Monitoring Public Transit Ridership Flow by Passively Sensing Wi-Fi and Bluetooth Mobile Devices

Ziyuan Pu, *Member, IEEE*, Meixin Zhu, Wenxiang Li, Zhiyong Cui, Xiaoyu Guo, and Yinhai Wang, *Senior Member, IEEE*

***Abstract*—Real-time public transit ridership flow and origin–destination (O–D) information is essential for improving transit service quality and optimizing transit networks in smart cities. The effectiveness and accuracy of the traditional survey-based methods and smart card data-driven methods for O–D information inference have multiple disadvantages in terms of biased results, high latency, insufficient sample size, and the high cost of time and energy. By considering the ubiquity of smart mobile devices in the world, monitoring public transit ridership flow can be accomplished by passively sensing Wi-Fi and Bluetooth (BT) mobile devices of passengers. This study proposed a system for monitoring real-time public transit passenger ridership flow and O–D information based on customized Wi-Fi and BT sensing device. By combining the consideration of the assumed overlapping feature spaces of passenger and nonpassenger media access control address data, a three-step data-driven algorithm framework for estimating transit ridership flow and O–D information is proposed. The observed ridership flow is used as the ground truth for evaluating the performance of the proposed algorithm. According to the evaluation results, the proposed algorithm outperformed all selected baseline models and the existing filtering methods. The findings of this study can help to provide real time and precise transit ridership flow and O–D information for supporting transit vehicle management and the quality of service enhancement.**

***Index Terms*—Origin–destination (O–D) information, real-time monitoring system, transit ridership flow, Wi-Fi and Bluetooth (BT) passive sensing.**

I. Introduction

PUBLIC transit ridership flow and origin–destination (O–D) information are crucial for transit network planning, routes optimization, service quality improvements and travel scheduling [1]–[4]. It is also an essential data input of Internet of Vehicles (IoV) in transit systems [5]. Traveler surveys have been employed for acquiring such information in most of the previous research [6]. However, the effectiveness and efficiency of the survey-based methods are questionable in terms of high time and energy cost, and biased results [7]. As the smart card becoming widely used, some researchers have developed methods for estimating ridership flow and O–D information based on transit smart card data [8]–[13]. However, most of the transit smart card system only requires tap-in action, it is hard to infer alighting stops of each passenger real timely. Nowadays, it is stated that more than 80% of individuals carried at least one Wi-Fi and Bluetooth (BT) mobile device in daily life [14], [15]. Thus, estimating transit ridership flow and O–D information based on Wi-Fi and BT sensing data has the greatest potential to be a more reliable method.

The basic idea of the Wi-Fi and BT passive sensing technology is to capture the hardware media access control (MAC) address of the discoverable mobile devices through Wi-Fi management frame and BT slave response message [16]. If the Wi-Fi or BT communication function of a mobile device is turned on and no existing connections with access points or other devices through Wi-Fi or BT protocol, then the mobile device is in the discoverable mode. The hardware MAC address is a globally unique identifier, it is easy to monitor the boarding and alighting information of a passenger by identifying the MAC address of the passenger's mobile devices.

However, there are still two uncertainties that cause errors for Wi-Fi and BT sensing-based transit ridership flow monitoring. First, since the detection range of Wi-Fi and BT sensor is usually larger than the inside space of transit vehicles, the mobile devices outside transit vehicles are still possible to be detected. Thus, separating passengers' MAC address and nonpassengers' MAC address is crucial for estimating ridership flow and O–D information from Wi-Fi and BT sensing data. Previously, several studies shed light on solving this problem based on filtering methods [17], [18]. Several empirically predefined thresholds were selected to filter out the MAC addresses potentially coming from outside of transit vehicles. However, the results of most previous studies are barely convinced due to the lack of validation based on ground-truth

Manuscript received February 10, 2020; revised April 29, 2020 and May 27, 2020; accepted June 19, 2020. Date of publication July 27, 2020; date of current version December 21, 2020. This work was supported by Pacific Northwest Transportation Consortium (PacTrans) and the research project titled "Cooperative Perception of Road-Side Unit and Onboard Equipment with Edge Artificial Intelligence for Driving Assistance" under Award 69A3551747124, which is funded by Connected Cities with Smart Transportation (C2SMART) Center. *(Corresponding author: Yinhai Wang.)*

Ziyuan Pu, Meixin Zhu, Zhiyong Cui, and Yinhai Wang are with the Department of Civil and Environmental Engineering, University of Washington, Seattle, WA 98195 USA (e-mail: ziyuanpu@uw.edu; meixin92@uw.edu; zhiyongc@uw.edu; yinhai@uw.edu).

Wenxiang Li is with the Business School, University of Shanghai for Science and Technology, Shanghai 200093, China (e-mail: lwxffff@gmail.com).

Xiaoyu Guo is with the Zachry Department of Civil and Environmental Engineering, Texas A&M University, College Station, TX 77843 USA (e-mail: xiaoyuguo@tamu.edu).

Digital Object Identifier 10.1109/JIOT.2020.3007373

data. Since collecting ground truth of O–D information is costly and labor intensive, only a few studies provided the comparison of observed ridership flow and the estimated values [19]–[21]. The obvious gaps between the observed data and the filtering results indicated that the considerable errors are caused by such hard-threshold filtering methods. Hence, an accurate and effective method for separating the MAC address data belonging to passengers and nonpassengers is indemand.

The main disadvantages of the filtering method are the assumed clear boundary between the feature spaces of passenger and nonpassenger data, and the values determination of the thresholds. To our knowledge, the feature space of passengers and nonpassengers' MAC address data is overlapping. For example, when a transit vehicle travels with another vehicle side by side for a distance, the features of the passengers' MAC address data could be similar to the features of the mobile devices in the other vehicle. Thus, a Fuzzy C-Means (FCM) clustering algorithm is proposed for separating passenger and nonpassenger MAC addresses in this study. FCM is one of the most popular fuzzy-based clustering algorithms which is suitable for separating the clusters with ambiguous boundaries [22]. Unlike hard or crisp clustering algorithms, e.g., K-Means clustering, FCM allows objects to have the possibility for belonging to all groups with a certain degree of membership. Previously, FCM was used to deal with the ambiguous clusters in several scenarios in the intelligent transportation engineering area, e.g., ship trajectories clustering [23] and traffic volume-based road groups clustering [24].

Second, since only partial transit passengers carry discoverable mobile devices, a method targeting on estimating the population ridership flow is essential. Previously, several methods were implemented to estimate the population based on a sample of Wi-Fi and BT sensing data, including scaling with a fixed number [25], linear regression [26], power function, and Fourier function-based methods [27]. Among the existing methods, Lesani and Miranda-Moreno [27] conducted a performance comparison of power function and Fourier function for estimating the population number of pedestrians based on the detected Wi-Fi and BT MAC addresses. The proposed power function achieved a relatively high R-squared value than the Fourier function. In addition, the R-squared value of the proposed power function is much higher than the linear regression methods in other studies [28], which could be an indicator of the nonlinear relationship between the population and the number of detected MAC addresses. Thus, to handle nonlinearity among data sets, a random forest (RF) regression model [29] is proposed for estimating the population ridership flow in this study, including the number of onboard, boarding, and alighting passenger. For the performance comparison purpose, linear regression is selected as the baseline model to indirectly demonstrate the effectiveness of the RF model.

The primary objective of this study is to establish a system for monitoring real-time public transit ridership flow based on the Wi-Fi and BT sensing technology. A three-step data-driven algorithm framework for estimating the real-time transit ridership flow is proposed. The target parameters include the number of onboard, boarding, and alighting passenger, and O–D information. The main contribution of this research can be summarized as follows.

1) The system for monitoring real-time public transit ridership flow is designed based on the customized Wi-Fi and BT sensing devices.
2) A three-step data-driven algorithm framework for mining real-time transit passenger ridership flow from Wi-Fi and BT sensing data is proposed.
3) The proposed system is implemented on three transit routes in Seattle. The ground-truth data is collected manually for validating the performance of the proposed algorithms.
4) The performance of the proposed algorithm is compared with the existing filtering methods. The experimental results indicate the proposed algorithm can highly improve the estimation accuracy.

The remainder of this article is organized as follows. Section II introduces the system framework and the detailed information about the Wi-Fi and BT sensing device. Section III presents the proposed three-step data-driven algorithm. Section IV describes the experimental design and the numerical results are presented in Section V. This article is summarized by concluding the research findings and future research topics.

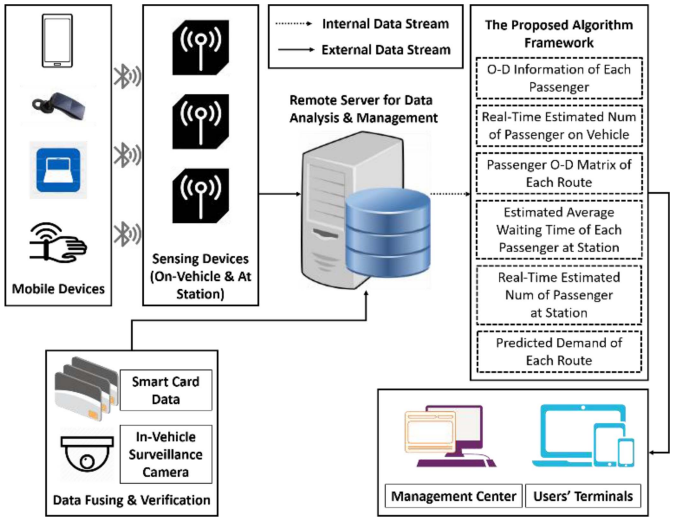

Fig. 1. System architecture of the real-time transit ridership flow monitoring system.

II. System Design

In this section, the proposed system and the customized Wi-Fi and BT sensing device are introduced in detail.

A. System Architecture

The system architecture is presented in Fig. 1. The discoverable Wi-Fi and BT mobile devices within the detection range of sensing devices can be detected, including mobile phones, laptops, BT earphones, etc. The real-time data will be transmitted from sensing devices to the remote data management and analysis server through cellular networks or Ethernet

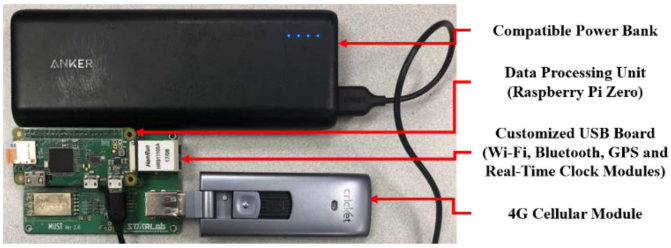

Fig. 2. Wi-Fi and BT sensing device.

cable connections. Besides, transit smart card data and in-vehicle surveillance camera data can be transmitted to the remote server for providing the ground-truth information to the system. Then, transit ridership flow and O–D information can be estimated through the proposed algorithm framework which will be introduced in Section III. By deploying the proposed system, the real-time information of transit operation can be delivered to the public in time, so that the public can optimize their travel plan based on real-time information. Furthermore, the real-time transit ridership flow also can be used to optimize vehicle dispatching and trip schedule. Generally, Wi-Fi and BT sensing devices can be installed in transit vehicles for monitoring passenger ridership flow or at transit stations for monitoring passenger waiting time and estimating the number of waiting passengers at stations. In this study, we only installed the Wi-Fi and BT sensing device in vehicles for monitoring passenger ridership flow. The detailed description of the customized Wi-Fi and BT sensing device will be introduced in Section II-B.

B. Customized Wi-Fi and Bluetooth Sensing Device

The Wi-Fi and BT sensing device are one of the most significant components of the proposed system in terms of hardware. Some existing products can sniff Wi-Fi and BT signals for traffic analysis at intersections, e.g., Acyclica RoadTrend [30]. However, it is quite different in the hardware part when the sensing device is installed on a moving vehicle, for instance, GPS recording, data communication, and power supply. Thus, a customized Wi-Fi and BT sensing device is designed which is presented in Fig. 2. The detailed information of each component is described as follows, which can guide the researchers and practitioners to build their implementation.

The Wi-Fi and BT sensing device are composed of four components, including sensing modules, data processing unit, communication module, and power supply.

Sensing Modules: There are four sensing modules are necessary, Wi-Fi module, BT module, GPS module, and real-time clock. In this study, all sensing modules are integrated into a customized PCB board which connects with the data processing unit through GPIO pins.

1) *Wi-Fi Module:* To capture the MAC address of Wi-Fi management frames, the Wi-Fi 802.11b/g/n module needs to set in the monitor mode [31]. In this study, Ralink 5370 Wi-Fi chipset is used. Its detection range is about 60 m and its frequency range is 2.4–2.4835 GHz.
2) *BT Module:* For sensing the MAC address in BT slave response messages, the BT module needs to keep sending out inquiry requests. BT 4.0 BCM20702 chipset is used in this study. The detection range is about 20 m.
3) *GPS Module:* To record moving features of the sensing device, a GPS module is employed to record the high-resolution latitude and longitude. U-blox 7020 chipset with −162 dBm tracking sensitivity is employed. The GPS module stores one data point per half-second. Each data point includes latitude, longitude, and timestamp.
4) *Real-Time Clock:* The data collection programs are parallelly running on the data processing unit through automatic start-up scripts. The MAC address data and GPS location matching are based on the timestamp. Most of the single-board computers have an embedded clock for time recording. However, once the power is off, the clock will stop running. If no Internet connection or manual time synchronization is employed, the clock will not be synchronized when the computer restarts. Even the GPS module can help with time synchronization, it still can ruin the data quality due to the signal related issues. Thus, the DS 3231 RTC real-time clock module was employed in this study to avoid the problems caused by time synchronization.

Data Processing Unit: Raspberry Pi Zero is employed as the data processing unit in this, which is a single-board computer with a 1.0-GHz single-core CPU and 512-MB RAM [32]. Other Internet-of-Things (IoT) device can be used for this kind of implementation, e.g., NVIDIA Jetson NANO, Asus Tinker Board, and Arduino Uno R3.

Communication Modules: 4G LTE Huawei USB Modem E397u-53 is employed as the hardware part of the data communication module. The T-Mobile data SIM card is plugged into the 4G modem, and the modem connects with the customized USB board via the USB interface. The connection of the device to the cellular network is activated via the Network Manager API in the software which allows automatic network connection upon start-up and automatic reconnection to the Internet whenever the connection fails. On-vehicle Ethernet or Wi-Fi service can be used as alternatives to the data communication module.

Power Supply: Portable Charger Anker PowerCore 20100-mAh power bank is used in this study. As the low energy consumption of the Raspberry Pi Zero, the employed power bank can support it for a one-day operation. Instead of power banks, the in-vehicle power supply also can be used with a voltage transformer.

III. Proposed Methodology

A. Algorithm Framework

The proposed algorithm framework is designed for mining real-time transit ridership flow using Wi-Fi and BT sensing data (see Fig. 3). Generally, the proposed algorithm is a three-step data-driven approach. Step one aims to extract the features and the vehicle moving features during the detection time of each MAC address. Then, MAC address data with their extracted features are used as the input of step two in which the FCM clustering algorithm is employed to cluster the MAC addresses into the passenger and nonpassenger clusters.

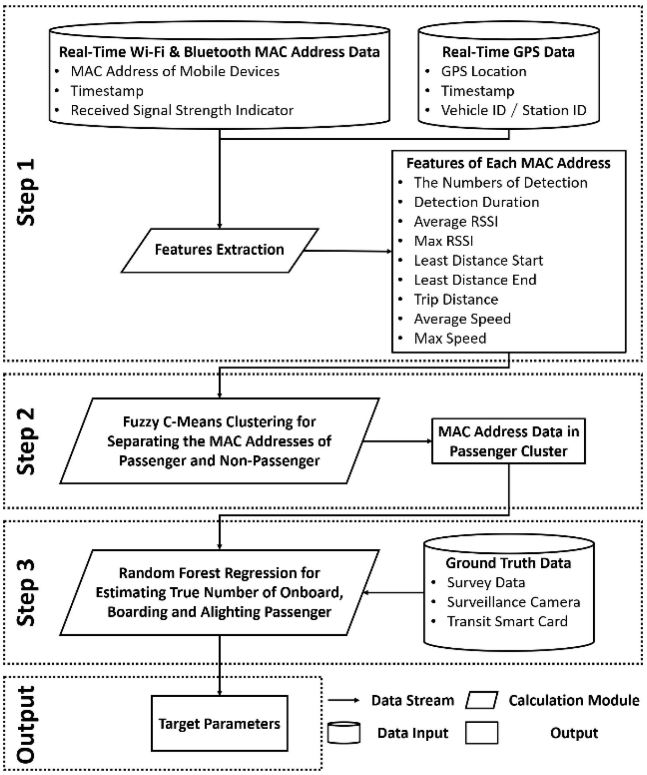

Fig. 3. Algorithm framework.

In step 3, the ridership flow population is estimated by the proposed RF regression model using the clustered passenger MAC addresses. The manual counted ground-truth data is used for training the proposed RF regression model in this study. Other data sources also can be alternatives for training algorithms, e.g., in-vehicle surveillance cameras and smart card data. In this study, the proposed algorithm framework estimates the population numbers of onboard, boarding, and alighting passengers of each stop. In addition, by accurately clustering the MAC addresses of passengers, the OD matrix of partial passengers also can be achieved. The following sections introduce the details of each step.

B. Feature Extraction

Since the detection range is not exactly the inside space of transit vehicle, the MAC addresses come from the mobile devices outside the vehicle can be detected as well. In the following situations, the MAC address of nonpassenger mobile devices can be detected by the onboard sensing device.

1) Fixed Wi-Fi or BT devices within the sensing range.
2) Mobile devices of the passengers standing at stations.
3) Mobile devices of the pedestrians or bicyclists within the sensing range.
4) Mobile devices in other vehicles within the sensing range.

For the fixed devices and the devices carried by the passengers standing at stations, the MAC address features are quite different with the passenger MAC address. Intuitively, the MAC address should be detected by only a few times and in a short time period. For the mobile devices in other vehicle or of pedestrian and bicyclists, even they travel parallelly with transit vehicle, the MAC address features also would be different from passenger MAC addresses, e.g., the location of the first and the last detection could be far away from the nearest stations.

To depict the features of each MAC address, nine features were extracted from the MAC address and GPS data which are presented in Table I. The features were categorized into two parts, MAC address features and vehicle moving features, respectively. MAC address features contain detection times, detection duration, average RSSI, and maximum RSSI. Travel distance, average speed, maximum speed, and the distances of to the nearest station when the MAC address is first and last detected are the five features that describe the vehicle moving features during the detection time of each unique MAC address. In this study, matching MAC address data and GPS data is finished on the edge side. Then, the MAC address point with GPS location is transmitted from the sensing device to the remote server for feature extractions. In this step, the output is the vectors of each MAC address and its features.

C. Separating Passenger and Nonpassenger MAC Addresses Using Fuzzy C-Means Clustering

Other than hard or crisp clustering algorithms, the fuzzy clustering algorithm assigns a certain degree of membership to a data point for all clusters, which indicates the data point can belong to any cluster [33]. Thus, fuzzy clustering algorithms usually are useful when the boundaries among clusters are ambiguous [34], which satisfies the characteristics of the overlapped feature spaces of passengers and nonpassengers. FCM clustering is one of the most popular fuzzy clustering algorithms. Previously, scholars implemented original and modified FCM clustering algorithms in multiple applications, e.g., image segmentation [35], sensor network optimization [36], stock performance prediction [37], and medical analysis [38]. It attempts to minimize the cost function J in (1) which is the summation of the membership function of each data point. The membership function only depends on the distance to the center of each cluster. Then, assign each data point to the closest cluster in terms of the membership function. Let $\chi = (X_1, X_2, X_3, \ldots, X_N)$ denotes a set of N MAC address to be partitioned into C clusters. $X_j = (x_1, x_2, x_3, \ldots, x_n)$ denotes n features of each MAC address. Then, the cost function J would be calculated as the following equation:

$$J = \sum_{j=1}^{N} \sum_{i=1}^{C} u_{ij}^{m} \left\| X_j - v_i \right\|^2 \tag{1}$$

where m is the parameter for controlling the fuzzification, $u_{ij} \in [0, 1]$ is the membership function of the jth data point in cluster i, which represents the possibility that the jth MAC address data point whether belongs to a passenger or not, thus, $\sum_{i=1}^{C} u_{ij} = 1 (j = 1, 2, \ldots, N)$. v_i is the center of the ith cluster, and $\|*\|$ is the similarity function of data point X_j and the cluster center v_i. For the similarity function selection, the Euclidean distance is employed as the similarity function in this study since it can reflect the attributes of the most extracted features. All extracted features are normalized for the similarity calculation. The cost function J is minimized

TABLE I
EXTRACTED FEATURES FOR CHARACTERIZING EACH UNIQUE MAC ADDRESS

Categories	Features	Definition
MAC Address Features	Detection Times	The number of times a unique MAC address is detected (Times)
	Detection Duration	The total amount of time for a unique mac to be detected (Seconds)
	Average RSSI	The average value of received signal strength indicator of each MAC (dBm)
	Maximum RSSI	The maximum value of received signal strength indicator of each MAC (dBm)
Vehicle Moving Features	Least Distance Start	The distance to the nearest station when MAC address is first detected (Meters)
	Least Distance End	The distance to the nearest station when MAC address is last detected (Meters)
	Travel Distance	The total travel distance of the vehicle between the first and the last detection of a unique MAC address (Meters)
	Average Speed	The average speed of the vehicle between the first and the last detection of a unique MAC address (Meters/Second)
	Maximum Speed	The largest speed of the vehicle between the first and the last detection of a unique MAC address (Meters/Second)

Algorithm 1 Training Process of FCM

Initialization:
The number of clusters C
The maximum number of iterations L
The fuzzification parameter m
Randomly Select the values of each cluster center $v_i^{(0)}$
Estimate $U^0 = \left[u_{ij}^0\right]$ using (2), U^0 is a $C \times N$ matrix
Repeat:
Update $v_i^{(t)}$ using (3) based on U^{t-1}
Compute U^t using (2) based on $v_i^{(t)}$
Until: $\left\|U^t - U^{t-1}\right\| \leq \varepsilon$ or $t \geq L$

when the data points closer to the center of their clusters are assigned with higher membership values than the assigned values of the data points far from the centroid. The solution of the minimized cost function J can be achieved by the following equations:

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{\|X_j - v_i\|}{\|X_j - v_k\|}\right)^{2/(m-1)}} \tag{2}$$

$$v_i = \frac{\sum_{j=1}^{N} u_{ij}^m X_j}{\sum_{j=1}^{N} u_{ij}^m}. \tag{3}$$

Initially, the centers of each cluster are randomly selected. Then, the membership function and the centers are updated until the cost function is converged. The training process of FCM clustering is presented in Algorithm 1. For the parameter settings, previous studies demonstrated that user-defined parameters highly influence the performance of algorithms, and several existing methods were designed to determine the hyperparameter settings for achieving optimal performance [39]–[41]. In this study, the parameters are set based on the results of previous studies. The fuzzification parameter m is set as 2 [42], the number of clusters C is set to 2 representing the clusters of passengers and nonpassengers. ε and L in the algorithm are set as 0.001 and 1000, respectively.

D. Estimating the Population Number of Onboard, Boarding, and Alighting Passengers of Each Stop Using Random Forest Regression

Considering that only partial passengers have discoverable mobile devices, the ridership flow population needs to be estimated based on the clustered passenger MAC address data. In this study, the RF regression model was employed for the estimation task. RF regression is a widely used nonparametric machine learning regression algorithm. As shown in the previous study, RF regression can capture the nonlinear relationship in the data set which results in a better goodness of fit than linear regression [43]–[45]. The general concept of RF is introduced by Breiman in 2001 [29]. In this study, the classification and regression trees (CART) algorithm [46] is used for trees development. Once a CART tree has been built, the branches which do not contribute to the predictive performance of the tree will be pruned for avoiding overfitting. However, if the CART trees are used in the RF, the pruning process will be ignored since the generalization error of a RF will always converge.

For the RF regression model developed in this study, five variables were selected as the regressors, including the day of week, the hour of the day, the minute of the hour, the dummy variable of whether the current stop is the last stop of the trip, and the number of passenger MAC addresses.

E. Algorithms Evaluation

1) Evaluation of the Results of Fuzzy C-Means Clustering: To evaluate the FCM clustering algorithm, Gaussian mixture model (GMM) and a Bayesian approach to GMM (BGM) were selected as the baseline models, since GMM and BGM are mixture density-based clustering algorithms which are also suitable for the data set with ambiguous boundaries. GMM is good at forming smooth approximation to arbitrarily shaped of the probability density and at scaling with the dimensionally of data [47]. BGM optimizes the selection of the number of components in the model as well as the partition data sets by automatically penalizing the overcomplex model [48], which could further improve the performance of the GMM model. In addition, BGM can avoid overfitting by eliminating parameters using integration [49]. The model specification can be found in [47] and [50]. Then, four metrics are employed for evaluating clustering performance. The following paragraphs introduce the metrics in detail.

External and internal clustering validation are two main categories of clustering validation methods [51]. The major difference is whether external information would be used for validation. For unsupervised clustering algorithms, internal

clustering validation is the only option due to the lack of available labeling information [52]. Compactness and separation are two main criteria for evaluating cluster similarity. Compactness measures the intradistance of each cluster and separation measures interdistance [53], [54]. The following metrics are employed to measure compactness and separation, including Silhouette coefficient (SC), Dunn's index, Davies–Bouldin (DB) index, and Beta CV measurement.

SC [55] evaluates the performance of clustering result based on the pairwise difference of inter and intra distances of clusters, which is simply expressed as

$$\mathrm{SC} = \frac{b(j) - a(j)}{\max\{a(j), b(j)\}} \tag{4}$$

where $a(j)$ is the average distance between the ith sample and all samples which are included in a given cluster C_j, and $b(j)$ is the minimum average distance between the ith sample and all samples of a given cluster $C_k (k \neq j)$. The value of SC ranges in $[-1,\ 1]$. A large SC value infers better clustering results.

Dunn's (DU) index is dedicated for identifying sets of compact and well separated clusters by maximizing intercluster distances whilst minimizing intracluster distances [56]. Dunn's validation index is calculated as

$$\mathrm{DU} = \min_{1\le i\le c}\left\{\min_{\substack{1\le i\le C\\ j\neq i}}\left\{\frac{\delta\left(C_i, C_j\right)}{\max_{1\le k\le c}\{\Delta(X_k)\}}\right\}\right\} \tag{5}$$

where $\delta(C_i, C_j)$ measures the intercluster distance between C_i and C_j, $\Delta(X_k)$ defines the intracluster distance of X_k, and C is the number of clusters. A larger value of Dunn's index implies better clustering results.

DB index [57] is the ratio of the sum of intracluster distance to intercluster separation, which is expressed by

$$\mathrm{DB} = \frac{1}{k}\sum_{i=1}^{k}\max_{i\neq j}\frac{\left(D(C_i) + D\left(C_j\right)\right)}{D\left(v_i, v_j\right)} \tag{6}$$

where $D(v_i, v_j)$ is the intercluster distance between the centers of clusters C_i and C_j and $D(C_i)$ is the intracluster diameter of the cluster C_i. The lower the DB value, the better the clustering results.

Beta CV measure (Beta CV) [**?**] is a measurement of clustering validation based on the ratio of the mean intracluster distance to the mean intercluster distance which can be calculated as [58]

$$\text{Beta CV} = \frac{\mathrm{Distance}_{\mathrm{intra}}/N_{\mathrm{intra}}}{\mathrm{Distance}_{\mathrm{inter}}/N_{\mathrm{inter}}} \tag{7}$$

where N_{intra} is the number of distinct intracluster edges, N_{inter} is the number of distinct intercluster edges.

2) Evaluation of the Results of Random Forest Regression: To evaluate the performance of the RF regression model, traditional linear regression model is developed based on the same variables for the comparison purpose. Mean absolute error (MAE), mean square error (MSE), and mean absolute percentage error (MAPE) are used as the evaluation metrics. The following equations present the metrics formulation:

$$\mathrm{MAE} = \frac{\sum_{i=1}^{N}\left|\hat{Y}_i - Y_i\right|}{N} \tag{8}$$

$$\mathrm{MSE} = \frac{\sum_{i=1}^{N}\left(\hat{Y}_i - Y_i\right)^2}{N} \tag{9}$$

$$\mathrm{MAPE} = \frac{1}{N}\sum_{i=1}^{N}\frac{\left|\hat{Y}_i - Y_i\right|}{Y_i} \times 100\% \tag{10}$$

where $\hat{Y}_i$ is the estimated number of onboard, boarding, or alighting passenger of stop i, Y_i is the ground-truth value, and N is the number of stops in the testing data set. Typically, the MAE presents a measure of the average misprediction of the model, the MSE is used to measure the error associated with a prediction, and the MAPE usually expresses accuracy as a percentage. The model with a smaller value of MAE, MSE, and MAPE performs better in the prediction of observed data.

3) Comparison With the Existing Filtering Methods: For the relevant existing studies, the filtering method was employed for processing the Wi-Fi and BT sensing data to estimate the public transit ridership flow [17]–[19], [59]–[63]. RSSI, the number of received packets of each MAC, detection duration of each MAC, distance of the first and the last detection to the nearest bus station, and vehicle speed while a MAC been detected were the main parameters for filtering the MAC address data. In order to compare the performance of the proposed algorithms, two filtering methods were selected as the representatives for the comparison purpose. The selected filtering methods were considered more comprehensive than others in terms of the number of filters and how the thresholds of each filter were determined. The number of onboard passengers at each stop was estimated by the proposed algorithm framework and two existing filtering methods. The detailed description of the selected filtering methods is presented as follows.

Filtering Method 1: Dunlap *et al.* [17] developed a three-step filtering method for separating passengers and nonpassengers. The MAC address which fits any following conditions would be considered as a nonpassenger MAC address: 1) detection times is lower than 3 for Wi-Fi MAC address and 1 for BT MAC address; 2) detection duration is less than 60 s; and 3) the distances of vehicle to the nearest station when the MAC address is first and last detected are larger than 600 ft (183 m) for Wi-Fi and 300 ft (91 m) for BT. The first and the last stops of the trip are determined by the stations which are the nearest stops to the vehicle when the MAC address is first and last detected.

Filtering Method 2: Mishalani *et al.* [18] defined a filtering method with four filters. If the features of a unique MAC address meet the following rules which is considered as a nonpassenger MAC address: 1) detection duration is less than 3 min; 2) maximum signal strength is lower than 20th of the cumulative distribution of observed signal strengths; 3) total travel distance is less than 900 ft (274 m); and 4) total number of detected signals per mile is less than 10. The first and last detected time of each MAC, the distance between the sensor

TABLE II
STATISTICAL SUMMARY OF THE DATA SET

Routes No.	Trip Date	Trip Start Time	Trip End Time	The Number of Stops	Number of Data Points		Number of Unique MAC	
					Wi-Fi	Bluetooth	Wi-Fi	Bluetooth
372	3/6/2018	7:35:00	8:32:00	21	2550	344	431	29
	3/6/2018	10:51:00	11:49:00	24	2055	344	854	53
	3/1/2018	11:03:00	11:51:00	28	3547	346	819	21
32	11/4/2018	16:55:37	17:21:57	12	904	172	294	8
	11/9/2018	18:40:49	19:26:26	24	2166	152	815	29
	11/9/2018	19:38:58	20:05:48	15	918	86	165	13
67	11/4/2018	15:05:15	15:47:26	27	1879	122	747	20
	11/8/2018	15:05:19	15:33:50	21	1351	88	555	18
	11/8/2018	15:38:10	16:04:44	19	657	125	179	14

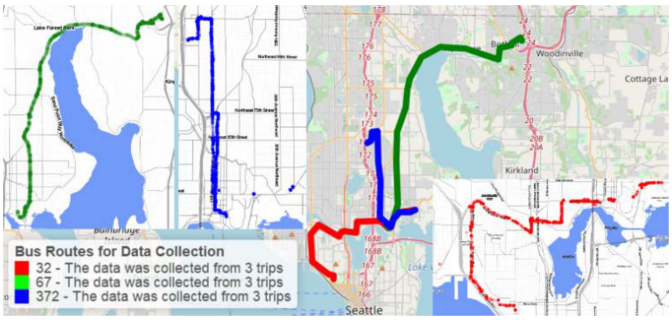

Fig. 4. Study area.

and stops nearby and a predefined threshold of the maximum sensor detection range of 200 ft (61 m) are used to determine the boarding and alighting stops for each MAC address.

IV. EXPERIMENTAL DESIGN

The data used in this study were collected from nine trips of three routes in Seattle. The detailed description of the study area and statistical summary of the data set are introduced in the following sections.

A. Study Area and Data Collection

The study area is three transit routes in the north of King County, including route 32, route 67, and route 372. Fig. 4 shows the three routes on the map and the GPS data points as well. Route 67 depicted in blue runs from University District to Northgate Transit Center, route 32 highlighted in red operates from Queen Avenue to Sand Point Way, and route 372, marked in green, provides service along the route from Bothell to University District. The data were collected from three trips of each route by the customized Wi-Fi and BT sensing device. For each trip, the sensing device was carried by a volunteer seating in the middle of the vehicle. The sensing device was powered on when the volunteer got seated and powered off once the vehicle arrived at the last stop or the volunteer took off the vehicle.

B. Statistical Summary of the Data Set

Table II shows the statistical summary of the data set. There are nine trips were traveled for collecting data. The number of stops is different from trip to trip. Since the vehicles only stop at the stations with waiting passengers or have onboard passengers requesting for taking off. Only the stations where the vehicle stopped were counted as stops in the data set. Besides the trip information, the amount of the MAC address collected from each trip also are introduced. Based on the statistical summary, 17 806 data points were collected, including 16 027 Wi-Fi data points and 1779 BT data points. The huge difference between the amount of Wi-Fi and BT data is caused by the amount of discoverable Wi-Fi and BT devices and data frame transmission frequency of Wi-Fi and BT protocols. Totally, 5064 unique MAC addresses were detected, including 4859 via Wi-Fi network and 205 via BT network. Based on the data set, averagely, one unique Wi-Fi MAC address is collected out of four Wi-Fi data points and one unique BT MAC address is collected out of ten BT data points.

TABLE III
EVALUATION OF CLUSTERING ALGORITHMS FOR SEPARATING PASSENGER AND NONPASSENGER MAC ADDRESS

Metrics	Fuzzy C-means	Bayesian Gaussian Mixture	Gaussian Mixture
SC	0.74289	0.65651	0.63654
DU	0.00021	0.00007	0.00005
DB	0.67708	0.79231	0.81318
Beta CV	0.16561	0.21994	0.23426

V. NUMERICAL RESULTS

A. Separating Passenger and Nonpassenger MAC Addresses Using Fuzzy C-Means Clustering

The raw Wi-Fi and BT MAC address data along with the GPS data were used to extract the proposed features of each MAC address. The FCM clustering was conducted to cluster each MAC address into passenger or nonpassenger clusters. The metrics of each model are presented in Table III. According to the evaluation metrics, the FCM clustering model outperformed all models in terms of achieving the highest value of SC and DU and the lowest value of Beta CV and DB, which indicates the clusters were separated well by the FCM clustering. The BGM and GM models had similar performance according to the closing value of all four metrics.

TABLE IV
STATISTICAL SUMMARY OF PASSENGER AND NONPASSENGER CLUSTERS

Features	Passenger				Non-Passenger			
	Min	Max	Mean	S.D.	Min	Max	Mean	S.D.
Detection Times	2.00	1021.00	20.89	69.24	1.00	31.00	1.26	1.71
Detection Duration (Seconds)	1.00	3060.00	418.78	679.08	0.00	1253.00	4.24	62.89
Average RSSI (dBm)	-88.00	-22.19	-56.65	14.17	-91.00	-39.00	-61.79	12.10
Max RSSI (dBm)	-85.00	-17.00	-50.62	15.01	-91.00	-37.00	-61.23	12.30
Least Distance Start (Meters)	1.64	2306.17	152.19	213.12	3.22	1064.48	324.49	213.49
Least Distance End (Meters)	2.18	1722.00	144.31	195.75	3.22	1064.48	325.57	213.74
Trip Distance (Meters)	8.94	20442.36	2409.68	4181.46	0.00	184.03	1.77	14.61
Average Speed (Meters/Second)	0.44	30.29	7.72	6.43	0.00	8.25	0.11	0.77
Max Speed (Meters/Second)	0.44	79.70	8.34	10.71	0.00	31.98	0.27	2.18

TABLE V
EVALUATION OF THE ESTIMATED NUMBER OF ONBOARD PASSENGERS

Methods	Fuzzy C-means			Bayesian Gaussian Mixture			Gaussian Mixture		
	MSE	MAE	MAPE	MSE	MAE	MAPE	MSE	MAE	MAPE
Linear Regression	20.29	3.26	28.96	23.54	3.46	33.86	27.79	3.43	34.49
Random Forest	14.61	2.08	11.27	22.61	3.25	31.02	10.36	2.50	32.09

Totally, 5064 unique MAC address were clustered by the FCM clustering algorithm into two clusters with 399 passenger MAC addresses and 4665 nonpassenger MAC addresses. Based on the FCM clustering results, the statistical summary of each feature is presented in Table IV. The mean values of the detection times and the detection duration of passenger MAC address are much larger than those of nonpassenger MAC address. The nonpassenger MAC addresses have 1.26 average detection times and 4.24-s detection duration which is consistent with the assumption that nonpassenger MAC address should be detected for few times and in a short time window. The average RSSI and the max RSSI of passenger MAC address is larger than those of nonpassenger for all four numbers, which is reasonable that the signal strength of nonpassenger's mobile device might be influenced by the bodyshell of the transit vehicle or the larger distance from the sensing device. The Least Distance Start and Least Distance End of passenger MAC addresses are about 200 m which is smaller than those of nonpassenger MAC addresses. It is explicable that the passenger MAC addresses are more likely to be detected around the station for the first and the last detection, and nonpassenger MAC addresses are more likely to be detected during the trip where the vehicle is far away from stations. However, since the nonpassengers waiting for other vehicles at the station are possible to be detected, several MAC addresses are close to the stations for the first and the last detections are still considered as nonpassenger. Other three vehicle moving features of passenger MAC address, including trip distance, average speed, and maximum speed, have higher mean values than those of nonpassengers for all four numbers. The mean values of these three features of nonpassengers' MAC addresses are close to zero, which indicates the vehicle almost halted during the time period when the MAC addresses of nonpassengers were detected. It is noted that the maximum values of average speed and the max speed of passenger are unreasonably high, which is caused by unstable GPS data.

B. Estimating the Population Number of Onboard, Boarding, and Alighting Passenger of Each Stop

After separating the passengers MAC address from the data set, the boarding and alighting stations of each passenger MAC address were assigned as the stations with the smallest distance to the vehicle for the first and the last detection. The total number of onboard, boarding, and alighting passengers of each stop were estimated based on the FCM clustering results. Then, the data were divided into training data and testing data with a portion of 7:3 for developing the proposed RF regression model as well as the linear regression model. The manual counting number of onboard, boarding, and alighting passengers of each stop was used as the ground truth for calculating MAE, MSE, and MAPE. In order to demonstrate the clustering results of the FCM, the total number of onboard, boarding, and alighting passenger of each stop was also counted based on the BGM and GM clustering results.

First, only the number of onboard passengers was estimated. The evaluation results are presented in Table V. According to the evaluation results, the estimated results based on the FCM clustering performed better than all other baseline models in terms of the smallest values of MSE, MAE, and MAPE for both the estimations of the linear regression and the RF regression except the MSE of the Gaussian mixture algorithm in RF regression case. The potential reason is that the passenger flow estimation based on the Gaussian mixture algorithm with RF regression might achieve more accurate results for the stations with a large number of passengers. However, since the overall estimated performance is not as accurate as FCM in the case of RF regression, the overall estimated error of the Gaussian mixture algorithm in the RF regression case is still higher than

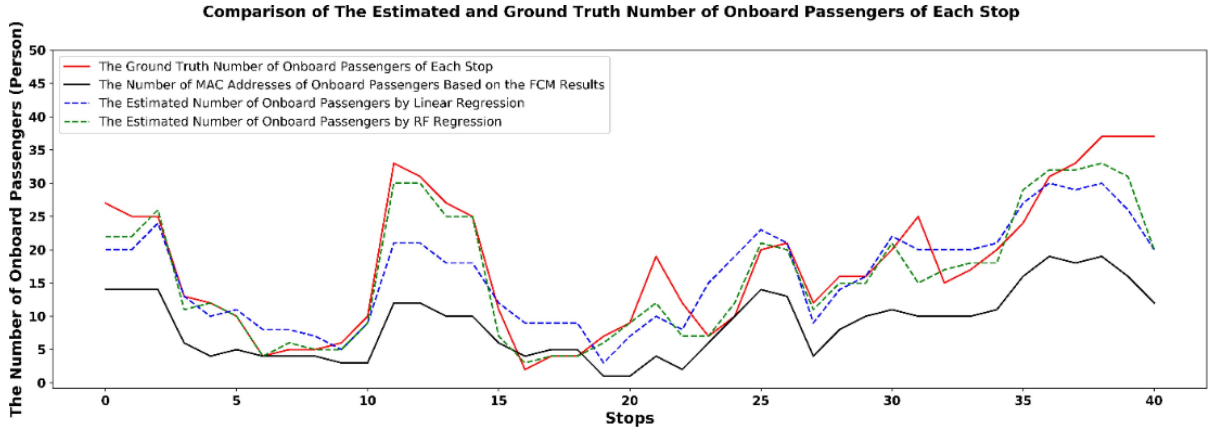

Fig. 5. Comparison of the number of clustered onboard passenger MAC addresses based on FCM clustering, estimation results of the population number of onboard passengers by linear regression and RF regression, and ground-truth data.

TABLE VI
COMPARISON OF THE PROPOSED ALGORITHM AND THE EXISTING FILTERING ALGORITHMS

Methods	Fuzzy C-means			Filtering Method 1 [17]			Filtering Method 2 [18]		
	MSE	MAE	MAPE	MSE	MAE	MAPE	MSE	MAE	MAPE
Linear Regression	26.29	3.26	28.96	52.05	5.32	51.25	67.78	6.62	58.62
Random Forest	14.61	2.08	11.27	35.16	3.84	36.47	30.03	3.76	27.5

the FCM in RF regression case. Furthermore, the estimated performance of the proposed RF regression algorithm is more accurate than that of the linear regression model for the estimation based on all three clustering algorithms. MSE, MAE, and MAPE of the RF regression model are highly smaller than those of the linear regression model. The estimated number of onboard passengers of each stop based on the FCM clustering results and the ground truth is visualized in Fig. 5.

The black solid line is the number of clustered passenger MAC addresses of each stop based on the FCM clustering. The red solid line is the ground-truth number of onboard passengers of each stop. For most of the stops, the number of passenger MAC addresses is a small proportion of the ground truth, and it can effectively reveal the trend of the ground truth. The blue dashed line presents the estimated number of onboard passengers based on the estimation of linear regression. By employing linear regression, the number of passenger MAC addresses were enlarged with a fixed proportion. The green dashed line shows the estimation results by RF regression, which is highly close to the ground truth and even superposed the red line for some stops. By capturing the nonlinear relationship between the number of passenger MAC addresses and the ground truth, the RF regression model achieved more accurate estimation of the population number of onboard passengers.

C. Comparison With the Existing Filtering Methods

The estimation results of RF regression and linear regression using the filtering results as the inputs are compared with the estimations based on FCM clustering results in this section. Table VI shows the evaluation results. Consistent with the previous evaluation results, the RF regression model performed better than the linear regression for all metrics. Among the existing filtering methods, Filtering Method 2 achieved a better performance than Filtering Method 1 in the case of RF regression, and the results in the case of linear regression is opposite. The estimation performance based on the FCM results improved a lot compared with the two existing filtering algorithms. It is demonstrated that the MAC addresses of passenger and nonpassenger are hard to be well separated by filters. By considering the overlapped feature spaces of passenger and nonpassenger, the FCM clustering algorithm effectively separated the MAC addresses of passenger and nonpassenger. Furthermore, the RF regression model effectively estimated the population number of onboard passengers by capturing the nonlinearity.

The scatter plots of the ground truth versus the estimated number of onboard passengers based on RF regression using FCM results and two filtering algorithms are presented in Fig. 6. According to the figure, the dots in the plots of filtering algorithms are dispersed around the diagonal line. For Filtering Method 1, most of the dots are above the diagonal line, which indicates the MAC addresses were more likely to be separated into the nonpassenger cluster so that the number of onboard was underestimated. The potential reason is that the Filtering Method 1 is inclined to separate passenger into the nonpassenger cluster, e.g., the GPS location was recorded every 20 s so that the distance of the vehicle to the nearest station is possible to be larger than the detection range for the first detection of a passenger MAC address. For the results of Filtering Method 2, most of the dots are beneath the line, which indicates the algorithm overestimated the number of onboard passengers. The explanation could be the filter for filtering signal strength was apt to separate the nonpassenger MAC addresses to the passenger cluster since the distribution of signal strength of nonpassenger MAC address is similar to the distribution of passenger MAC address. The rightmost

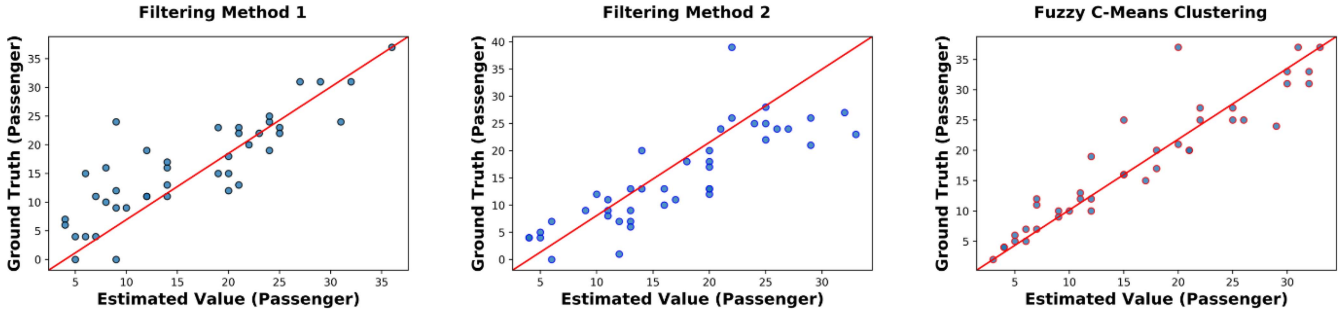

Fig. 6. Scatter plot of ground truth versus the estimated number of onboard passengers based on RF regression using FCM clustering and filtering algorithms' results.

TABLE VII
EVALUATION OF THE ESTIMATED NUMBER OF BOARDING AND ALIGHTING PASSENGER OF EACH STOP

Estimations	MSE	MAE	MAPE
Estimating the Number of Boarding Passengers of Each Stop	0.86	0.50	14.72
Estimating the Number of Alighting Passengers of Each Stop	0.96	0.54	17.41

scatter plot presents the estimation using the FCM results. The dots are more concentrated around the diagonal line than others. It is noted that the estimation is more accurate for the small number of passengers. As the number increased, the error became more considerable. The potential reason could be the insufficient data point with the large value in the training data set.

Besides the number of onboard passengers, the numbers of boarding and alighting passengers of each stop were also estimated based on the RF regression model using FCM clustering results. The estimation performance was evaluated by the three metrics calculated based on the manual counting numbers of boarding and alighting passengers of each stop, which is presented in Table VII. According to the evaluation results, the estimated numbers of boarding and alighting passengers are acceptable in terms of the small value of MSE, MAE, and MAPE. It is noted that the MAPE of estimated numbers of both boarding and alighting passengers are higher than the MAPE of the estimated number of onboard passengers, which is potentially caused by numerous zero values of the number of boarding and alighting passengers in the data set.

D. Estimating the Ridership Flow and O–D Information of the Selected Transit Trip

Based on the proposed algorithm framework, the transit demand can be monitored by the estimated numbers of onboard, boarding, and alighting passengers and O–D information from Wi-Fi and BT sensing data. In order to further demonstrate the feasibility of the proposed method, the ridership flow and O–D matrix of a selected trip were estimated based on the proposed algorithms. The results are presented in Tables VIII and IX. The selected trip was traveled on November 9, 2018 from 19:38:58 to 20:05:48. Totally, the transit vehicle stopped at 15 stations during the trip.

Table VIII presents the O–D matrix of the passenger MAC addresses by the FCM clustering algorithm. Even only partial O–D information can be achieved, the main trend of the travel demand can be achieved. Besides the O–D matrix, the numbers of boarding and alighting passengers were estimated using the RF regression model. The RF regression was trained by the data set which is collected from other trips. The ground-truth numbers of boarding and alighting passengers of each stop are also presented in the table. Table IX shows the estimated number of onboard passengers of each stop and the ground truth as well. The estimated errors are negligible for the most stops. However, the estimation errors were relatively large for the last two stops. Since the sensing device was powered off before the trip ended for the selected trip so that the MAC address data quality was influenced for the last two stops. Therefore, the zero number of MAC addresses for the last stops is the main reason for the large error.

By successfully capturing the partial O–D matrix, the numbers of onboard, boarding, and alighting passengers of each stop, the public transit demand can be achieved. Based on the output parameters of the proposed system, it is easy to observe which parts of the trip have more travel demands and which stops are more popular for the traveler.

VI. CONCLUSION

In summary, this study proposed a real-time system for monitoring public transit ridership flow based on the customized Wi-Fi and BT sensing device. For the methodology, a three-step data-driven approach is developed for mining the transit ridership flow and O–D information from Wi-Fi and BT sensing data, including feature extraction for characterizing MAC address data, FCM clustering algorithm for separating MAC address of passenger, and RF regression for estimating the population number of ridership flow. To demonstrate the effectiveness and efficiency of the proposed algorithm, GMM and BGM were selected as the baseline models for evaluating FCM clustering, and linear regression was selected for evaluating RF regression. The comparison of the proposed algorithm with the existing filtering methods was conducted as well. The MAC address data was collected by the customized Wi-Fi and BT sensing device from nine trips of three transit routes in Seattle. Multiple evaluation metrics were calculated based on ground-truth data and the estimates to quantitatively evaluate the estimation performance. According to the results,

TABLE VIII
O–D MATRIX OF THE SELECTED TRIP

Boarding \ Alighting	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Total Boarding MAC	Total Ground Truth Boarding	Total Estimated Boarding
1	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	2	2	3
2		0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	2	0
3			0	1	0	1	0	0	0	0	0	0	0	0	0	2	1	2
4				0	0	0	0	0	0	1	0	0	0	0	0	1	0	1
5					0	1	0	0	0	0	0	0	0	0	0	1	1	1
6						0	1	0	0	0	0	0	0	0	0	1	1	1
7							0	3	1	0	0	0	0	0	0	4	2	3
8								0	0	0	0	0	0	0	0	0	1	3
9									0	0	0	0	2	0	0	2	0	2
10										0	1	1	0	0	0	2	2	1
11											0	1	0	0	0	1	0	0
12												0	0	0	1	1	1	2
13													0	1	0	1	0	3
14														0	0	0	3	2
15															0	0	0	0
Total Alighting MAC	0	0	0	1	0	2	1	3	1	1	1	2	2	3	2	19 / 19		
Total Ground Truth Alighting	0	0	0	0	1	2	1	0	1	0	1	0	2	2	6		16 / 16	
Total Estimated Alighting	0	0	1	1	1	2	1	1	2	2	1	2	3	2	5			24 / 24

TABLE IX
NUMBER OF ONBOARD PASSENGERS OF THE SELECTED TRIP

Stops	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Ground Truth Onboard Passenger	2	4	5	5	5	4	5	6	5	7	6	7	5	6	6
Onboard MAC of Each Stop	2	3	5	5	6	5	8	5	6	7	7	6	5	2	0
Estimated Onboard Passenger	2	4	5	5	5	6	6	4	4	5	5	5	4	3	1

the proposed algorithm outperformed baseline models and the existing filtering methods.

The finding of this study can help to provide real-time accurate transit ridership flow and O–D information for supporting the transit network planning and improving quality of service. In addition, transit passengers could get a better understanding of the operational status of transit systems for optimizing their travel plan. In this study, the O–D information of a sample of passengers is achieved. The population O–D inference based on Wi-Fi and BT sensing data could be a future research direction.

REFERENCES

[1] L. Liu, L. Sun, Y. Chen, and X. Ma, "Optimizing fleet size and scheduling of feeder transit services considering the influence of bike-sharing systems," *J. Clean. Prod.*, vol. 236, Nov. 2019, Art. no. 117550.

[2] X. Hua, W. Wang, Y. Wang, and Z. Pu, "Optimizing phase compression for transit signal priority at isolated intersections," *Transport*, vol. 32, no. 4, pp. 386–397, 2017.

[3] V. R. Vuchic, *Urban Transit: Operations, Planning, and Economics*. Hoboken, NJ, USA: Wiley, 2017.

[4] V. R. Vuchic, *Urban Transit Systems and Technology*. Hoboken, NJ, USA: Wiley, 2007.

[5] W. Xu *et al.*, "Internet of Vehicles in big data era," *IEEE/CAA J. Autom. Sinca*, vol. 5, no. 1, pp. 19–35, Jan. 2018.

[6] M. Ben-Akiva, P. P. Macke, and P. S. Hsu, "Alternative methods to estimate route-level trip tables and expand on-board surveys," *Transp. Res. Rec. J. Transp. Res. Board*, vol. 1037, pp. 1–11, 1985.

[7] M. Wardman, "A comparison of revealed preference and stated preference models of travel behaviour," *J. Transp. Econ. Policy*, vol. 22, no. 1, pp. 71–91, 1988.

[8] X. Ma, Y.-J. Wu, Y. Wang, F. Chen, and J. Liu, "Mining smart card data for transit riders' travel patterns," *Transp. Res. C, Emerg. Technol.*, vol. 36, pp. 1–12, Nov. 2013.

[9] T. Kusakabe and Y. Asakura, "Behavioural data mining of transit smart card data: A data fusion approach," *Transp. Res. C, Emerg. Technol.*, vol. 46, pp. 179–191, Sep. 2014.

[10] X. Ma, Y. Wang, F. Chen, and J. Liu, "Transit smart card data mining for passenger origin information extraction," *J. Zhejiang Univ. Sci. C*, vol. 13, no. 10, pp. 750–760, 2012.

[11] L.-M. Kieu, A. Bhaskar, and E. Chung, "A modified density-based scanning algorithm with noise for spatial travel pattern analysis from smart card AFC data," *Transp. Res. C, Emerg. Technol.*, vol. 58, pp. 193–207, Sep. 2015.

[12] X. Ma, C. Liu, H. Wen, Y. Wang, and Y.-J. Wu, "Understanding commuting patterns using transit smart card data," *J. Transp. Geography*, vol. 58, pp. 135–145, Jan. 2017.

[13] C. Morency, M. Trépanier, and B. Agard, "Measuring transit use variability with smart-card data," *Transp. Policy*, vol. 14, no. 3, pp. 193–203, 2007.

[14] *Global Mobile Phone Internet User Penetration 2019 Statistic*. Accessed: Mar. 13, 2019. [Online]. Available: https://www.statista.com/statistics/284202/mobile-phone-internet-user-penetration-worldwide/

[15] *Smartphone Penetration in the U.S. (Share of Population) 2010–202 Statistic*. Accessed: Mar. 13, 2019. [Online]. Available: https://www.statista.com/statistics/201183/forecast-of-smartphone-penetration-in-the-us/

[16] Y. Malinovskiy, Y.-J. Wu, Y. Wang, and U. K. Lee, "Field experiments on bluetooth-based travel time data collection," in *Proc. 89th Annu. Meeting Transp. Res. Board*, 2010.

[17] M. Dunlap, Z. Li, K. Henrickson, and Y. Wang, "Estimation of origin and destination information from Bluetooth and Wi-Fi sensing for transit," *Transp. Res. Rec. J. Transp. Res. Board*, vol. 2595, no. 1, pp. 11–17, 2016.

[18] R. G. Mishalani, M. R. McCord, and T. Reinhold, "Use of mobile device wireless signals to determine transit route-level passenger origin–destination flows: Methodology and empirical evaluation," *Transp. Res. Rec. J. Transp. Res. Board*, vol. 2544, no. 1, pp. 123–130, 2016.

[19] A. Hidayat, S. Terabe, and H. Yaginuma, "WiFi scanner technologies for obtaining travel data about circulator bus passengers: Case study in Obuse, Nagano Prefecture, Japan," *Transp. Res. Rec.*, vol. 2672, no. 45, pp. 45–54 2018.

[20] T. Oransirikul, R. Nishide, I. Piumarta, and H. Takada, "Measuring bus passenger load by monitoring Wi-Fi transmissions from mobile devices," *Procedia Technol.*, vol. 18, pp. 120–125, Jan. 2014.

[21] Y. Ji, J. Zhao, Z. Zhang, and Y. Du, "Estimating bus loads and OD flows using location-stamped farebox and Wi-Fi signal data," *J. Adv. Transp.*, vol. 2017, May 2017, Art. no. 6374858.

[22] J. C. Bezdek, R. Ehrlich, and W. Full, "FCM: The fuzzy c-means clustering algorithm," *Comput. Geosci.*, vol. 10, nos. 2–3, pp. 191–203, 1984.

[23] S. Gan, S. Liang, K. Li, J. Deng, and T. Cheng, "Trajectory length prediction for intelligent traffic signaling: A data-driven approach," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 2, pp. 426–435, Feb. 2018.

[24] M. Gastaldi, G. Gecchele, and R. Rossi, "Estimation of annual average daily traffic from one-week traffic counts. A combined ANN-fuzzy approach," *Transp. Res. C, Emerg. Technol.*, vol. 47, pp. 86–99, Oct. 2014.

[25] D. C. Duives, W. Daamen, and S. P. Hoogendoorn, "How to measure static crowds? Monitoring the number of pedestrians at large open areas by means of Wi-Fi sensors," in *Proc. 97th Annu. Meeting Transp. Res. Board*, 2018, p. 6.

[26] V. Kostakos, T. Camacho, and C. Mantero, "Wireless detection of end-to-end passenger trips on public transport buses," in *Proc. 13th Int. IEEE Conf. Intell. Transp. Syst.*, Funchal, Portugal, 2010, pp. 1795–1800.

[27] A. Lesani and L. F. Miranda-Moreno, "Development and testing of a real-time WiFi-Bluetooth system for pedestrian network monitoring, Classification, and data extrapolation," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 4, pp. 1484–1496, Apr. 2019.

[28] V. Kostakos, T. Camacho, and C. Mantero, "Towards proximity-based passenger sensing on public transport buses," *Pers. Ubiquitous Comput.*, vol. 17, no. 8, pp. 1807–1816, 2013.

[29] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[30] *FLIR Acyclica RoadTrend*. Accessed: Mar. 24, 2019. [Online]. Available: https://www.flir.com/support/products/roadtrend#Specifications

[31] M. Cunche, "I know your MAC Address: Targeted tracking of individual using Wi-Fi," *J. Comput. Virol. Hacking Techn.*, vol. 10, no. 4, pp. 219–227, 2014.

[32] V. Tzivaras, *Raspberry Pi Zero W Wireless Projects*. Birmingham, U.K.: Packt Publ. Ltd, 2017.

[33] L. A. Zadeh, "Fuzzy sets," *Inf. Control*, vol. 8, no. 3, pp. 338–353, 1965.

[34] R. Xu and D. C. Wunsch, "Survey of clustering algorithms," *IEEE Trans. Neural Netw.*, vol. 16, no. 3, pp. 645–678, May 2005.

[35] C. Wang, W. Pedrycz, J. Yang, M. Zhou, and Z. Li, "Wavelet frame-based fuzzy C-means clustering for segmenting images on graphs," *IEEE Trans. Cybern.*, early access, Jul. 10, 2019, doi: 10.1109/TCYB.2019.2921779.

[36] J. Qin, W. Fu, H. Gao, and W. X. Zheng, "Distributed k-means algorithm and fuzzy *c*-means algorithm for sensor networks based on multiagent consensus theory," *IEEE Trans. Cybern.*, vol. 47, no. 3, pp. 772–783, Mar. 2017.

[37] M. J. Rezaee, M. Jozmaleki, and M. Valipour, "Integrating dynamic fuzzy C-means, data envelopment analysis and artificial neural network to online prediction performance of companies in stock exchange," *Phys. A, Stat. Mech. Appl.*, vol. 489, pp. 78–93, Jan. 2018.

[38] A. K. Dubey, U. Gupta, and S. Jain, "Comparative study of K-means and fuzzy C-means algorithms on the breast cancer data," *Int. J. Adv. Sci. Eng. Inf. Technol.*, vol. 8, no. 1, pp. 18–29, 2018.

[39] C. Xu, Z. Li, Z. Pu, Y. Guo, and P. Liu, "Procedure for determining the deployment locations of variable speed limit signs to reduce crash risks at freeway recurrent bottlenecks," *IEEE Access*, vol. 7, pp. 47856–47863, 2019.

[40] J. Wang and T. Kumbasar, "Parameter optimization of interval Type-2 fuzzy neural networks based on PSO and BBBC methods," *IEEE/CAA J. Autom. Sinca*, vol. 6, no. 1, pp. 247–257, Jan. 2019.

[41] S. Gao, M. Zhou, Y. Wang, J. Cheng, H. Yachi, and J. Wang, "Dendritic neuron model with effective learning algorithms for classification, approximation, and prediction," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 2, pp. 601–614, Feb. 2019.

[42] R. J. Hathaway and J. C. Bezdek, "Fuzzy *c*-means clustering of incomplete data," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 31, no. 5, pp. 735–744, Oct. 2001.

[43] W. Zhu, X. Liu, M. Xu, and H. Wu, "Predicting the results of RNA molecular specific hybridization using machine learning," *IEEE/CAA J. Autom. Sinca*, vol. 6, no. 6, pp. 1384–1396, Nov. 2019.

[44] Z. Pu, Z. Li, R. Ke, X. Hua, and Y. Wang, "Evaluating the nonlinear correlation between vertical curve features and crash frequency on highways using random forests," *J. Transp. Eng. A, Syst.*, vol. 146, no. 10, 2020, Art. no. 04020115.

[45] J. G. Brida, B. Lanzilotta, L. Moreno, and F. Santiñaque, "A non-linear approximation to the distribution of total expenditure distribution of cruise tourists in Uruguay," *Tour. Manag.*, vol. 69, pp. 62–68, Dec. 2018.

[46] L. Breiman, *Classification and Regression Trees*. Abingdon, U.K.: Routledge, 2017.

[47] D. Reynolds, "Gaussian mixture models," in *Encyclopedia of Biometrics*. Boston, MA, USA: Springer, 2015, pp. 827–832.

[48] S. J. Roberts, D. Husmeier, I. Rezek, and W. Penny, "Bayesian approaches to Gaussian mixture modeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1133–1142, Nov. 1998.

[49] H. Attias, "A variational Baysian framework for graphical models," in *Advances in Neural Information Processing Systems*. San Mateo, CA, USA: Kaufmann, 2000, pp. 209–215.

[50] D.-S. Lee, J. J. Hull, and B. Erol, "A Bayesian framework for Gaussian mixture background modeling," in *Proc. Int. Conf. Image Process. (Cat. No. 03CH37429)*, vol. 3. Barcelona, Spain, 2003, pp. 973–976.

[51] J. Wu, H. Xiong, and J. Chen, "Adapting the right measures for K-means clustering," in *Proc. 15th ACM SIGKDD Int. Conf. Knowl. Discover. Data Min.*, 2009, pp. 877–886.

[52] Y. Liu, Z. Li, H. Xiong, X. Gao, and J. Wu, "Understanding of internal clustering validation measures," in *Proc. IEEE Int. Conf. Data Min.*, Sydney, NSW, Australia, 2010, pp. 911–916.

[53] Y. Zhao and G. Karypis, "Evaluation of hierarchical clustering algorithms for document datasets," in *Proc. 11th Int. Conf. Inf. Knowl. Manag.*, 2002, pp. 515–524.

[54] P.-N. Tan, *Introduction to Data Mining*. Boston, MA, USA: Addison Wesley, 2018.

[55] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, pp. 53–65, Nov. 1987.

[56] J. C. Dunn, "Well-separated clusters and optimal fuzzy partitions," *J. Cybern.*, vol. 4, no. 1, pp. 95–104, 1974.

[57] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-1, no. 2, pp. 224–227, Apr. 1979.

[58] J. Han, *CS 412 Introduction to Data Mining*, The Grainger College Eng. Comput. Sci., Urbana, IL, USA, 2017.

[59] D. B. Paradeda, W. K. Junior, and R. C. Carlson, "Bus passenger counts using Wi-Fi signals: Some cautionary findings," *Transportes*, vol. 27, no. 3, pp. 115–130, 2019.

[60] L. Mikkelsen, R. Buchakchiev, T. Madsen, and H. P. Schwefel, "Public transport occupancy estimation using WLAN probing," in *Proc. 8th Int. Workshop Resilient Netw. Design Model. (RNDM)*, Halmstad, Sweden, 2016, pp. 302–308.

[61] T. Oransirikul, I. Piumarta, and H. Takada, "Classifying passenger and non-passenger signals in public transportation by analysing mobile device Wi-Fi activity," *J. Inf. Process.*, vol. 27, pp. 25–32, Jan. 2019.

[62] T. Oransirikul and H. Takada, "The practicability of predicting the number of bus passengers by monitoring Wi-Fi signal from mobile devices with the polynomial regression," in *Proc. Adjunct ACM Int. Joint Conf. Pervasive Ubiquitous Comput. Int. Symp. Wearable Comput.*, 2019, pp. 781–787.

[63] U. Mehmood, I. Moser, P. P. Jayaraman, and A. Banerjee, "Occupancy estimation using WiFi: A case study for counting passengers on busses," in *Proc. IEEE 5th World Forum Internet Things (WF-IoT)*, Limerick, Ireland, 2019, pp. 165–170.

Ziyuan Pu (Member, IEEE) received the B.S. degree in transportation engineering from Southeast University, Nanjing, China, in 2010, and the M.S. and Ph.D. degrees in civil and environmental engineering from the University of Washington, Seattle, WA, USA, in 2015 and 2020, respectively.

He is currently a Research Associate with the University of Washington. His research interests include transportation data science, smart transportation infrastructures, connected and autonomous vehicles, and urban computing.

Meixin Zhu received the B.Sc. and M.Sc. degrees in traffic engineering from Tongji University, Shanghai, China, in 2015 and 2018, respectively. He is currently pursuing the Ph.D. degree with the University of Washington., Seattle, WA, USA.

He also serves as a Research Assistant with the Smart Transportation Applications and Research Laboratory, University of Washington, advised by Prof. Y. Wang. His research interests include autonomous driving, artificial intelligence, big data analytics, driving behavior, traffic-flow modeling and simulation, and naturalistic driving study.

Wenxiang Li received the Ph.D. degree from the College of Transportation Engineering, Tongji University, Shanghai, China, in 2019.

He is currently an Assistant Professor with the Business School, University of Shanghai for Science and Technology, Shanghai. He was a visiting student with the Department of Civil and Environment Engineering, University of Washington, Seattle, WA, USA, from September 2017 to September 2018. His research focuses on sustainable transportation, which includes carbon emission trading of transport sector, low-carbon transport evaluation, and planning of electric vehicles and shared mobility.

Zhiyong Cui received the B.S. degree in software engineering from Beihang University, Beijing, China, in 2012, and the M.S. degree in software engineering and microelectronics from Peking University, Beijing, in 2015. He is currently pursuing the Ph.D. degree in civil and environmental engineering with the University of Washington, Seattle, WA, USA.

His research interests include deep learning, machine learning, traffic data mining, and intelligent transportation systems.

Xiaoyu Guo received the B.S. degree in mathematics from University at Buffalo, The State University of New York, Buffalo, NY, USA, in 2017, and the M.S. degree in civil engineering from the Zachry Department of Civil and Environmental Engineering, Texas A&M University, College Station, TX, USA, in 2019, where she is currently pursuing the Ph.D. degree in civil engineering.

Since 2018, she has been a Research Assistant for the Connected Infrastructure Group under the System Reliability Division, Texas A&M Transportation Institute, College Station. Her research interests include connected infrastructures and vehicles, transportation modeling and simulation, traffic operations, and traffic flow theory.

Yinhai Wang (Senior Member, IEEE) received the master's degree in computer science from the University of Washington (UW), Seattle, WA, USA, in 2002, and the Ph.D. degree in transportation engineering from the University of Tokyo, Tokyo, Japan, in 1998.

He is a Professor of transportation engineering and the Founding Director of the Smart Transportation Applications and Research Laboratory, UW. He also serves as the Director for Pacific Northwest Transportation Consortium (PacTrans), USDOT University Transportation Center for Federal Region 10. He has published over 100 peer reviewed journal articles and delivered more than 110 invited talks and nearly 200 other academic presentations. His active research fields include traffic sensing, e-science of transportation, and transportation safety.

Dr. Wang was the winner of the ASCE Journal of Transportation Engineering Best Paper Award for 2003. He serves as a member of the Transportation Information Systems and Technology Committee and Highway Capacity and Quality of Service Committee of the Transportation Research Board. He is currently a member of the steering committee for the IEEE Smart Cities and chaired the First IEEE International Smart Cities Conference in 2015. He was an elected member of the Board of Governors for the IEEE ITS Society from 2010 to 2013 and served on the Board of Governors for the ASCE Transportation & Development Institute 2013–2015. He is an Associate Editor of the *Journal of Intelligent Transportation Systems*, the *Journal of Computing in Civil Engineering*, and the *Journal of Transportation Engineering*.