ticket, magnetic stripe card, and smartcard. The unique ID assigned to each smartcard makes it possible to track the transit movements of each smartcard holder. The smartcard data have been used to derive passenger OD flows [11–15]. However, many AFC systems for bus transit are access-based. The systems only have records of passengers entering the system and do not have records of passengers leaving the system. As a result, additional information or assumptions are needed to infer the destinations of passengers.

Traditional electronic farebox equipment are still widely used to collect transit fares. Electronic fareboxes could provide transactional records. Each record includes fare category, fare medium (e.g., cash, card, or transfer), current identifiers (i.e., operator, route, and run number), and a time stamp [5, 16]. Similar information could also be extracted from AFC systems. But passengers who pay by cash are not counted in AFC systems. When integrated with Automatic Vehicle Location (AVL) systems, the fareboxes could offer location-stamped data, which can be used to infer passenger boarding at individual bus stops [16]. Although passenger boarding instances are useful to understand ridership trends, it is insufficient for schedule, planning, and control applications.

Navick and Furth proposed methods to estimate passenger miles, OD patterns, and bus loads using farebox data, assuming that the pattern of passenger alighting instances in one direction is symmetric with the pattern of passenger boarding in the opposite direction [5]. Using similar assumptions, Lu and Reddy developed algorithms to measure daily bus passenger miles using farebox data [17]. Nevertheless, the symmetry assumption tends to be invalid if passengers use different bus routes or modes for their return trips.

Electronic fareboxes are more prevalent than APC techniques in bus systems. For example, Shanghai bus system served approximately 7.3 million passengers daily in 2014 using 16,717 buses on 1,354 bus routes [18]. Almost all buses are equipped with Global Position System (GPS) based AVL systems and fareboxes consisting of smartcards POS machines and coin fare collectors. However, only 265 of them are equipped with APC systems. On some bus routes, passenger boarding at individual stops has become available in real time, thanks to the integration of the farebox and AVL systems. However, efficient and economic solutions to obtain loads and OD flows are still unavailable.

Recently, some researchers attempted to collect passenger OD flows using Wi-Fi signal sensors [19, 20]. The sensors detect the unique Wi-Fi media access control (MAC) address of each device (e.g., smartphone). By detecting the MAC address at multiple locations over time, the origin and destination stops of the corresponding device can be inferred. Nevertheless, the detected devices only represent a sample of passengers. The resulting OD flow matrices may not be sufficient to capture the variation of passenger demand across bus trips.

In this study, we develop a hierarchical Bayesian model to estimate bus trip-level OD flow matrices and a period-level OD flow matrix based on farebox and Wi-Fi signal data. The period-level OD flow matrix represents passenger travel patterns on a bus route in a period. Bus loads and average journey length on each bus trip are derived directly from the estimated trip-level OD flow matrices. The performance of the proposed method is evaluated empirically on an operational bus route.

The remainder of this paper is organized as follows. The Wi-Fi signal sensor is introduced in Section 2 and the methodology of incorporating the farebox and Wi-Fi signal data is presented in Section 3. In Section 4, we show the empirical study followed by a discussion of the results and directions for future research.

2. Wi-Fi Signal Sensor

Wi-Fi signal sensors detect the signals emitted from Wi-Fi modules installed in various mobile devices. The Wi-Fi modules are defined based on IEEE 802.11 standard [21]. The basic unit for information exchange between devices is frame. The standard protocol defines multiple types of frames, such as Beacon, Acknowledgment (ACK), Data, and Probe. An access point periodically sends Beacon frames to announce its presence. When a mobile device is connected to an access point, information is exchanged via Data or ACK frames. If a mobile device is not connected to an access point, it would send Probe frames to search for available access points. The information detected by Wi-Fi signal sensors includes MAC addresses of the mobile device and the access point, frame type, time stamp, and signal strength. Signal strength is correlated with the distance between Wi-Fi sensor and mobile device.

It is worth mentioning that Apple Inc. introduced random MAC addresses during the release of the iOS 8 mobile operating system to protect user location privacy [22]. The iOS 8 system provides a random address when the device is searching for a Wi-Fi network. Nevertheless, the feature works only on iPhone 5S and iPhone 6 when the phone wakes up from a sleep mode and when the phone is not associated with a Wi-Fi network [23]. In reality, the phone is seldom in sleep mode. Many applications that use location services keep the phone awake despite the screen being switched off. In addition, a delicate combination of settings is involved to make the feature of random MAC addresses work. Therefore, unless the user turns off Wi-Fi or switches to airplane mode, he or she is still likely to be tracked.

3. Methodologies

3.1. A Hierarchical Bayesian Model. On a given bus route, it is assumed that farebox system, AVL system, and Wi-Fi signal sensor have been installed on buses. For each bus trip, farebox data provide stop-level boarding and Wi-Fi signal data provide sampled OD flows. Based on passenger boarding and sampled OD flows, we develop a hierarchical Bayesian model to estimate trip-level OD flow matrices and a period-level OD flow matrix. The period-level OD flow matrix is defined by an Alighting Probability (AP) matrix p, where $p(i, j)$ represent the probability of a passenger alighting at stop j conditional on having boarded at stop i. The AP matrix is assumed to be stable over bus trips in a homogeneous period (see Ji et al. [24] for the definition of homogeneous periods). The volume OD flow matrix for bus trip l is denoted by T_l, where $T_l(i, j)$ represents the number of passengers travelling from stop i to stop j. Note that T_l is unobservable in current problem, and we observed only two related quantities, the row totals of T_l in the form of passenger boarding obtained using farebox system and a sample of T_l observed using Wi-Fi signal sensors.

It is reasonable to assume that the estimates of T_l's are correlated and the relationship could be captured through taking T_l's as samples from a common population distribution with parameter p, the AP matrix. Therefore, we model the current problem hierarchically. That is, the observations are modelled conditional on parameters such as the trip-level OD flow matrices T_l's and these parameters are modelled conditional on other hyperparameters such as the AP matrix p.

Some notations are necessary to present the hierarchical model. Let T be the collection of T_l on all bus trips. Denote $b_l(i)$ as passenger boarding at stop i on bus trip l, b_l as the collection of $b_l(i)$ on bus trip l, and b as the collection of b_l. Let $x_l(i, j)$ be the number of sampled passengers boarding at stop i and alighting at stop j on bus trip l, x_l be the sampled OD flow matrix on bus trip l, and x be the collection of x_l.

The trip-level volume OD flow matrix T_l is determined by the AP matrix p and passenger boarding b_l. The associations among observed data on different bus trips are captured using a joint probability distribution for the volume OD flow matrices on different bus trips.

The hierarchical model structure could be better understood in the context of assumed distributions on variables and observed data. The distributions adopted in this study have been widely used in the literature to depict the randomness of the observed data or process. Specifically, (1) conditional on the alighting probabilities and passenger boarding at a given stop on a given bus trip, the OD flows originating from the given stop and destined to downstream stops on the given bus trip are assumed to follow multinomial distribution [7, 25]; (2) conditional on the volume OD flows on a given bus trip, the sampled OD flows are assumed to follow hypergeometric distribution [26]; (3) the prior distribution of the alighting probabilities is assumed to be Dirichlet, which is a conjugate prior distribution of the multinomial distribution [27].

Based on the above assumptions, the joint posterior likelihood of p and T is given by

$$\begin{aligned} f(p,T \mid x,b) &\propto f(x \mid T)\, f(T \mid p,b)\, f(p) \\ &\propto \prod_l f(x_l \mid T_l)\, f(T_l \mid p,b_l)\, f(p) \\ &\propto \prod_l \prod_i \prod_j \frac{p(i,j)^{T_l(i,j)}}{\Gamma(T_l(i,j)-x_l(i,j)+1)} \\ &\cdot \prod_i \prod_j p(i,j)^{\mu(i,j)-1}, \end{aligned} \tag{1}$$

where Γ represents the gamma function. $\mu(i,j)$ represents the hyperparameter of the prior distribution of the alighting probabilities and is positive for any feasible OD pair.

The hyperparameter $\mu(i,j)$ can be seen as the number of observations for each OD pair (i,j) that we have already observed. Prior information about the probability OD matrix, such as a model derived OD matrix and historical OD flow data, is incorporated through the hyperparameter μ. When no prior information is available, a uniform prior distribution is taken by setting $\mu(i,j) = 1$ for all feasible OD pairs. That is, we assign equal probability to vector with entries sum to one.

3.2. Estimates of OD Flow Matrices. Point estimates of p and T are valuable for various applications in practice. It is natural to choose the marginal posterior mode as the point estimate [28]. The derivations are presented in the following.

The marginal posterior likelihood of p is derived from (1) by summing it over all feasible OD matrices T_l for each bus trip:

$$f(p(i,) \mid x,b) \propto \prod_j p(i,j)^{\sum_l x_l(i,j)+\mu(i,j)-1}, \tag{2}$$

where $p(i,)$ represents the alighting probabilities at stops downstream of stop i for passengers having boarded at stop i. The marginal posterior likelihood of $p(i,)$ is determined by the prior and the sampled OD flows on all bus trips. Equation (2) is the density of Dirichlet distribution [27], and its mode is given by

$$p(i,j) = \frac{\sum_l x_l(i,j) + \mu(i,j)}{\sum_j \left(\sum_l x_l(i,j) + \mu(i,j)\right) - K_i}, \tag{3}$$

where K_i represents the number of stops downstream of stop i.

The marginal posterior likelihood of T derived from (1) by integrating over all feasible p is given by

$$\begin{aligned} f(T \mid x,b) \propto \prod_i \prod_j \left(\prod_l \frac{1}{\Gamma(T_l(i,j)-x_l(i,j)+1)} \right) \\ \cdot \frac{1}{\Gamma\left(\sum_l T_l(i,j) + \mu(i,j)\right)}. \end{aligned} \tag{4}$$

The mode of the likelihood of (4) can be obtained by solving the following maximization problem:

$$\begin{aligned} \max_{T \geq 0} \quad & -\sum_i \sum_j \left(\sum_l \left(T_l(i,j) - x_l(i,j)\right) \log\left(T_l(i,j) - x_l(i,j)\right) - T_l(i,j) \right) \\ & + \left(\sum_l T_l(i,j) + \mu(i,j) \right) \log\left(\sum_l T_l(i,j) + \mu(i,j) \right) - \sum_l T_l(i,j) \Bigg) \end{aligned} \tag{5}$$

$$\text{s.t.} \quad \sum_j T_l(i,j) = b_l(i) \quad \forall l. \tag{6}$$

The objective function in (5) equals approximately the logarithm of (4). Stirling's approximation is applied to the logarithm of the gamma function [29]. Analyzing the first-order necessary conditions of the model yields the optimal value of T:

$$T_l(i,j) = \left(\left(b_l(i) + \sum_{h \neq l} e_h(i) + \sum_k \mu(i,k) - K_i \right) x_l(i,j) \right.$$

$$+ \left(b_l(i) - e_l(i)\right)\left(\sum_{h \neq l} x_h(i,j) + \mu(i,j) - 1\right)\Bigg) \times \frac{1}{\sum_h e_h(i) + \sum_k \mu(i,k) - K_i}, \tag{7}$$

where $e_h(i)$ represents the sum of the sampled OD flows originating from stop i on bus trip h. Equation (7) reveals that the sampled OD flows on bus trips other than trip l also provide valuable information for the estimation of the OD flows on bus trip l.

The trip-level passenger alighting instances, bus loads, and average journey length can be derived from the trip-level OD flow matrices. Specifically, the alighting count, $a_l(j)$, at stop j on bus trip l is given by

$$a_l(j) = \sum_{i<j} T_l(i,j). \tag{8}$$

Bus load, $g_l(k)$, between stop k and stop $k+1$ on bus trip l is given by

$$g_l(k) = \sum_{i \leq k} \left(b_l(i) - a_l(i)\right) \tag{9}$$

And the average journey length, w_l, on bus trip l is given by

$$w_l = \frac{\sum_i \left(a_l(i)\, d(i) - b_l(i)\, d(i)\right)}{\sum_i b_l(i)}, \tag{10}$$

where $d(i)$ is the cumulative distance of stop i from the departure terminal. The numerator of (10) represents the total distance all passengers on bus trip l travelled. Note that it is not necessary to know the origin and destination stops of each passenger to obtain the total distance. The denominator represents the total number of passengers travelling on bus trip l.

3.3. Inferring OD Flows from Wi-Fi Data. If a mobile device emits signals in a high frequency, we could infer the origin and destination stops for the device with high confidence. However, the time intervals between consecutive signals are random and could be long. Thus, we propose a probabilistic method to quantify the uncertainties of the OD pairs that the detected passenger may travel along. For illustration, we consider a given device on a general bus trip and let m represent the number of signals emitted from the device and s_h represent the time interval between the hth and $(h+1)$th signals. Note that the subscript indicating bus trips is omitted for convenience in the following.

It is assumed that s_h follows a distribution with the parameter of λ. Conditional on Wi-Fi signals s and passenger boarding b, the posterior likelihood that the passenger carrying the given device boards at stop i and alights at stop j is given by

$$f(O = i, D = j \mid s, b) = c \times \int_\lambda f(s \mid O = i, D = j, \lambda)\, \pi(\lambda)\, f(O = i \mid b) \cdot f(D = j), \tag{11}$$

where O and D represent the origin and destination stops of the given passenger, respectively. c is the proportionality constant that satisfies the totality axiom. $\pi(\lambda)$ represents the prior distribution of λ. $f(O = i \mid b)$ represents the probability of the given passenger originating from stop i, conditional on passenger boarding b. $f(D = j)$ represents the probability of the given passenger destined to stop j. And,

$$f(s \mid O = i, D = j, \lambda) = f\left(s \geq s_0(i) \mid \lambda\right) \prod_{h=1}^{m-1} f(s_h \mid \lambda)\, f\left(s \geq s_m(j) \mid \lambda\right), \tag{12}$$

where $s_0(i)$ is the time interval between the first Wi-Fi signal and bus arrival time at stop i. $s_m(j)$ is the time interval between bus arrival time at stop j and the last Wi-Fi signal.

Substituting (12) into (11), (11) can be expressed by

$$f(O = i, D = j \mid s, b) = c \times \int_\lambda f\left(s \geq s_0(i) \mid \lambda\right) \cdot \prod_{h=1}^{m-1} f(s_h \mid \lambda)\, f\left(s \geq s_m(j) \mid \lambda\right) \pi(\lambda) \times f(O = i \mid b)\, f(D = j). \tag{13}$$

The proportionality constant c is a value such that the summation of the posterior likelihoods over all feasible stop pairs equals one. Let $f_k(i,j)$ represent the posterior likelihood of travelling between stop i and stop j for passenger k. The sampled passenger flow between stop i and stop j is estimated by aggregating $f_k(i,j)$ over k:

$$x(i,j) = \sum_k f_k(i,j). \tag{14}$$

For simplicity, (14) is used in (3) and (7) to estimate the period-level AP matrix and trip-level OD flow matrices. How to incorporate the uncertainty in Wi-Fi OD flows in the estimation of OD flow matrices is reserved for future research.

4. Empirical Evaluation

4.1. Data. The data used in this study were collected on Route Jiahuang in Jiading district of Shanghai in the period between 8 and 9 am on weekdays in June of 2016 (see Figure 1 for the route map). The bus route is 19 km long. Buses operate with the headway of 18 minutes. Along the bus route, Huangdu, Laozhai, and Fangtai are three densely populated towns and Downtown of Jiading is the center of Jiading District. This study focuses on passengers travelling from Huangdu town to North Jiading. The route in this direction has 20 bus stops and 190 feasible OD pairs.

Passenger demand on 12 bus trips was collected. A surveyor carrying a Wi-Fi sensor and a GPS logger rode a bus to count boarding and alighting instances at each stop. At the same time, the Wi-Fi signal sensor detected and recorded Wi-Fi signals and the GPS logger recorded bus locations automatically. The total number of passengers on 12 bus trips

Table 1: Passenger and device observations by bus trip.

Trips	Number of passengers	Devices	Number of devices after filtering			Rate
			Step 1	Step 2	Step 3	
1	26	842	35	15	7	26.9%
2	52	1036	52	36	20	38.5%
3	70	1025	87	54	31	44.3%
4	64	1009	50	32	14	21.9%
5	73	1803	73	42	27	37.0%
6	80	984	85	63	31	38.8%
7	52	1144	121	65	37	71.2%
8	31	977	22	17	8	25.8%
9	84	1818	74	61	39	46.4%
10	85	841	95	64	37	43.5%
11	69	1853	61	40	14	20.3%
12	58	1202	36	23	19	32.8%
Total	744	14534	791	512	284	38.2%

Figure 1: Illustration of Wi-Fi signals on the route map.

is 744, while 14,534 mobile devices were detected by Wi-Fi sensors. Obviously, most of the detected devices do not belong to onboard passengers.

4.2. Data Cleaning. The detection range of the Wi-Fi sensors used in the experiment is approximately 100 meters. The sensors would inevitably pick up Wi-Fi signals from nonpassengers, especially when buses dwell at bus stops or stop near intersections. The sensor could also detect stationary Wi-Fi routers. When traffic is congested, the sensors may receive signals from devices on general vehicles for a relatively long time.

A procedure is developed to filter out erroneous data. For signals with the same MAC address, the following steps are carried out to check the validity of the data:

(1) Low signal strength provides clues that the signals may come from mobile devices outside of the buses. We exclude the MAC address if the median signal strength is less than −60 dBm. We use the threshold of −60 dBm because it approximately indicates that the distance between the detected device and the Wi-Fi sensor is 10 meters, which is considered reasonable given the size of a bus.

(2) Signals from nonpassengers usually last for relative short distance. We exclude the MAC address if the distance between the location where the MAC address is initially detected and the location where the MAC address is lastly detected is less than 400 meters. We adopt the threshold of 400 meters because passengers tend to choose walking for such a short journey. In addition, the adopted threshold is reasonable since the shortest distances between two consecutive stops is 370 meters and most of them are longer than 900 meters.

(3) Onboard passengers would be detected multiple times. Small number of signals indicates that the signals come from nonpassengers. We exclude the MAC address with less than 10 signals.

Table 1 summarizes the observed data by bus trip. Eventually, 94.6%, 2.0%, and 1.6% of the detected devices are excluded, respectively, by Steps (1)–(3) above. The valid sample rate, which is defined as the ratio of the final sample size to the number of passengers, varies from 20.3% to 71.2% across bus trips. The whole valid sample rate is 38.2%. The data after filtering are used to derive sampled OD flows on each bus trip.

The effects of the screening criteria on the final sample sizes are analyzed using one-at-a-time (OAT) technique. That is, we analyze the effect of one screening criterion at a time, keeping the other criteria fixed. The effects of the screening criteria on the final sample size are presented in Table 2. As can be seen, the final sample size is sensitive to the screening criteria in Step 1 and Step 3 but is insensitive to the criterion in Step 2. In Step 1, when the threshold of the median signal strength is increased from −70 to −55, the sample size is reduced by 116. The insensitivity of the sample size to the distance is due to the filter in Step 3. The mobile device detected for a short distance usually results in small number of signals. In Step 3, when the threshold of the number of signals is increased from 5 to 20, additional 145 devices would be filtered.

Table 2: Sensitivity of the final sample size to the screening criteria.

Step 1: signal strength			
Criteria (dBm)	−70	−60	−55
Sample size	333	284	217
Step 2: distance			
Criteria (m)	100	400	1000
Sample size	284	284	278
Step 3: number of signals			
Criteria	5	10	20
Sample size	349	284	204

4.3. Inferring Wi-Fi OD Flows. The observed data showed that the time intervals between consecutive signals are random and their lengths depend on the type and brand of the devices. Figure 1 illustrates the Wi-Fi signals for two devices on the route map. The signal frequency of device A is high. It is clear that the origin is stop 8 and the destination is stop 15. Nevertheless, the signal frequency of device B is low and the variation of the time intervals are relatively high, which bring high uncertainty about the origin and destination stops of device B. In our dataset, the mean time intervals of 62.0% of the valid MAC addresses are less than 30 seconds. But the time intervals of some MAC addresses are quite long. The mean time interval is above 2 minutes for 5.9% of the valid MAC addresses.

The fitness of the time intervals between consecutive Wi-Fi signals to various distributions (i.e., exponential, gamma, and lognormal) was evaluated using Kolmogorov–Smirnov test [30]. The evaluation revealed that the exponential, gamma, and lognormal distributions provide good fitness (p values are greater than 0.05) for 46%, 51%, and 48% of mobile devices, respectively. Further investigations suggest that the estimates of bus loads and OD flows are insensitive to the distributions adopted.

For simplicity, we assume that the time intervals between consecutive signals follow exponential distribution with the parameter λ. The exponential distribution is commonly used to describe the time between events [31]. Noninformative prior distribution is used for λ. That is, $\pi(\lambda) \propto 1/\lambda$. Conditional on passenger boarding $b(i)$, it is assumed that the probability that the given passenger originates from stop i is proportional to $b(i)$. A uniform prior is used for the destination of the given passenger, which represents that the passenger is equally likely to alight at any downstream stop. Based on the above assumptions, (13) can be expressed by

$$\begin{aligned} f\left(O=i, D=j \mid s, b\right) &= c \times \int_{\lambda} f\left(s \geq s_0(i) \mid \lambda\right) \\ &\cdot \prod_{h=1}^{m-1} f\left(s_h \mid \lambda\right) f\left(s \geq s_m(j) \mid \lambda\right) \frac{1}{\lambda} \times b(i) = c \\ &\times \frac{b(i)}{\left(s_0(i) + s_m(j) + (m-1)\,\bar{s}\right)^{m-1}}, \end{aligned} \tag{15}$$

where $\bar{s}$ represents the mean of the time intervals between consecutive signals. The posterior likelihood of travelling between each feasible stop pair is obtained using (15). Wi-Fi OD flows for each bus trip is obtained using (14).

4.4. Performance Metrics. The objective of this study is to obtain good estimates of bus loads and OD flows using farebox and Wi-Fi signal data. The true trip-level bus loads and average journey length are available since we have manually collected the boarding and alighting instances at each stop for each bus trip (see (9) and (10)). Nevertheless, the true OD flow data were not collected due to the fact that OD surveys are costly and labor intensive. Thus, we evaluate the performance of the proposed method by comparing the estimates of bus loads and average journey length with the observed values. Average journey length has been considered to quantify the accuracy of OD flow estimates [10, 26]. Poor estimates of bus loads and average journey lengths indicate low accuracy of OD flow estimates.

Specifically, the estimation error of bus load is quantified by the mean of the maximum absolute difference between the estimated and observed bus loads, R, where the mean is taken over bus trips:

$$R = \frac{\sum_l \max_k \left|\widehat{g}_l(k) - g_l(k)\right|}{L}, \tag{16}$$

where L represents the number of bus trips. $\widehat{g}_l(k)$ represents the estimated bus load between stop k and stop $k+1$ on bus trip l. Higher value of R indicates poorer estimates of bus loads.

The estimation error of the average journey length is quantified by the weighted average of the absolute difference between the estimated and true average journey lengths, W, where the average is taken over bus trips and the weight is the number of passengers on each bus trip:

$$W = \frac{\sum_l \left|\widehat{w}_l - w_l\right| \times \sum_i b_l(i)}{\sum_l \sum_i b_l(i)}, \tag{17}$$

where $\widehat{w}_l$ represents the estimate of the average journey length on bus trip l. Higher values of R and W indicate lower accuracy of OD flow estimates.

For comparison, a simple method that is straightforward and could be easily implemented in practice is also considered in the evaluation. The simple method assumes that observed data are independent across bus trips. For bus trip l, passenger flow from stop i to stop j is estimated by scaling

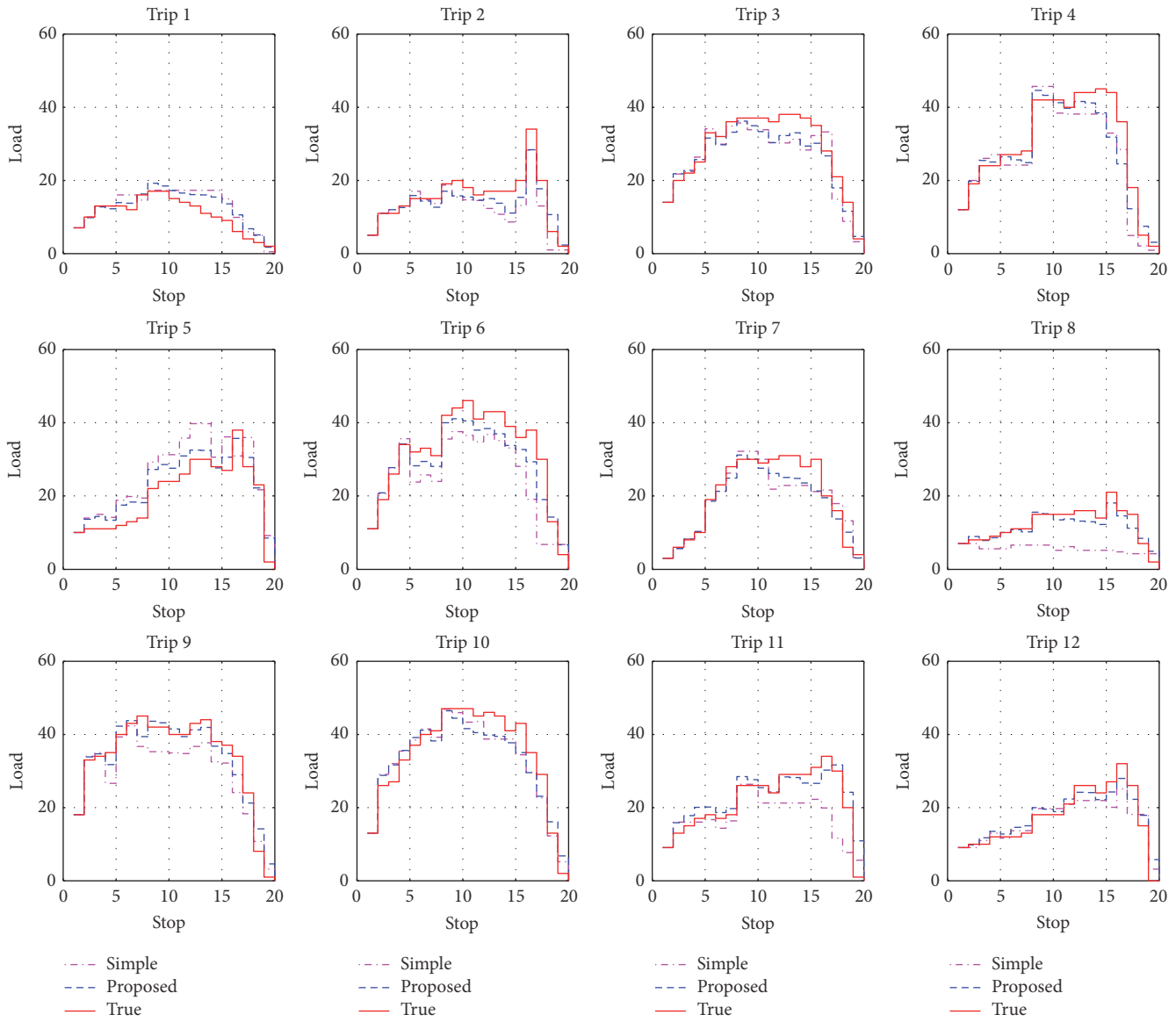

Figure 2: Estimated and true bus load profiles.

up the Wi-Fi OD flow $x_l(i, j)$ by the ratio of the true boarding count to the boarding count sampled by Wi-Fi sensor at stop i:

$$T_l(i, j) = b_l(i) \times \frac{x_l(i, j)}{\sum_k x_l(i, k)}. \quad (18)$$

The bus loads and average journey lengths estimated by the simple method could be obtained using (8)–(10). In addition, the simple method estimates the AP matrix by the pooled Wi-Fi OD flows. The resulting estimate is equivalent to the one produced by the proposed method if the uniform prior is used (see (3)).

4.5. Results. In this part, we compare the performance of the simple and proposed methods in terms of bus loads and average journey lengths. Uniform prior distribution for the alighting probabilities is used in the proposed method. As a result, the proposed method and the simple method produce the same AP estimate.

Figure 2 presents the estimated and true bus load profiles for each bus trip. As can be seen, the bus load profile varies greatly across bus trips. Since the data were collected in the same period over multiple days, day-to-day variation partially leads to the relatively large trip-level demand variation. In addition, the bus route is located in suburban area. The variety of passengers also contributes to the variation of the trip-level passenger demand.

As shown in Figure 2, the maximum load may be observed on any segments between stop 7 and stop 17. The load profiles derived from the proposed method show relatively higher degree of similarity with the corresponding true ones. The variations of bus loads across segments and bus trips are well captured by the load profiles resulting from the proposed method. In contrast, the simple method may

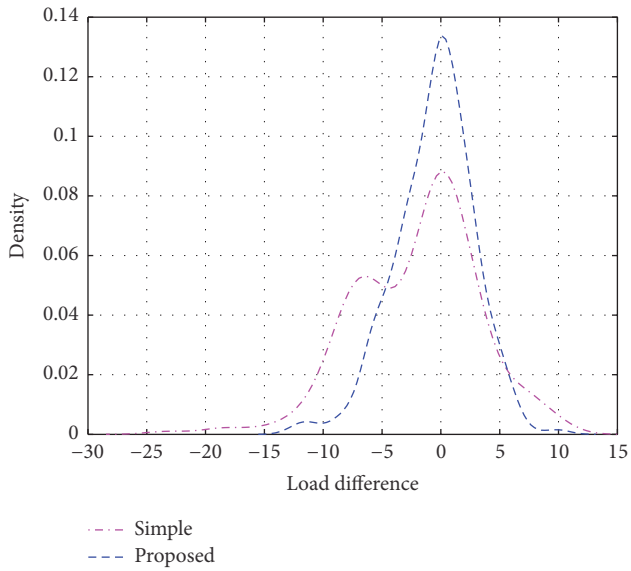

Figure 3: Distributions of the difference of the estimated and true loads.

produce poor load estimates on some bus trips, such as bus trip 8.

To quantify the estimation errors, Figure 3 presents the distributions of the differences between the estimated and true bus loads. Figure 3 is obtained by pooling the differences between the estimated and true bus loads over all segments and all bus trips. Positive difference indicates the load is overestimated. Figure 3 reveals that the proposed method provides better bus load estimates than the simple method. The absolute load difference is less than 5 passengers on 86.3% of route segments for the proposed method, while it is 62.1% for the simple method. In addition, the maximum absolute load differences are 12 and 23 passengers, respectively, for the proposed and simple methods. Overall, the load estimation error, R, equals 7.6 for the proposed method and 11.7 for the simple method. Figure 3 also shows that the simple method tends to underestimate bus loads. The estimated bus loads are lower than the true ones by more than 5 passengers on 30.4% of route segments.

The underestimation of the bus loads produced by the simple method stems from the well-known "structural zeros" problem [26]. That is, when the sampled flows in some OD pairs are zero, the updated flows in the corresponding pairs are also zero. Therefore, when no passengers boarding at a given stop are detected by Wi-Fi sensor, the estimated OD flows originated from the given stop would be zero if the simple method is used. The "structural zeros" problem is more serious on bus trips with low sample size, such as bus trips 1, 8, and 11. In contrast, the "structural zeros" problem is greatly alleviated in the hierarchical model since it takes into account the sampled OD flows on other bus trips when estimating OD flows for a bus trip (see (7)).

Figure 4(a) presents the estimated and true average journey lengths on each bus trip. The journey length is measured by number of bus stops. Figure 4(a) demonstrates that the proposed method produces better estimates of the average journey length on most bus trips. The average journey lengths estimated by the simple method are remarkably different from the true ones on some bus trips. For example, when the simple method is used, the differences on bus trips 1, 8, and 11 between the estimated and true values are 3.9, 4.4, and 3.2, respectively. By contrast, when the proposed method is applied, the difference goes down to 1.4, −0.5, and 0.3 on the respective bus trip. The poor performance of the simple method on bus trips 1, 8, and 11 is very likely due to the low sample size and the "structural zeros" problem. Overall, the estimation error of the average journey length, W, equals 0.4 for the proposed method and equals 1.2 for the simple method. When passengers on all bus trips are considered, the true average journey length is 7.0 stops. The average journey lengths resulting from the simple and proposed methods are 7.6 and 6.9 stops, respectively.

The distribution of passengers' journey length, which is obtained by pooling the estimated trip-level estimated OD flow matrices, is presented in Figure 4(b). The distributions produced by the simple method and the proposed method show similar patterns. Nevertheless, the simple method tends to produce lower proportion of short trips (e.g., less than 6 stops). Consider the distribution produced by the proposed method, the journey lengths of 53.5% of passengers are less than 5 stops, and 12.4% of passengers board at a stop and alight immediately at next stop. This result is understandable since the bus route is located in the suburban area and the distances between consecutive stops are relatively long. It is also found that the journey lengths of 10.8% of passengers are over 15 stops.

The period-level AP matrix resulting from the proposed method reveals some interesting travel patterns. Figure 5 presents the largest three probabilities that passengers alight at downstream zones, conditional on having boarded at stop 1 or stop 2. As can be seen, passengers having boarded at stop 1 and stop 2 have different travel patterns. Passengers having boarded at stop 1 are more likely to take long journeys. Specifically, passengers having boarded at stop 1 have high probabilities of alighting at Outskirts of Jiading zone and Downtown of Jiading zone, while passengers having boarded at stop 2 have high probabilities of alighting at Laozhai and Zhaoxiang zones.

The results in Figure 5 are explainable. Stop 1 and stop 2 are over 800 meters apart. Stop 1, which is a terminal stop, lies to the east of Huangdu town, and stop 2 is located at the center of Huangdu town where population is concentrated. Passengers originating from Huangdu town could wait at stop 2 or walk to stop 1. Since passengers who travel further have greater stimulus to chase seats [32], passengers with long journeys are more likely to walk to stop 1 to have a good chance to get a seat. Nevertheless, passengers who travel to nearby stops tend to wait for buses at stop 2.

5. Conclusion and Future Research

In this study, we propose a hierarchical Bayesian model to estimate transit route-level passenger travel patterns using farebox and Wi-Fi signal data. The proposed model captures

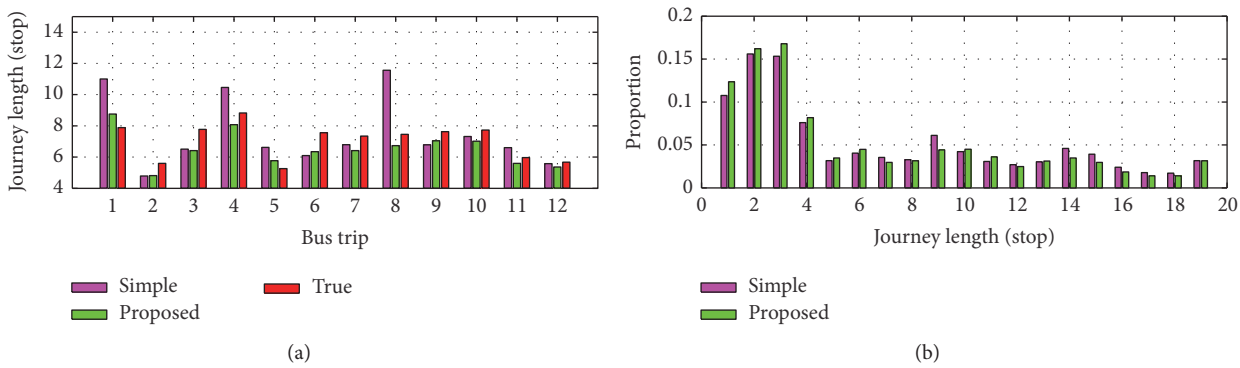

Figure 4: Passengers' journey length: (a) average journey length by bus trip; (b) distribution of journey length.

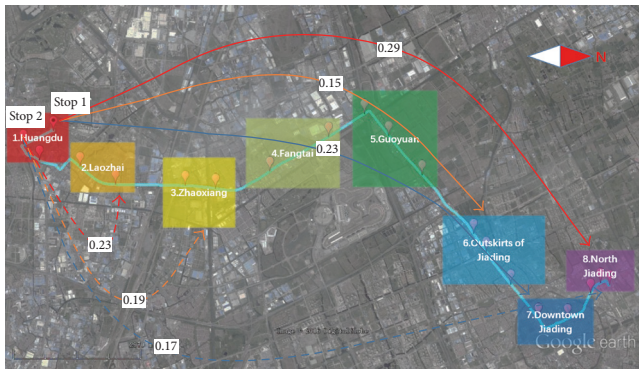

Figure 5: Top three alighting probabilities for boarding stops 1 and 2.

the associations among observed data on different bus trips. When estimating OD flows for a bus trip, the sampled OD flows on other bus trips are also taken into account. Doing so not only improves the accuracy of the OD flow estimates but also reduces the likelihood of encountering the "structural zeros" problem. Procedures are developed to filter noisy Wi-Fi signal data and determine Wi-Fi OD flows for each bus trip. Empirical evaluation on an operational bus route demonstrates that the proposed method outperforms a simple method that assumes observed data are independent across bus trips. The empirical results demonstrate the feasibility of combining farebox and Wi-Fi signal data to obtain good and detailed transit route-level passenger demand information.

Although promising, additional research would be necessary before the method can be applied in operational use. One limitation of our study is that direct evaluation of the OD flow estimates is lacking. Instead, we used measures based on bus load and average journey length to reflect indirectly the accuracy of the OD flow estimates. It would be useful to conduct additional study to evaluate the performance of the proposed method by comparing the estimated OD flows with the field-observed OD flows.

In addition, this study uses error-free boarding information. Nevertheless, passenger boarding provided by fareboxes contains errors, which may result from mechanical failures, imperfect interactions between passengers and fareboxes and between drivers and fareboxes, and problems of matching farebox and AVL data. It is valuable to conduct a large scale empirical study with actual farebox data to quantify the effect of boarding errors on the accuracy of bus load and OD flow estimates.

Moreover, after filtering the Wi-Fi signal data, the remaining data may still contain data from nonpassengers. Additional potential problems of the Wi-Fi signal data include that one passenger may carry more than one mobile device and that the youth may be oversampled since they are more likely to use smart phones. Furthermore, as discussed before, the Wi-Fi OD flows are of uncertainty. How to consider multiple sources of errors and the uncertainty in Wi-Fi OD flows in the estimation of OD flow matrices is worth investigating in the future.

Despite the need for additional investigations, the method proposed in this paper is promising. The method is data driven and does not require expert intervention. It is also computationally efficient and can be extended for real-time applications. In real-time scenario, historical data would be considered as a priori information for the estimation of OD flows and bus loads on a bus trip. The real-time estimation of bus trip-level OD flows and bus loads is useful for transit controls and schedule coordination.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (51308410, 71361130013), Shanghai Science and Technology Committee (15DZ1204402), and the 13th Five-Year National Key Research and Development Plan (2016YFB1200402). The work of the first author was supported by Program for Changjiang Scholars and Innovative Research Team in University.

References

[1] A. Ceder, "Bus frequency determination using passenger count data," *Transportation Research Part A-Policy and Practice*, vol. 18, no. 5-6, pp. 439–453, 1984.

[2] P. G. Furth et al., "Using archived Avl-Apc data to improve transit performance and management," in *Transit Cooperative Research Program 113*, 2006.

[3] C. E. Cortés, S. Jara-Díaz, and A. Tirachini, "Integrating short turning and deadheading in the optimization of transit services," *Transportation Research Part A: Policy and Practice*, vol. 45, no. 5, pp. 419–434, 2011.

[4] A. Tirachini, C. E. Cortés, and S. R. Jara-Díaz, "Optimal design and benefits of a short turning strategy for a bus corridor," *Transportation*, vol. 38, no. 1, pp. 169–189, 2011.

[5] D. Navick and P. Furth, "Estimating passenger miles, origin-destination patterns, and loads with location-stamped farebox data," *Transportation Research Record: Journal of the Transportation Research Board*, no. 1799, pp. 107–113, 2002.

[6] P. G. Furth and D. S. Navick, "Bus route O-D matrix generation: relationship between biproportional and recursive methods," *Transportation Research Record: Journal of the Transportation Research Board*, no. 1338, pp. 14–21, 1992.

[7] Y. Li and M. J. Cassidy, "A generalized and efficient algorithm for estimating transit route ODs from passenger counts," *Transportation Research Part B: Methodological*, vol. 41, no. 1, pp. 114–125, 2007.

[8] Y. Ji, R. G. Mishalani, and M. R. McCord, "Estimating transit route OD flow matrices from APC data on multiple bus trips using the IPF method with an iteratively improved base: method and empirical evaluation," *Journal of Transportation Engineering*, vol. 140, no. 5, Article ID 04014008, pp. 1–8, 2014.

[9] Y. Ji, R. G. Mishalani, and M. R. McCord, "Transit passenger origin-destination flow estimation: efficiently combining onboard survey and large automatic passenger count datasets," *Transportation Research Part C: Emerging Technologies*, vol. 58, pp. 178–192, 2015.

[10] Y. Ji, Q. You, S. Jiang, and H. M. Zhang, "Statistical inference on transit route-level origin-destination flows using automatic passenger counter data," *Journal of Advanced Transportation*, vol. 49, no. 6, pp. 724–737, 2015.

[11] J. Zhao, A. Rahbee, and N. H. M. Wilson, "Estimating a rail passenger trip origin-destination matrix using automatic data collection systems," *Computer-Aided Civil and Infrastructure Engineering*, vol. 22, no. 5, pp. 376–387, 2007.

[12] J. Barry, R. Freimer, and H. Slavin, "Use of entry-only automatic fare collection data to estimate linked transit trips in New York city," *Transportation Research Record*, no. 2112, pp. 53–61, 2009.

[13] J. J. Barry et al., "Origin and destination estimation in New York city with automated fare system data," *Transportation Research Record*, no. 1817, pp. 183–187, 2002.

[14] D. Uniman, J. Attanucci, R. Mishalani, and N. Wilson, "Service reliability measurement using automated fare card data: application to the London underground," *Transportation Research Record*, no. 2143, pp. 92–99, 2010.

[15] W. Wang, J. P. Attanucci, and N. H. M. Wilson, "Bus passenger origin-destination estimation and related analyses using automated data collection systems," *Journal of Public Transportation*, vol. 14, no. 4, pp. 131–150, 2011.

[16] P. G. Furth, "Integration of fareboxes with other electronic devices on transit vehicles," in *Transportation Research Record*, pp. 21–27, 1996.

[17] A. Lu and A. Reddy, "Algorithm to measure daily bus passenger miles using electronic farebox data for national transit database section 15 reporting," *Transportation Research Record*, no. 2216, pp. 19–32, 2011.

[18] *Comprehensive Transportation Annual Report of Shanghai*, 2015.

[19] M. Dunlap, Z. Li, K. Henrickson, and Y. Wang, "Estimation of origin and destination information from bluetooth and Wi-Fi sensing for transit," *Transportation Research Record*, no. 2595, pp. 11–17, 2016.

[20] R. G. Mishalani, M. R. McCord, and T. Reinhold, "Use of mobile device wireless signals to determine transit route-level passenger origin–destination flows," *Transportation Research Record*, no. 2544, pp. 123–130, 2016.

[21] I. W. Group, "Ieee standard for information technology–telecommunications and information exchange between systems–local and metropolitan area networks–specific requirements–Part 11: Wireless Lan Medium Access Control (Mac) and Physical Layer (Phy) Specifications Amendment 6: Wireless Access in Vehicular Environments," *IEEE Std*, pp. 802-11p, 2010.

[22] Apple Inc., http://devstreaming.apple.com/videos/wwdc/2014/715xx4loqo5can9/715/715_user_privacy_in_ios_and_os_x.pdf

[23] S. Coman, *Say Zebra*, LA Theatre Works, 2012.

[24] Y. Ji, R. Mishalani, M. McCord, and P. Goel, "Identifying homogeneous periods in bus route origin-destination passenger flow patterns from automatic passenger counter data," *Transportation Research Record*, no. 2216, pp. 42–50, 2011.

[25] M. L. Hazelton, "Statistical inference for transit system origin-destination matrices," *Technometrics*, vol. 52, no. 2, pp. 221–230, 2010.

[26] M. Ben-Akiva, P. Macke, and P. Hsu, "Alternative methods to estimate route-level trip tables and expand on-board surveys," *Transportation Research Record*, no. 1037, pp. 1–11, 1985.

[27] A. Gelman et al., *Bayesian Data Analysis*, Chapman & amp, Boca Raton, Florida, 2nd edition, 2004, edition, Chapman Hall/CRC.

[28] J. Berger, *Statistical Decision Theory and Bayesian Analysis*, Springer-Verlag, New York, USA, 1985.

[29] G. Marsaglia and J. C. Marsaglia, "A new derivation of Stirling's approximation to N," *American Mathematical Monthly*, pp. 826–829, 1990.

[30] M. Hollander and D. A. Wolfe, *Nonparametric Statistical Methods*, John Wiley & Amp; Sons, Inc: Danvers, MA, 2 edition, 1999.

[31] A. Karlsson, *Mathematical Statistics and Data Analysis*, vol. 22, 2007.

[32] Y. Ji, X. Yang, and Y. Du, "Optimal design of a short-turning strategy considering seat availability," *Journal of Advanced Transportation*, vol. 50, no. 7, pp. 1554–1571, 2016.