Developing Wireless Sensing Methods and Technologies for Enhanced Transit Rider and Non-Motorized Traffic Data

Ziyuan Pu

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2020

Reading Committee:

Yinhai Wang, Chair

Xuegang (Jeff) Ban

Ed McCormack

Program Authorized to Offer Degree:

Civil & Environmental Engineering

University of Washington

Abstract

Developing Wireless Sensing Methods and Technologies for Enhanced Transit Rider and Non-Motorized Traffic Data

Ziyuan Pu

Chair of the Supervisory Committee:
Yinhai Wang, Full Professor
Department of Civil and Environmental Engineering

Real-time traffic data is essential for the advancement of emerging data-driven transportation technologies, e.g. deep learning-based traffic modeling approaches, autonomous vehicles, and urban computing. The existing sensing technologies work properly well for identifying the mobility patterns of motorized vehicles. However, it is still a big challenge for transportation agencies to obtain reliable data of transit riders and non-motorized travelers in today's practice with existing traffic sensing technologies. To fulfill the data needs of understanding and modeling the mobility of transit riders and non-motorized traffic (bicycling, and walking), device-based wireless sensing methods and technologies have been developed to acquire relevant data. The basic idea of device-based wireless sensing technology is to capture the Media Access Control (MAC) address of Wi-Fi or Bluetooth enabled mobile devices. The MAC address can be used as a global

unique identifier to re-identify mobile devices at different sensing locations, and thus travelers can be detected by identifying their mobile devices instead of detecting travelers directly. Such a data acquisition method certainly provides a novel means for transit riders and non-motorized traffic data collection. Nevertheless, the limitations still exist for wireless sensing technologies due to the uncertainties caused by the sensing mechanism, including traffic mode uncertainty, localized spatial uncertainty, and population uncertainty. Such uncertainties bring considerable errors that can generate significant biases in the extracted traffic parameters from wireless sensing data.

The major objective of this dissertation is to mitigate the impacts of the uncertainties based on the proposed wireless sensing methods and technologies for transit rider and non-motorized traffic data acquisition. large-scale field tests are conducted to evaluate the efficiency and accuracy of the proposed methodology. Besides the proposed methodology, a general method for establishing a wireless sensing system is presented for guiding implementations. This dissertation fills up the gap about effective traffic data acquisition methods for transit rider and non-motorized traffic and thus supporting the transportation systems with reliability, equality, and sustainability.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

It is a long journey to complete this dissertation and the subsequent Ph.D. degree, probably the most challenging activity of my first 30 years of my life. The best and worst moments of my doctoral journey have been shared with many people. It has been a great privilege to spend six years in the Department of Civil and Environmental Engineering at the University of Washington, and its members will always remain dear to me.

Firstly, I want to express my gratitude to my adviser, Prof. Yinhai Wang. He patiently provided me the vision, encouragement, and advice to struggle with the graduate program and complete my dissertation. I want to thank him not only for being a distinguished and supportive adviser to me throughout my graduate school career but also for being a role model as a great human being.

Special thanks to my supervisory committee, Prof. Jeff (Xuegang) Ban, Prof. Edward D. McCormack, and Prof. Anne Vernez Moudon for their support, guidance, and helpful suggestions. Their guidance has served me well and I owe them my heartfelt appreciation.

Members of the STAR Lab at the University of Washington also deserve my sincerest thanks, their friendship and assistance have meant more to me than I could ever express. I could not complete my work without invaluable friendly assistance from Mr. Zhiyong Cui, Mr. Ruimin Ke, Mr. John Ash, Mr. Yifan Zhuang, Mr. Meixin Zhu, Mr. Frank (Hao) Yang, and Mrs. Summer (Meng) Xia. I also should mention Prof. Xiucheng Guo and Prof. Zhibin Li from Southeast University, Mr. Chenglong Liu and Mr. Duo Zhang from Tongji University, Mr. Tianming Liu from the University of Michigan, Ms. Qiannan Zhang from Shanghai Maritime University, and Ms. Shuo Wang from Nanjing University of Science and Technology for their valuable suggestions and crucial help with the research.

My friends in the U.S., China, and other parts of the world were sources of laughter, joy, and support. Special thanks go to Zehao Qin, Mian Sun, Ziheng Wu, Xiao Wang, Xiao Han, and Ming Yang. I am very happy that, in many cases, my friendships with you have extended well beyond our shared time. I also wish to thank my entire family for providing a loving environment for me.

Lastly, and most importantly, I wish to thank my parents, Ye Chi and Anjian Pu. They bore me, raised me, supported me, taught me, and love me. To them, I dedicate this dissertation.

DEDICATION

To my parents.

Chapter 1. INTRODUCTION

1.1 PROBLEM STATEMENT

This research is mainly motivated by two issues. The first is the increasing needs of transit rider and non-motorized (bicycling and walking) traffic data. As the interest in providing better infrastructure support for non-motorized and public transit users increases, relevant data is getting more attraction to improve the understanding of infrastructure usage patterns based on data-driven approaches. However, it still lacks an effective means to collect traffic data of transit riders and non-motorized travelers, which turns into the second issue, that being the limitations of the existing sensing methods and technologies on transit rider and non-motorized data acquisition. In this chapter, each of the issues presented is described in greater detail in the following subsections.

1.1.1 *Traffic Data Needs of Transit Riders and Non-motorized Travelers*

Due to the rapid increase in use of data-driven traffic modeling approaches, such as. deep learning-based algorithms for traffic prediction [1], [2], mobility pattern recognition [3]–[5], and traffic signal timing optimization [6], the entire transportation system has experienced numerous improvements during the recent decades in terms of safety, mobility, sustainability, and efficiency [7]. In general, the goal of data-driven approaches is to mine the data and discover any underlying patterns in the datasets [8]. A traffic dataset with high volume, variation, veracity, and intensive information is essential for developing data-driven approaches. To obtain representative real-time traffic data across a given geographic area, thousands of traffic sensors are deployed on roadway networks and in vehicles, e.g., surveillance cameras, loop detectors, microwave traffic sensors, and in-vehicle sensing systems. Based on the existing traffic sensor networks, network-wide traffic parameters, e.g., travel time, traffic volume, traffic speed, and parking activities, can be extracted

to support traffic status inference, traffic control applications, and traffic management systems. As the desire to pursue green and sustainable transportation is growing, interest in providing better infrastructures support for non-motorized (bicycling and walking) and public transit users is a topic generating more attention. Volumes and other basic traffic information on transit riders and non-motorized travelers are indispensable in supporting and understanding infrastructure usage and traffic demand. For example, transit ridership flow and Origin-Destination (O-D) information are essential for transit route planning, transit vehicle dispatch optimization, and trip scheduling [9]–[12]. Traffic counts and travel time data for non-motorized traffic are critical for travel information systems, non-motorized traffic signal control, crowd management strategies, and for use in emergency and evacuation scenarios. However, it is still a big challenge for transportation agencies to obtain reliable data of transit riders and non-motorized travelers in today's practice with existing traffic sensing technologies.

1.1.2 *Limitations of The Existing Methods for Transit Rider and Non-motorized Traffic Data Acquisition*

Currently, collecting traffic data describing transit riders and non-motorized travelers is a task that still primarily relies on traditional methods and off-the market sensing technologies. Traveler surveys [13] and smart card data are the major data sources for inferring transit ridership flow and O-D information [14]–[19]. However, inferring transit ridership flow based on survey data requires substantial manual work and as a result suffers from high latency, biases, and costliness. Even for smart card data, it is hard to infer real-time ridership flow due to the lack of alighting locations of transit passengers. For non-motorized traffic data collection, analysts mainly rely on traditional data acquisition methods, such as surveys, manual counting, infrared sensing, and eco-counter [20]. To increase the availability of data describing non-motorized traffic, video-based sensing

technologies are employed to extract real-time traffic information of non-motorized traffic modes, e.g., traffic speed, and travel time [21], [22]. However, the accuracy of the extracted information cannot satisfy the needs for understanding and identifying traffic status due to the impacts from environmental factors, such as illumination conditions and weather [23], [24]. In addition, the excessive computational cost and privacy issues restrict the large-scale implementation of video-based sensing technology. Therefore, there is a clear need of novel data sources and acquisitions technologies to support the understanding of and enable improvements to be made to public transit systems and non-motorized traffic mobility [25]–[27].

1.2 RESEARCH BACKGROUND

With the ubiquitous usage of wireless devices increases, scholars developed traffic sensing methods based on wireless sensing technologies. The basic idea of wireless sensing technology is to detect the signals being transmitted by the wireless device in vehicles or being carried by travelers, e.g., Wi-Fi or Bluetooth-enabled mobile device (WBM device), and on-vehicle wireless devices. Typically, wireless signals contain global unique identifiers which can be used to precisely re-identify a specific wireless device. Then, traffic parameters can be obtained by identifying the movements of wireless device instead of tracking travelers and vehicles directly. For acquiring traffic data on transit riders and non-motorized travelers, the existing wireless sensing methods usually utilized the probe request frames or slave messages being transmitted by the WBM device of transit riders and non-motorized travelers. However, the extracted traffic parameters based on the existing wireless sensing methods still contain considerable biases caused by three uncertainties, including traffic modes uncertainty, localized spatial uncertainty, and population uncertainty.

Localized spatial uncertainty: A WBM device can be detected at any location within the detection range of a wireless sensor. The exact detection location cannot be determined based on the wireless sensing data. Usually, the wireless sensor's location is used for traffic parameters calculation, e.g., traffic speed. The difference between detection location and sensor's location cause considerable errors within the extracted traffic parameters

Traffic modes uncertainty: Since the detected wireless signals can be transmitted by the wireless devices of travelers in any traffic modes, e.g., pedestrian and bicyclist, and transit passengers and non-passengers, the traffic modes cannot be directly observed in wireless sensing data. Thus, the extracted traffic parameters can be extremely biased if traffic modes are not correctly identified and separated properly.

Population uncertainty: Since not every traveler has a wireless device in the discoverable mode, it is difficult to ensure that the detected sample came from enough representatives of the population to reflect the global trend. Usually, the population uncertainty varies from location to location and, also, changes in the temporal dimension. Thus, developing a method to estimate the population based on the detected samples, or to utilizing device-free wireless sensing technologies is necessary in order to mitigate the impacts of penetration rate of discoverable wireless devices.

1.3 RESEARCH OBJECTIVES

The major objective of this dissertation is to develop innovative wireless sensing methods and technologies to enhance transit rider and non-motorized traffic data by addressing the three aforementioned uncertainties. The specific objectives corresponding to addressing each uncertainty are presented as follows.

- Localized spatial uncertainty

1) By utilizing wireless signal strength measurements, an innovative method is established to reduce the estimated traffic speed errors caused by localized spatial uncertainty.

- Traffic modes uncertainty
  2) For transit rider data, a new algorithm is developed to separate wireless signals transmitted by the wireless devices of transit passengers and non-passengers.
  3) For non-motorized traffic data, an approach is designed to identify wireless sensing data from different traffic modes, including car, bike, and walk modes.
- Population uncertainty
  4) For transit rider data, a method is constructed to estimate the population of transit passengers.
  5) For non-motorized traffic data, a device-free wireless sensing technology is developed for detecting pedestrian presence and moving features that also mitigates the impacts of population uncertainty.

In addition, this research proposes a general framework for establishing a wireless sensing system. This methodology includes wireless sensing equipment, the supporting system architecture, and a data analysis and visualization platform. To demonstrate the feasibility and the effectiveness of the proposed wireless sensing theory and framework, two case studies are implemented for transit ridership flow monitoring and non-motorized traffic monitoring, respectively.

1.4 DISSERTATION ORGANIZATION

The remainder of the dissertation contains six chapters. Figure 1-1 shows the framework of all chapters, Chapter 1-7. The current chapter, Chapter 1, introduces the problem statement and background knowledge. Chapter 2 presents the literature review results in terms of state of the art

and practice. Chapter 3 proposes a generalized methodology for wireless sensing. The proposed methodology includes two parts: the theoretical methodology for addressing the introduced uncertainties and the common method for implementations. Chapters 4 and 5 propose specific methods and technologies for transit ridership flow monitoring and non-motorized traffic monitoring based on the proposed generalized methodology. In addition, these two chapters present the prototype products developed based on the proposed common framework for implementations. Chapter 6 discusses the impacts of MAC address randomization as well as the potential implementation scenarios of the proposed wireless sensing methods. Finally, Chapter 7 concludes the dissertation with a summary of the contributions and recommendations for future research.

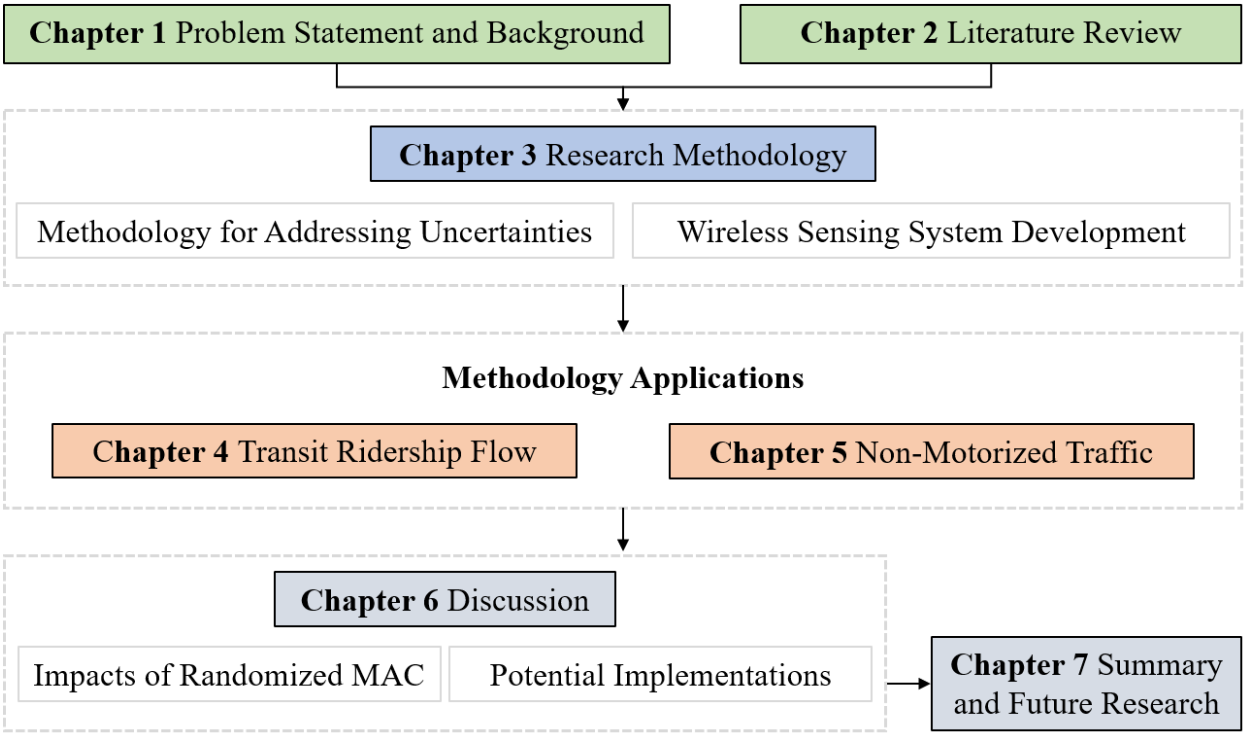


Figure 1-1 Dissertation Framework

Chapter 2. STATE OF THE ART AND PRACTICE

The major goal of this research is to enhance transit rider and non-motorized traffic data through new sensing methods and technologies. A thorough understanding of the state-of-the-art and state-of-the-practice related to transit rider and non-motorized traffic data collection is essential. Thus, this chapter presents the literature review results on the topics of traffic sensing theory, wireless sensing theory, and the existing implementations.

2.1 TRAFFIC SENSING THEORY

In general, traffic parameters, e.g., traffic speed, can be detected by two sensing methods, those being point detection and re-identification. For point detection, traffic sensors utilize the physical features (e.g., visual features) of objects (vehicles and travelers) to detect their presence. Localized traffic parameters, such as speed, volume, and occupancy, can be extracted based on point detection results. The detected physical features are similar from object to object, and thus such methods generally cannot support re-identify objects at different sensing locations. Representative point detection technologies include inductive loop detectors [28], microwave sensing, video-based sensing device [29]–[31], 3D laser-based sensing [32], radar-based detection [30], [33], and infrared sensors [34]. For re-identification, the main goal is more than to simply detect the presence of objects. Rather, objects need to be re-identified at different sensing locations based on either extracted visual features or global unique identifiers, e.g., license plates of vehicles, Radio-Frequency Identification (RFID), and Media Access Control (MAC) addresses. Then, traffic parameters can be inferred based on the timestamp when objects were detected at different sensing locations. Basically, the result of re-identification method conveys more valuable information than that of point detection. For example, travel time is a type of traffic parameter that can convey the

traffic status of a road segment. Re-identification method can detect travel time, while it cannot be achieved by point detection method. Among the existing sensing technologies, some are enabled with re-identification functions for identifying vehicles or individual travelers, such as video-based sensing technology utilizing visual features and license plate information [35]–[38], phone app-based sensing [39], [40], RFID-based sensing [41], and smart card-based methods [42]. For transit riders and non-motorized travelers (bicyclists and pedestrians), only a few technologies are available for re-identification, including video-based sensing technology, phone app-based sensing, and smart card-based methods. However, all of these methods face challenges and limitations in terms of sensing accuracy and implementation feasibility. Video-based sensing technology is highly sensitive to environmental factors, e.g., illumination and weather conditions. The sensing accuracy will heavily drop under unsatisfactory environmental conditions. Considering the visual features of transit riders and non-motorized travelers are not evident among individuals, the accuracy of video-based sensing technology would be even worse for re-identifying transit riders and non-motorized travelers. In addition, the privacy concerns and high computational cost limit the feasibility of large-scale implementations [43]. In terms of other methods, phone app-based sensing methods suffer from low penetration rate and data availability issues, and smart card-based methods for transit ridership monitoring also have penetration concerns and issues with the lack of alighting information. Therefore, a novel traffic sensing technology with a re-identification function that can accommodate the disadvantages of the existing methods is in-demand to provide robust, reliable, and efficient sensing results.

2.2 Wireless Sensing Theory

As the usage of wireless devices (e.g., smart phones, and on-vehicle communication devices) increases rapidly, wireless signals being transmitting by a huge amount of wireless devices is an

excellent data resource for traffic sensing. Basically, there are two kinds of wireless sensing methods, those being device-based wireless sensing and device-free wireless sensing. The goal of device-based wireless sensing is to capture the hardware unique identifier of the wireless device being carried by travelers. Then, the hardware unique identifier can be used for re-identification. Device-free wireless sensing utilizes the perturbation in wireless communication caused by moving travelers to detect their presence and moving features. The following sections will introduce the existing studies on such topics and their pros and cons.

2.2.1 *Device-Based Wireless Sensing*

2.2.1.1 Sensing Technology

Device-based wireless sensing requires that a discoverable wireless device be carried by travelers and vehicles so that they can be detected by tracking the discoverable wireless device based on the hardware unique identifier. According to the literature, several wireless signals were utilized for device-based wireless traffic sensing, including RFID [44], Dedicated Short-Range Communication (DSRC) [45], Wi-Fi, and Bluetooth (BT) [46]. Among them, RFID and DSRC were usually used for motorized vehicle detection, a process which required vehicles to install a tag containing an RFID or DSRC chipset. Then, the signal being transmitted by the chipset can be captured by wireless sensors. However, for transit riders and non-motorized travelers, it is difficult to force each traveler to carry a chipset to be detected. WBM devices are common electronic devices that many people use in their daily lives. Nowadays, it is reported that more than 80% of individuals carried at least one WBM device in their daily life [47], [48]. For each WBM device, a Media Access Control (MAC) address is assigned as the hardware unique identifier. A Wi-Fi and BT MAC address is a 48-bit global unique identifier assigned by The Institute of Electrical and Electronics Engineers (IEEE) for use as a network address in communications with a network

segment [49]. As seen from Figure 2-1, the MAC address of WBM devices can be captured by either connecting the device with an access point or deploying a MAC address detector to passively capture the Wi-Fi probe request frames or BT slave response messages in the air (also known as passive Wi-Fi and BT sensing). Typically, building connections between the device and an access point requires actions being conducted from uses. This requirement makes the impacts of population uncertainty even worse. For the passive sensing method, when WBM devices do not connect with access points, they keep sending out probe request frames or slave response messages to find potential connections. If the access points are set in monitor mode, they can capture the frames and messages in the air so that they capture the MAC address in the header of the frames. Previously, scholars utilized passive Wi-Fi and BT sensing technology to establish methods for transit ridership flow monitoring and multi-modal traffic speed estimation. The following two sections will introduce the existing studies and potential improvements.

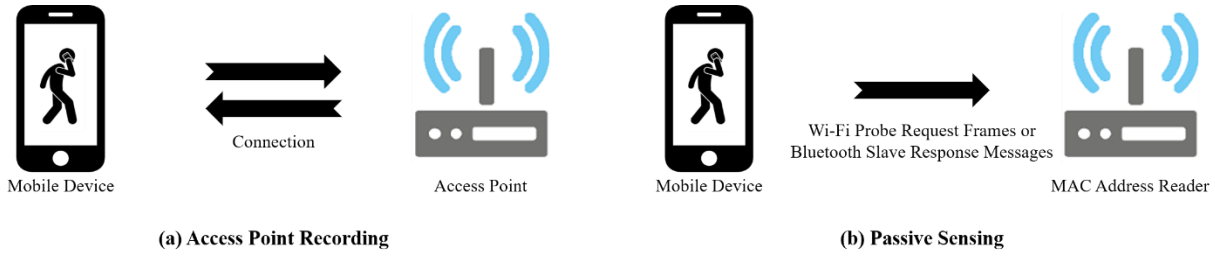


Figure 2-1 Device-Based Wi-Fi and BT Sensing. (a) Access Point Recording, and (b) Passive Sensing

2.2.1.2 Transit Ridership Flow Motoring

For monitoring transit ridership flow based on passive Wi-Fi and BT sensing, two major uncertainties need to be solved, those being traffic mode uncertainty and population uncertainty. For traffic mode uncertainty, since the detection range of a MAC address detector is usually larger than the inside space of transit vehicles, it is possible to detect WBM devices outside of transit vehicles. Thus, separating passengers' MAC addresses and non-passengers' MAC addresses is

crucial. Previously, several studies shed light on solving this problem based on filtering methods [50], [51]. Basically, several empirically predefined thresholds were used to filter out the data potentially coming from the outside of transit vehicles, e.g., detection duration. However, the results of the majority of those studies are barely convincing due to the lack of validation. Since collecting ground truth O-D information is costly and labor intensive, only a few studies provided the comparison of observed ridership flow and the estimated ridership flow [52]–[54]. The obvious gaps between the observed data and the filtering results indicated that the considerable errors are caused by the use of hard-threshold filtering methods. Hence, an accurate and effective method for separating the MAC address data belonging to passengers and non-passengers is in-demand. For population uncertainty, only partial transit passengers carry WBM devices, a method targeting estimation of the population ridership flow based on the number of separated passengers is needed. Previously, several methods were implemented to estimate the population, including scaling with a fixed number [55], linear regression [56], and power function and Fourier function based methods [57]. Among the existing methods, Lasani and Miranda-Moreno (2016) [57] conducted a performance comparison of power function and Fourier function based methods for estimating the population of pedestrians based on the detected Wi-Fi and BT MAC addresses. The proposed power function achieved a relatively higher R-squared value than the Fourier function. In addition, the R-squared value of the proposed power function is also much higher than that from linear regression method in other studies [58], which could be an indicator of the non-linear relationship between the population and the number of detected MAC addresses.

2.2.1.3 Multi-Modal Traffic Speed Estimation

The previous studies demonstrated that traffic speed monitoring based on passive Wi-Fi and BT sensing is cost-effective and relatively accurate compared with other means [59]. Numerous

studies shed light on investigating sources of bias and improving the accuracy of passive Wi-Fi and BT sensing-based travel time estimation from multiple perspectives. Jason et al. proposed a method for real-time travel time estimation using Media Access Control address matching in 2008 [46]. Their study was the first case in which passive Wi-Fi and BT sensing technology was used in travel time acquisition. The existing studies demonstrated the reliability and modeled the error sources of the travel time estimation based on passive Wi-Fi and BT sensing data [59]–[69]. The accuracy of such methods can be impacted by the communication of BT devices with sensors, the size of the detection range of the MAC address detector, the average traffic speed on a road segment, vertical sensor placement, and the type of antennas used for collecting data. To improve the accuracy of traffic speed estimation based on passive Wi-Fi and BT sensing data, other data sources, e.g., loop detector data, GPS data, were fused with passive Wi-Fi and BT sensing data [70]–[72]. Hyoshin and Ali proposed a two-stage stochastic model for determining the optimal number and location of MAC address detectors [73]. A traffic mode identification model was developed [74], and the Received Signal Strength Indicator (RSSI) of Wi-Fi and BT communication was used for developing an improved travel time estimation method [75], [76]. Based on the understanding of this technology, several implementations were developed for monitoring traffic in different scenarios, including travel time prediction [77], [78], arterial traffic congestion analysis [79], bicycle travel time estimation [80], travel time delay monitoring in work zones [81], roadway system assessment [82], freeway travel time monitoring system development [83], pedestrian network monitoring [84], and crowd mobility pattern exploration [85]. Most of them demonstrated the feasibility and effectiveness for monitoring motorized traffic on freeways, and only a few of them conducted research on multi-modal traffic monitoring in road networks within urban areas. In case of solely monitor motorized traffic on freeways, the challenges may

not be as hard as those posed in monitoring traffic for urban areas which have more complicated roadway networks and traffic composition. There are still several issues that potentially limit the implementation of such technology to monitor multi-modal traffic speed in urban areas including the localized spatial uncertainty, and the traffic mode uncertainty. For the localized spatial uncertainty, one previous studies explored methods to mitigate errors by evaluating the errors of the estimated traffic speed calculated by the first or the last detection [76]. Another study explored the detection point closest to the sensors based on RSSI values [86]. However, those simple methods cannot effectively correct the biases of the estimated traffic speed caused by the detection range of the sensors. For the traffic mode uncertainty, several studies have established algorithms to identify traffic modes based on supervised machine learning algorithms [74], filtering methods based on pre-defined thresholds [87], and logit models [84]. However, the accuracy of the existing methods highly relies on the empirical information extracted from a large set of labeled data or pre-defined thresholds. Considering that labeled data is often hard to obtain, and the optimal values of pre-defined thresholds are difficult to determine, the existing methods may still need to be improved to remove their dependency on large amounts of labeled data or pre-defined thresholds.

2.2.2 *Device-Free Wireless Sensing*

As mentioned earlier, passive Wi-Fi and BT sensing technology suffers from population uncertainty. For some implementation scenarios, e.g., advanced pedestrian crossing light systems and advanced driver-assistance systems [88], [89], precise pedestrian detection is required. False pedestrian detection could potentially result in pedestrian-related traffic crashes. Thus, a device-free wireless sensing technology is in-demand to alleviate the impacts of population uncertainty. In 802.11b/g/n standards, Wi-Fi Channel State Information (CSI) is a fine-grained channel response which describes the amplitude and phase information for Orthogonal frequency-division

multiplexing (OFDM) subcarriers [90]. Wi-Fi CSI is mainly impacted by the static environment (e.g., buildings in the outdoor environment, or furniture in the indoor environment) and moving objects (e.g., pedestrians or bicyclists) within the detection range (see Figure 2-1). Previously, scholars utilized Wi-Fi CSI to monitor microscopic human movements in indoor environments, including gait recognition, fall detection, gesture recognition, etc. [91]–[94]. The existing studies demonstrate Wi-Fi CSI is capable of sensing detailed information on human movements in the indoor environment. However, for implementations in the transportation domain, pedestrian detection is usually conducted in an outdoor environment. Thus, it is valuable to demonstrate whether the Wi-Fi CSI is feasible and reliable means for pedestrian detection in outdoor environments.

2.3 State of the Practice

The state-of-the-art describes the theory of how existing methods based on passive Wi-Fi and BT sensing technology can help with traffic sensing, especially for transit rider and non-motorized traffic monitoring. That said, for implementing the state-of-the-art, use of sensing equipment and systems beyond just equations and algorithms are essential. In general, only a few studies are available on guiding the implementation. The following paragraphs introduce the limited sources in terms of MAC address detector, and wireless sensing system development.

For MAC address detector development, some existing products are available on the market, e.g. Acyclica RoadTrend [95]. These devices are mainly used as a fixed MAC address detector for traffic analysis at intersections. However, MAC address detector also can be used as mobile sensing devices in applications, such as sensing transit riders on transit vehicles. The necessary hardware and software of the MAC address detector are quite different depending on whether as fixed sensing or mobile sensing paradigm is used, e.g., GPS recording, data communication, and

power supply. Thus, customized MAC address detector is critical to extend the implementation scenarios. In addition, the stability of the MAC address detector is significant for large-scale implementations. To maintain the stable operation of the equipment, advanced software frameworks and other minor hardware components, e.g., a real-time clock and watch dog module, are critical to support cost-effective and sustainable sensing system. However, no existing literature comments on the guidance of software framework design and hardware integration. To facilitate the implementations, a comprehensive guide for development of MAC address detector is also important. For wireless sensing system design, limited studies proposed system architecture design for specific implementations, including pedestrian monitoring [84], and freeway travel time monitoring [83]. Basically, these studies proposed a conceptual system architecture with the description of essential system components and data streaming among them. However, the proposed system lacks a means of data management and visualization, and these components are an important part to help make analysis results beneficial for road users and traffic manages.

According to the preceding considerations, besides the proposed methods for addressing uncertainties, this research will propose a general method for establishing a wireless sensing system. This method will include a sensing system design, sensing equipment development, and tools for data management and visualization. The proposed method will be introduced in Chapter 3.

Chapter 3. RESEARCH METHODOLOGY

3.1 METHODOLOGY FOR ADDRESSING UNCERTAINTIES

3.1.1 *Localized Spatial Uncertainty*

Typically, the detection range of MAC address detectors is about 50 - 80 meters for Wi-Fi and 10 - 20 meters for BT. The MAC address of a WBM device may be detected anywhere within the detection range, however, the exact detection location cannot be observed in the raw data. Such localized spatial uncertainty causes errors within the extracted traffic parameters, e.g. traffic speed. As seen in Figure 3-1, the two red circles represent the location of two MAC address detectors marked as s and e, and their detection range is represented by the blue circles. While the pedestrian is walking from the sensing location s to the sensing location e, the MAC address of the WBM device carried by the pedestrian is detected at P_1 and P_2. The traffic speed of the pedestrian moving from s to e can then be estimated with Equation (3-1),

$$Speed_{original} = \frac{d_{device}}{TT_{P_1-P_2}} \qquad (3-1)$$

where d_{device} is the distance between sensing points s and e, and $TT_{P_1-P_2}$ is the travel time between P_1 to P_2. Obviously, the estimated traffic speed using Equation (3-1) is biased and the error can be calculated with Equation (3-2),

$$Speed_{error} = \frac{\left|d_{device} - d_{P_1-P_2}\right|}{TT_{P_1-P_2}} \qquad (3-2)$$

where $d_{P_1-P_2}$ is the distance between P_1 and P_2. The magnitude of the error depends on the size of the detection range. Sometimes the detection range can be enlarged to one thousand meters by

adding a powerful antenna to the MAC address detector in order to capture the MAC addresses of WBM devices in vehicles driving with high speed [66].

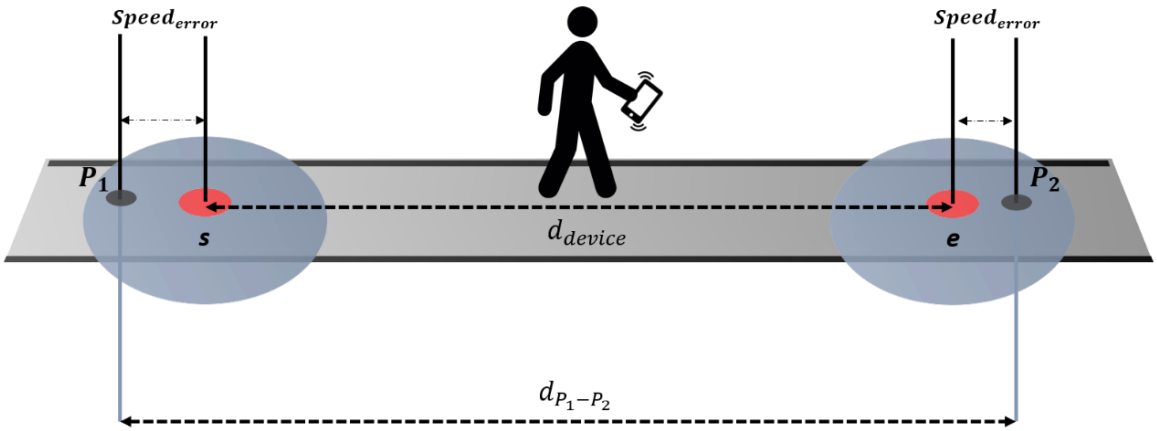


Figure 3-1 Estimated Traffic Speed Error Caused by Detection Range

For each detected Wi-Fi and BT probe request frame, there is an integer ranging from -120 to -30 that tells the signal strength of a Wi-Fi or BT signal. This integer is called the Received Signal Strength Indicator (RSSI). In previous studies, the distance between MAC address detectors and WBM devices has been demonstrated as one of the major influential factors of RSSI [96]. Functions mapping RSSI to distance were developed and utilized for indoor localization at the centimeter-level [97], [98].

In this study, the RSSI is utilized to mitigate the impacts of localized spatial uncertainty in traffic speed estimation. To explore the relationship between RSSI and distance, experiments are conducted to collect RSSI measurements under varying distances from the MAC address detector. Figure 3-2 shows the boxplot of RSSI measurements at different distances for Wi-Fi and BT signals, respectively. According to the figures, it is obvious that the values of RSSI increase as the WBM device gets closer to the MAC address detector. The correlations are fitted with exponential functions which are shown as red dashed lines. The R-squared value is higher than 0.95 for both

Wi-Fi and BT data, and this indicates an outstanding goodness of fit of the fitted functions. Then, the distance can be estimated by Equations (3-3) and (3-4) based on RSSI measurements,

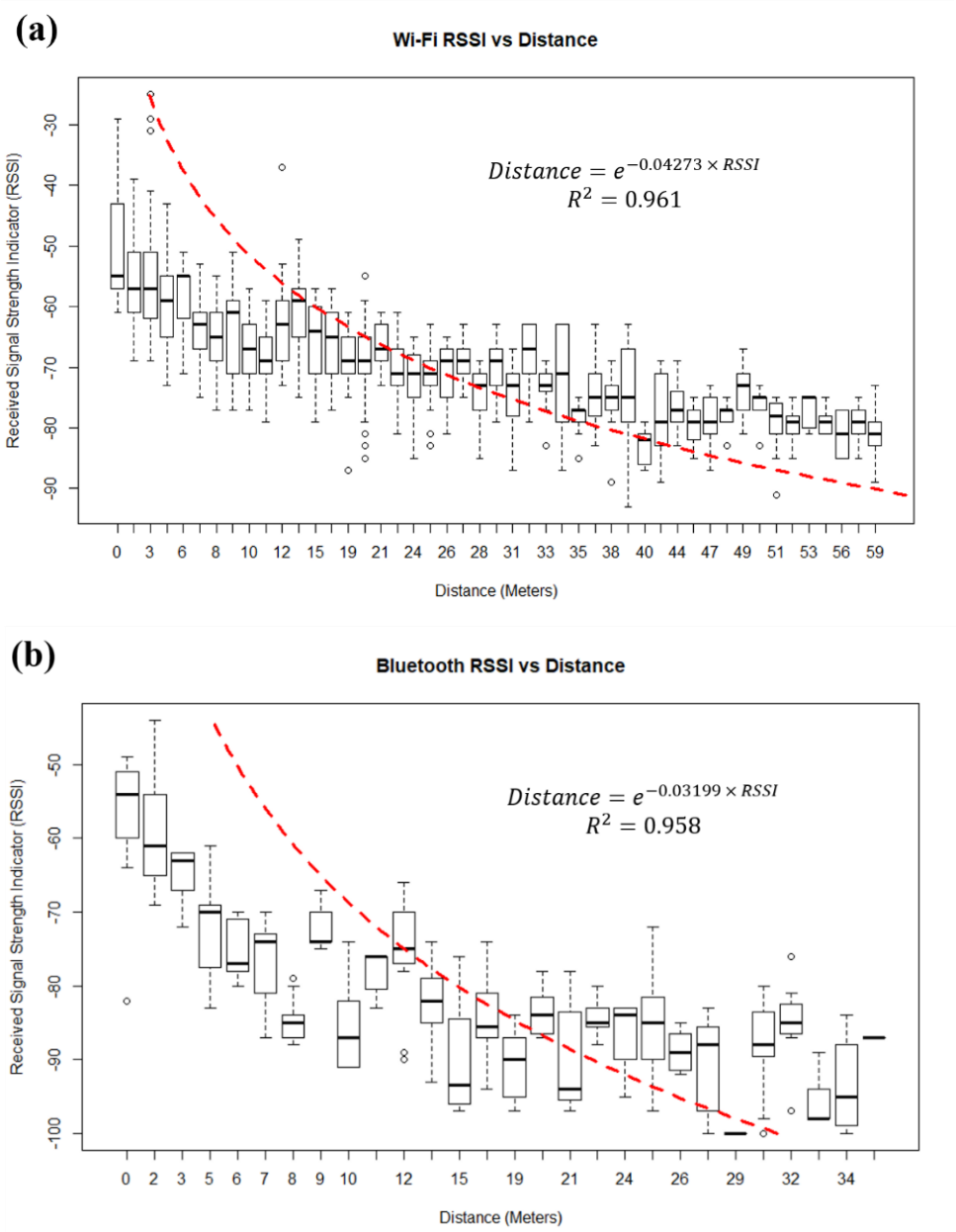


Figure 3-2 RSSI vs Distance with Regression Line. (a) Wi-Fi, and (b) BT

$$d_{Wi-Fi} = e^{-0.04273 \times RSSI} \quad (3-3)$$

$$d_{Bluetooth} = e^{-0.03199 \times RSSI} \quad (3-4)$$

where d_{Wi-Fi} and $d_{Bluetooth}$ are the distance between a WBM device and the MAC address detector for Wi-Fi and BT, respectively. To correct the errors based on RSSI measurements, the relative position of the detected WBM device and the MAC address detector is another required piece of information. Figure 3-3 illustrates the relative position when a WBM device is detected once, twice, and n times within the detection range of a MAC address detector.

In Figure 3-3, Scenario 1-1 and 1-2 present the case when a WBM device is detected once while within the detection range, where P_1 is the detection location and d_{P_1} is the distance between the MAC address detector and P_1. In this case, P_1 is either detected before or after the traveler passes the MAC address detector, and the relative position cannot be determined. When a WBM device has two detection points for a single trip, there are four scenarios which are presented in Scenario 2-1 through Scenario 2-4. If P_1 and P_2 are detected in chronological order with $RSSI_{P_1} < RSSI_{P_2}$, the relative position of P_1 to the MAC address detector can be determined such that P_1 is always detected before the WBM device passes the MAC address detector and vice versa. In general, if a WBM device is detected n times, we will always have a detected point with the highest RSSI measurement which is considered as the closet detection point to the MAC address detector. Then, the relative position of the rest of detection points to the MAC address detector can be determined by the following rules:

1) The detection points that are detected before the detection point with the highest RSSI measurement are determined to be the detection points being detected before the WBM device passes the MAC address detector.
2) The detection points that are detected after the detection point with the highest RSSI measurement can be determined to be the detection points being detected after the WBM device passes the MAC address detector.

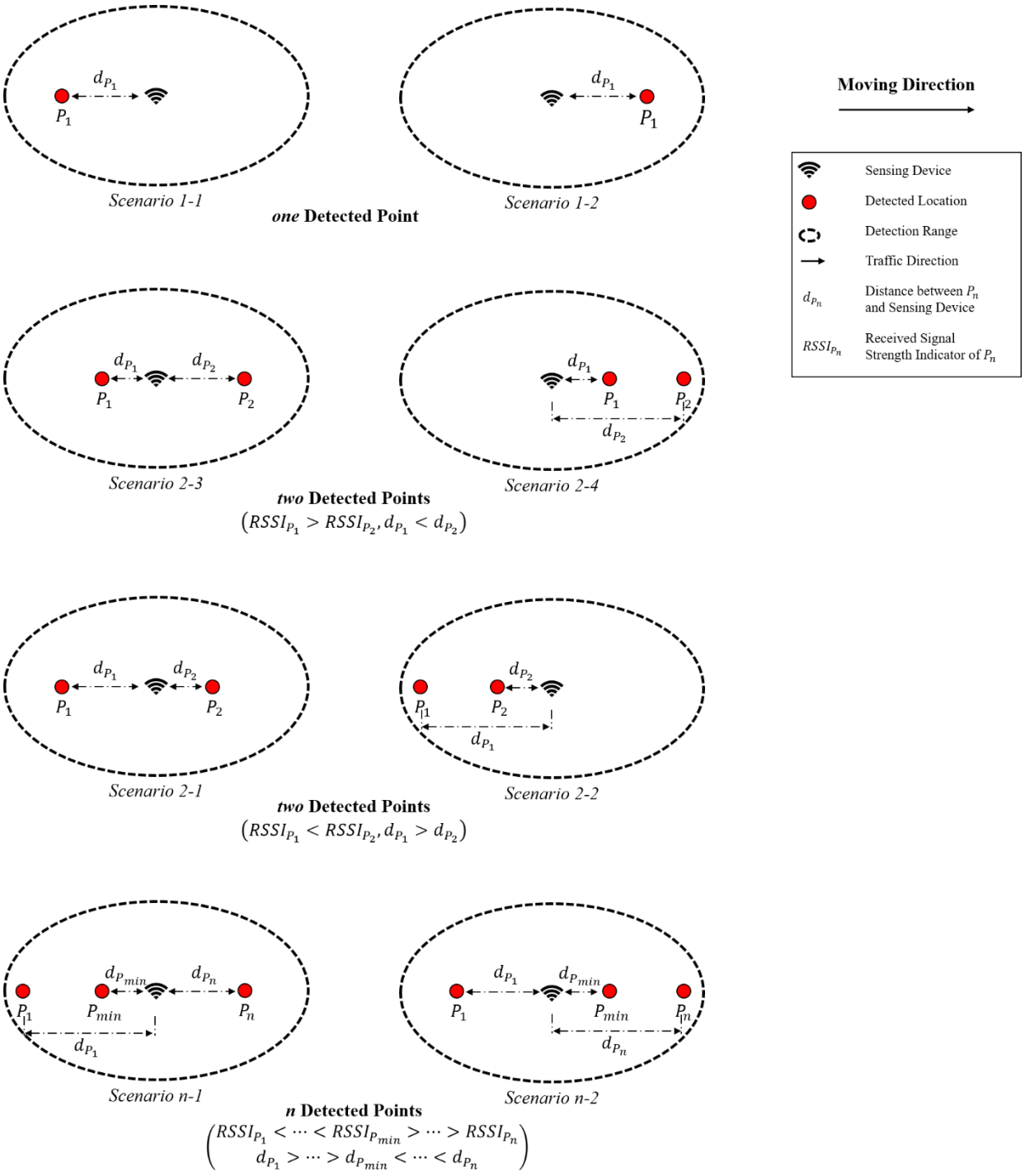


Figure 3-3 Relative Position of Detection Locations and MAC address detectors

Thus, if a WBM device is detected more than once either at the start or the end sensor location, the estimated traffic speed can be corrected by Equation (3-5) based on a single pair of detection points at the start and the end sensing locations,

$$Corrected\ Speed_{s-e} = \begin{cases} \frac{d_{device} - d_s + d_e}{TT_{s-e}}, if\ P^s = \{P_n^s | n \in (min, n]\}\ and\ P^e = \{P_n^e | n \in (min, n]\} \\ \frac{d_{device} + d_s - d_e}{TT_{s-e}}, if\ P^s = \{P_n^s | n \in [1, min)\}\ and\ P^e = \{P_n^e | n \in [1, min)\} \\ \frac{d_{device} + d_s + d_e}{TT_{s-e}}, if\ P^s = \{P_n^s | n \in [1, min)\}\ and\ P^e = \{P_n^e | n \in (min, n]\} \\ \frac{d_{device} - d_s - d_e}{TT_{s-e}}, if\ P^s = \{P_n^s | n \in (min, n]\}\ and\ P^e = \{P_n^e | n \in [1, min)\} \end{cases} \quad (3-5)$$

where d_{device} is the distance between start and end sensing locations. P^s and P^e are the detection points at the start and end sensing locations. d_s and d_e are the distances between P^s and P^e to the MAC address detector, and TT_{s-e} is the detection time difference between the timestamps P^s and P^e. The traffic speed can be corrected by adding or subtracting d_s and d_e depending on whether P^s and P^e are detected before or after the detection point P_{min} which is the detection point that has the highest RSSI measurement. Typically, a specific MAC address can be detected multiple times at a sensing location. Thus, the estimated traffic speed of a trip based on MAC address matching can be corrected by averaging the corrected speed of all pairs of detection points at the start and end sensing locations based on Equation (3-6),

$$Corrected\ Speed_{trip} = \frac{\sum_i^{\boldsymbol{S}} \sum_j^{\boldsymbol{E}} Corrected\ Speed_{i-j}}{\boldsymbol{S} \times \boldsymbol{E}} \quad (3-6)$$

where S and E are the total numbers of detection points at the start and end sensing locations, i represents the detection point at the start sensing location, and j represents the detection point at the end sensing location. As previously discussed in the above, the proposed method for traffic speed correction is not feasible when a WBM device is detected only once at both the start and end sensing locations. However, Figure 3-4 (a) shows the boxplot of detection times of WBM devices when they are separated into car, bike, and walking modes. The average detection times is larger than once for all three modes, which indicates the impact of having only one detection point is trivial, and the proposed traffic speed correction method can be implemented for addressing localized spatial uncertainty.

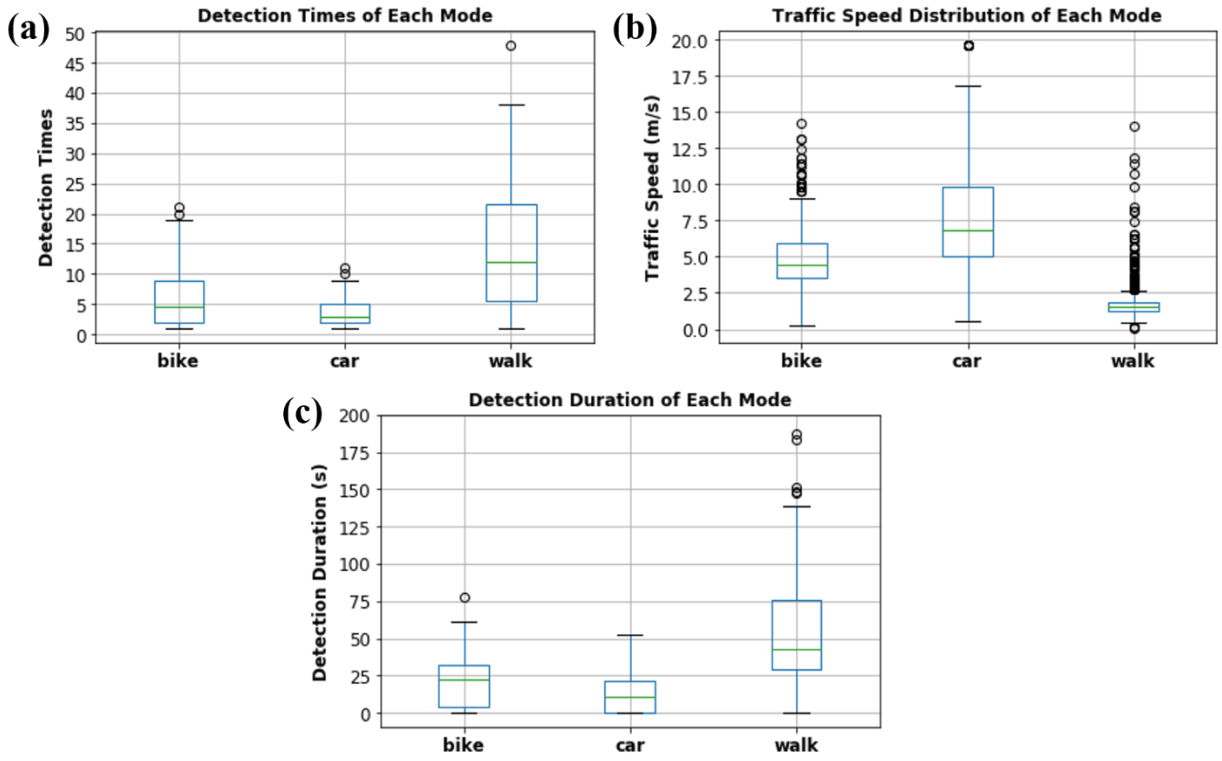


Figure 3-4 Boxplot of Key Features. (a) Detection Times, (b) Ground Truth of Traffic Speed of Each Mode, and (c) Detection Duration

3.1.2 *Traffic Mode Uncertainty*

To identify the traffic mode through MAC address data, pre-defined hard thresholds of extracted features (e.g. estimated traffic speed) were used in previous studies. However, as it is introduced in the literature review, this method has issues with both accuracy and efficiency. In this research, fuzzy clustering is proposed to identify the traffic mode based on the extracted features. The following sections will introduce the feature extraction and the proposed fuzzy clustering, respectively.

3.1.2.1 Features Extraction

MAC address detectors can be used as either fixed sensing equipment or mobile sensors. For using them as fixed sensing equipment, the features only can be extracted from MAC address data. When

WBM devices are carried by travelers in different traffic modes, they will stay within the detection range for different time durations, and the travel times between adjacent sensing points will be different as well. Thus, the number of detections, detection duration, detection time difference, and the estimated traffic speed are the potential features. Besides MAC address data, high-resolution GPS location data from mobile sensors can provide useful information as well. The potential features include travel distance, average speed, maximum speed, and the location when a MAC address is detected at the first and last time. All in all, there are two data sources that can be used for feature extraction, and the features need to be selected depending on specific implementations.

3.1.2.2 Fuzzy Clustering

The disadvantages of the traffic mode identification method based on pre-defined thresholds can be summarized in two parts. On one hand, since the values of the extracted features vary according to time and location, it is hard to develop a general method to determine the values of the thresholds. On the other hand, for the hard thresholding, it is assumed the feature spaces among different traffic modes are well separated, a condition which might not be satisfied in reality. As seen from Figure 3-4, all three features are overlapping for different traffic modes. Thus, to accommodate these two disadvantages, fuzzy clustering is proposed for traffic mode identification.

Clustering algorithms are a class of unsupervised machine learning algorithms. The goal of clustering is to separate a finite unlabeled data set into discrete underlying data structures. In this case, no pre-defined values of features need to be provided as the input. As shown in Figure 3-5, there are two types clustering algorithms in terms of the types of boundaries among clusters, those being hard clustering and fuzzy clustering. In general, hard clustering is good at coping with data sets with clear boundaries, e.g. K-Means clustering. The membership function of each data point

belonging to a specific cluster equals either 1 or 0, which means a data point is assigned to only one cluster. However, for fuzzy clustering, the data points can be assigned to all clusters with a certain degree of membership which reflect the possibility of a data point belonging to a specific cluster. This is quite useful for dealing with data sets with ambiguous boundaries. Among the class of fuzzy clustering algorithms, many existing algorithms have been demonstrated to be effective for image segmentation [99], sensor network optimization [100], stock performance prediction [101], and medical analysis [102]. The previous studies employed different fuzzy clustering algorithms with modifications based on specific challenges. For addressing traffic mode uncertainty in this research, fuzzy clustering algorithms need to be selected based on specific implementations, and the algorithms need to be modified to fit the nature of specific problems. The proposed fuzzy clustering algorithms for identifying traffic modes in different scenarios will be introduced in Chapters 4 and 5.

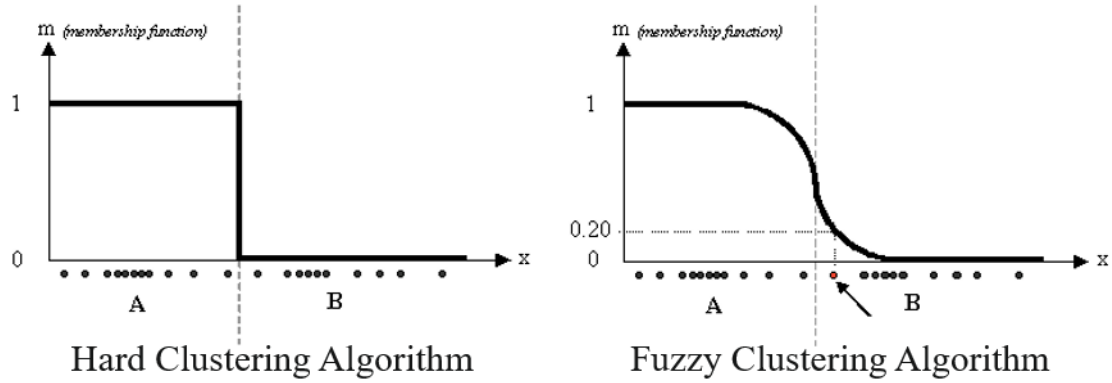


Figure 3-5 Comparison of Hard Clustering and Fuzzy Clustering

3.1.3 *Population Uncertainty*

To address population uncertainty of the passive Wi-Fi and BT sensing method, two methods are considered effective, one is population estimation and another one is device-free wireless sensing technology. For population estimation, a set of labelled data is required to train an estimation

function. For device-free wireless sensing technology, additional sensors need to be deployed on-site. The method selection needs to be re-considered based on specific challenges. The following section will introduce these two methods separately.

3.1.3.1 Population Estimation

Once the traffic modes of WBM devices are identified, a sample of the population is obtained as the outcome. Then, it is required to estimate the entire population based on the detected sample. Basically, the detection rate of WBM devices depending on the penetration rate of discoverable WBM devices at a specific location. It is assumed that the population shares the same penetration rate at a specific location. Therefore, the detected sample can be considered as representative of the total population, which means the main trend of variation can be reflected by the detected sample. Then, Equation 3-7 can be used to estimate the population based on the detected sample.

$$F(X_i) = P_i \tag{3-7}$$

where $F(\cdot)$ is the estimation function which maps the detected sample to the total population, X_i is the detected sample at timestamp i, and P_i is the estimated population at timestamp i. For the estimation function, a set of labelled data is essential to train the coefficients of the estimation function. The function can be selected as a linear or non-linear function according to the specific implementations. Previously, several methods were demonstrated to be effective for estimating the population based on a sample of MAC address data, including scaling with a fixed number [55], linear regression [56], and power function and Fourier function based methods [57]. Among the existing methods, the non-linear functions performed better than linear functions for multiple implementations, e.g. traffic volume estimation [84]. Thus, the non-linearity and linearity of the estimation function should be considered to fit the intrinsic attributes of the problem.

3.1.3.2 Device-Free Wireless Sensing using Wi-Fi Channel State Information

In this research, a device-free wireless sensing technology utilizing Wi-Fi CSI is proposed. Wi-Fi CSI is a high-resolution signal strength indicator which is primarily influenced by the static environment and moving objects in surrounding areas. To capture Wi-Fi CSI signals, Wi-Fi CSI sensor is consisted of two devices, those being Wi-Fi frame transmitter and the receiver. During the sensing process, the transmitter keeps sending Wi-Fi frames to the receiver through a Line of Sight (LoS). The Wi-Fi communication will be influenced when an object is moving around the LoS. Figure 3-6 presents an example of sensing pedestrians in an indoor environment. Basically, the Wi-Fi frame communication is reflected off the moving pedestrian so that Wi-Fi CSI will generate a perturbation which can be used for sensing the presence and moving features of the object. In this case, the detection no longer relies on sensing WBM devices of travelers, and thus, the population uncertainty issue can be solved. In previous studies, Wi-Fi CSI has been demonstrated to be effective for multiple sensing objectives in indoor environments, e.g., indoor localization at the decimeter-level [103], human presence detection [104], indoor crowd counting [93], human activities recognition [105], and fall detection [106]. No existing studies shed light on pedestrian detection in outdoor environments. Thus, in this research, a device-free mobile sensing technology for pedestrian detection is developed for addressing population uncertainty of passive Wi-Fi and BT sensing technology. Details of the proposed sensing algorithm will be introduced in Chapter 5.

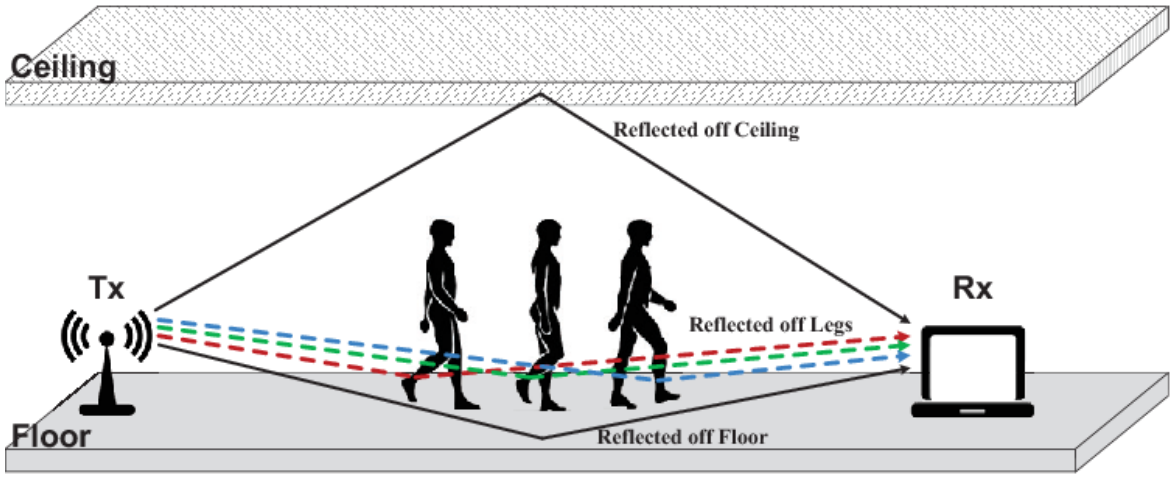


Figure 3-6 Pedestrian Monitoring by Wi-Fi Channel State Information

3.2 Wireless Sensing System Development

To implement the proposed wireless sensing methods and technologies, components beyond a set of data analysis and method are required. A thorough wireless sensing system based on passive Wi-Fi and BT sensing is indispensable for data collection, data communication, data analysis and data visualization. In this section, a generalized method for establishing a wireless sensing system is introduced. The wireless sensing system architecture, MAC address detector, and data management and visualization platforms will be introduced separately.

3.2.1 *Wireless Sensing System Architecture*

A wireless sensing system primarily consists of two parts, the wireless sensors for data collection, and the remote server for data analysis, management, and visualization. Figure 3-7 shows the architecture of a wireless sensing system. The major role of wireless sensors on the edge side is to sense discoverable WBM devices. Real-time wireless sensing data is transmitted to a remote server either by cellular network or Ethernet communication. On the serve side, an advanced database is deployed for managing wireless sensing data and other relevant data sources. Data analysis

modules are used to address uncertainties and for traffic parameter extraction. Finally, the results are visualized and broadcasted to transportation managers and road users through a visualization platform, and, also, the platform will be responsible for dealing with the users' requests. Each component of the system can be customized to fit the needs of a specific implementation. For example, wireless sensors can be either fixed or in-vehicle MAC address detectors, and the data analysis modules can be targeted to extract transit ridership flow or multi-modal traffic speed. In either case, the system architecture will remain the same for different implementations.

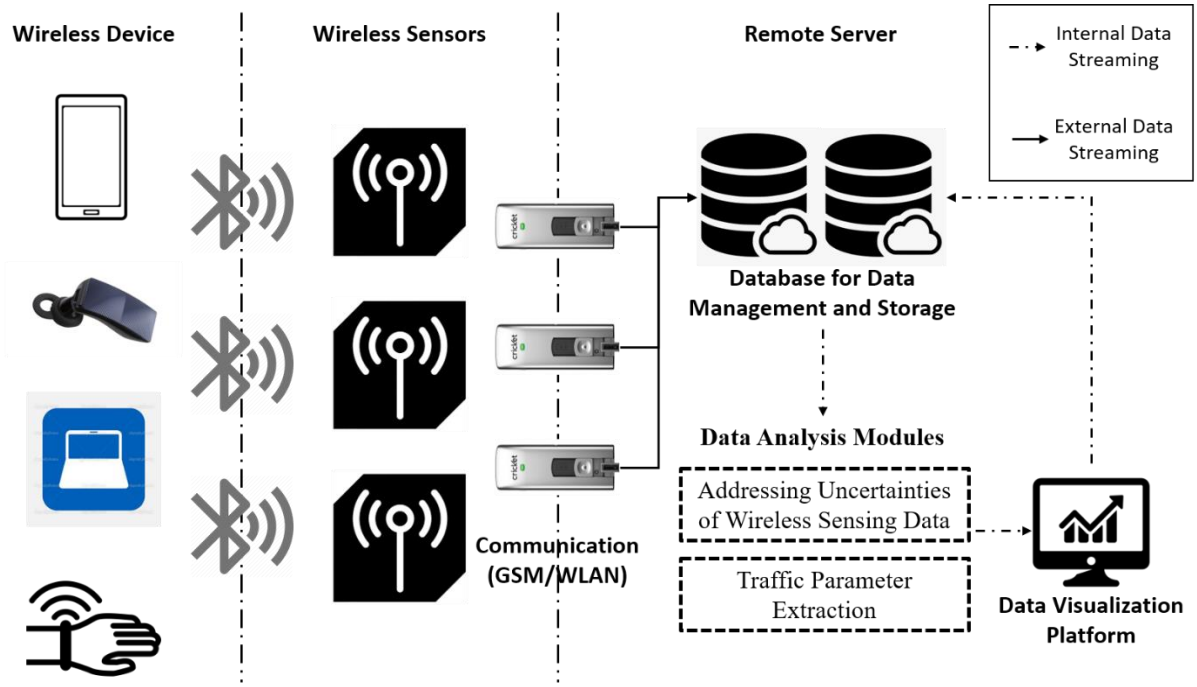


Figure 3-7 Sensing System Architecture

The real-time data streaming between the sensor-side and the server side is supported by remote communication protocols, including Transmission Control Protocol (TCP) and User Datagram Protocol (UDP) [107]. These two communication protocols are main members of the internet protocol suite, and each has clear advantages and disadvantages. TCP is a connection-oriented protocol, and thus data is guaranteed to be received if a connection exists. Consequently, the data packet header of TCP is rather large resulting in higher communication costs. UDP is a

connectionless protocol meaning a connection between the terminal and server is not required while data is transferring. That said, the delivery of data is not guaranteed under the UDP protocol, and the data reception rate is influenced by communication bandwidth, data generation rate, and terminal performance [108]. However, since the UDP data packet has a smaller header than the TCP header, it is more suitable for scenarios with limited communication sources.

3.2.2 *Customized MAC Address Detector*

3.2.2.1 Hardware Design

The MAC address detector is one of the most significant components of the proposed wireless sensing system. The customized MAC address detector is composed of four components, including sensing modules, the data processing unit, the communication module, and the power supply. All sensing modules are integrated into a customized USB board which connects with the data processing unit through GPIO pins (see Figure 3-8).

Sensing Modules: Four sensing modules are necessary, Wi-Fi module, BT module, GPS module and real-time clock.

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 monitor mode [109]. In the design, the Ralink 5370 Wi-Fi chipset is proposed. Its detection range is about 60 meters, and the 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. The BT 4.0 BCM20702 chipset is used in this study. The detection range is about 20 meters.
3) GPS module: to use as a mobile sensor, a GPS module is employed to record the high-resolution latitude and longitude. The 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 sensing programs run in parallel on the data processing unit via automatic start-up scripts. The MAC address data and GPS location matching is based on the timestamp. Most 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 conducted, the clock will not be synchronized. Even through the GPS module can help with time synchronization, it still can ruin the data quality due to 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 study. The Pi Zero is a single-board computer with a 1.0 GHz Single-Core CPU and 512 MB RAM [110]. Other Internet of Things (IoT) devices can be used for this kind of implementation, e.g. NVIDIA Jetson NANO, Asus Tinker Board, and Arduino Uno R3.

Communication Modules: A 4G USB Modem is a necessary hardware component in order to enable data communication using a cellular network. A cellphone data SIM card is plugged into the 4G modem, and the modem connects with the sensor via a USB interface. The connection of the MAC address detector to the cellular network is activated via the Network Manager API in the software which allows automatic network connection upon start-up and automatic re-connection to the Internet whenever the connection fails. Ethernet communication or Wi-Fi service can be used as alternatives for the data communication module.

Power Supply: Both power wire and portable power bank are ideal for supporting the sensing equipment's operation. As the data processing unit has low energy consumption, a portable power bank can support the operation for a whole day.

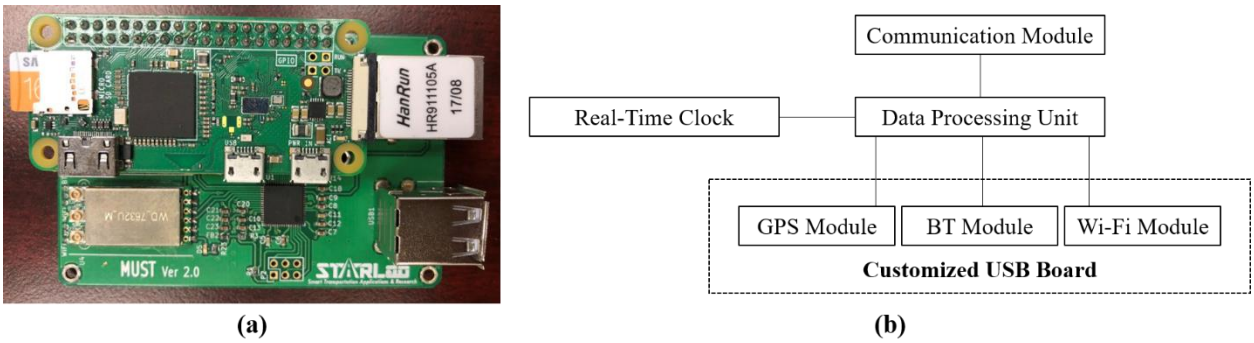


Figure 3-8 Customized MAC Address Detector. (a) Integrated Equipment, and (b) Conceptual Framework of Customized USB Board

3.2.2.2 Software Architecture

Within the software running on the customized MAC address detector, there are four main functions are designed including sensing programs, sensor status monitoring programs, data communication programs, and a local log. Figure 3-9 presents the four functions and their relationships. The sensing programs are designed to communicate with sensing modules (Wi-Fi, BT, and GPS) for setting the modules' mode, defining hyper-parameters of sensing, and passing data to the communication socket and local log sub-system. Sensor status monitoring programs take the responsibility of checking the status of the hardware and the operation system. The monitoring history will be recorded in a local log, and the recording of abnormal actions will pass to the communication packet to warn the managers on the server-side. The communication programs are developed for dealing with two directional communication with the remote server, including transmitting real-data and warning messages, and allowing remote access control function from managers. The local log is responsible for recording the sensor's operation history

and storing real-time data locally in log files. In summary, the designed software framework can support the customized MAC address detector's operation with reliability, stability, and efficiency.

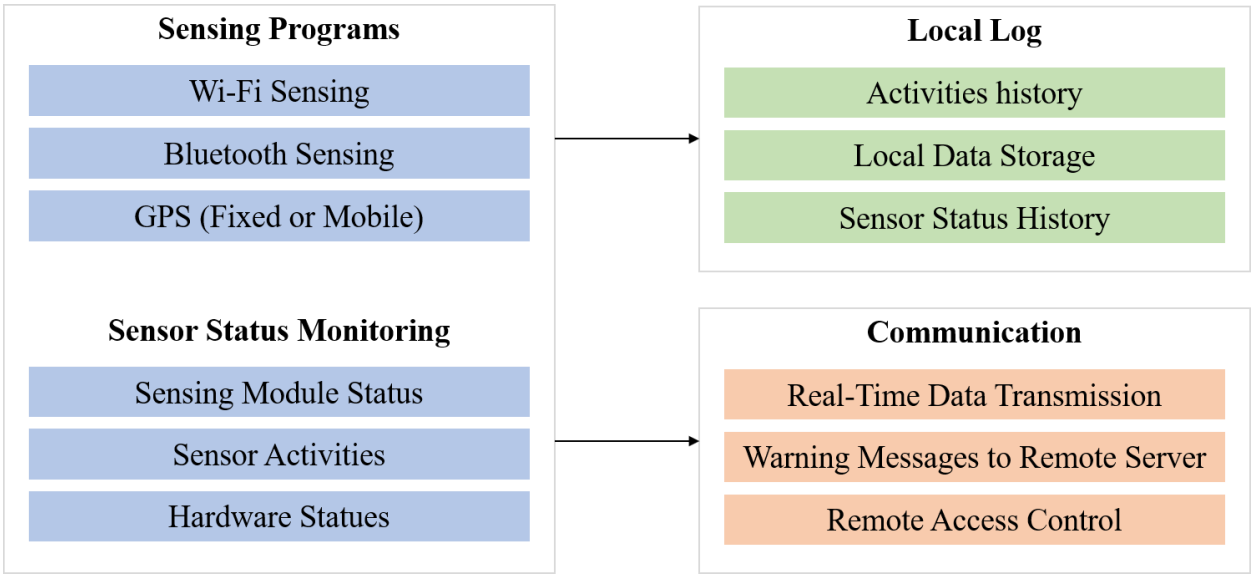


Figure 3-9 Framework of Sensing Software

3.2.3 *Data Management and Visualization*

3.2.3.1 Database Design

For visualizing traffic parameters, spatial data and non-spatial data need to be matched in the database. Basically, spatial data are stored as shapefiles [111], which usually contain geospatial information, such as polylines, points and polygons. The real-time wireless sensing data can be treated as non-spatial data. To present non-spatial data in a GIS system, one must link the spatial data and non-spatial data together. To solve the data matching task, a relational database was created based on Open Street Map (OSM) geospatial data [112]. In the designed database, geospatial data are encoded in the manner of well-known binary (WKB) format which is defined by the Open Geospatial Consortium (OGC); the binary content is used to encode vector geometry. The GPS location (e.g., latitude and longitude) and projection information are encoded in a hexadecimal string and can be decoded to a sequence of latitude and longitude pairs to view in any

GIS or mapping system for data visualization purposes. Figure 3-10 shows the proposed database structure. There are two spatial data tables and four non-spatial data tables in the database. Foreign keys were used to link tables for cross-referencing. The wireless sensing data table is responsible for store the real-time data which are collected by wireless sensors. The latitude and longitude of wireless sensing data are the foreign keys to find the corresponding sensor ID and trip ID. The vertices' OSM IDs are the keys for matching spatial and non-spatial data. In this manner, relevant geospatial information can be easily extracted from spatial data tables and matched with non-spatial data for transportation data analysis.

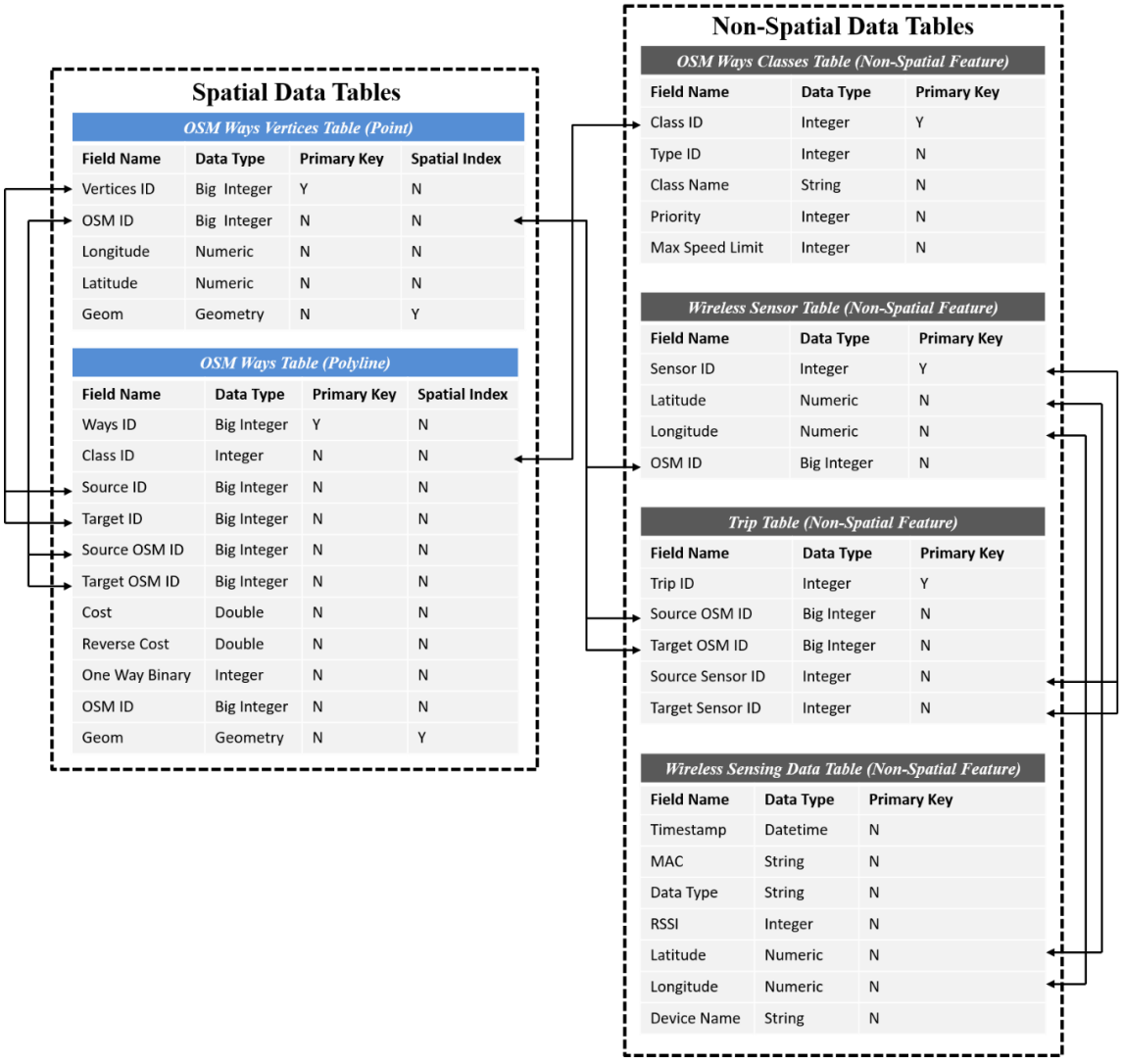


Figure 3-10 Database Structure Design

3.2.3.2 Analysis and Visualization Platform Design

To visualize real-time traffic parameters being extracted from wireless sensing data, a multi-tier data management, analysis and visualization platform is developed. Figure 3-11 presents the conceptual design. As shown in the figure, there are three tiers in the platform: the data tier, the data matching and computational tier, and the presentation tier.

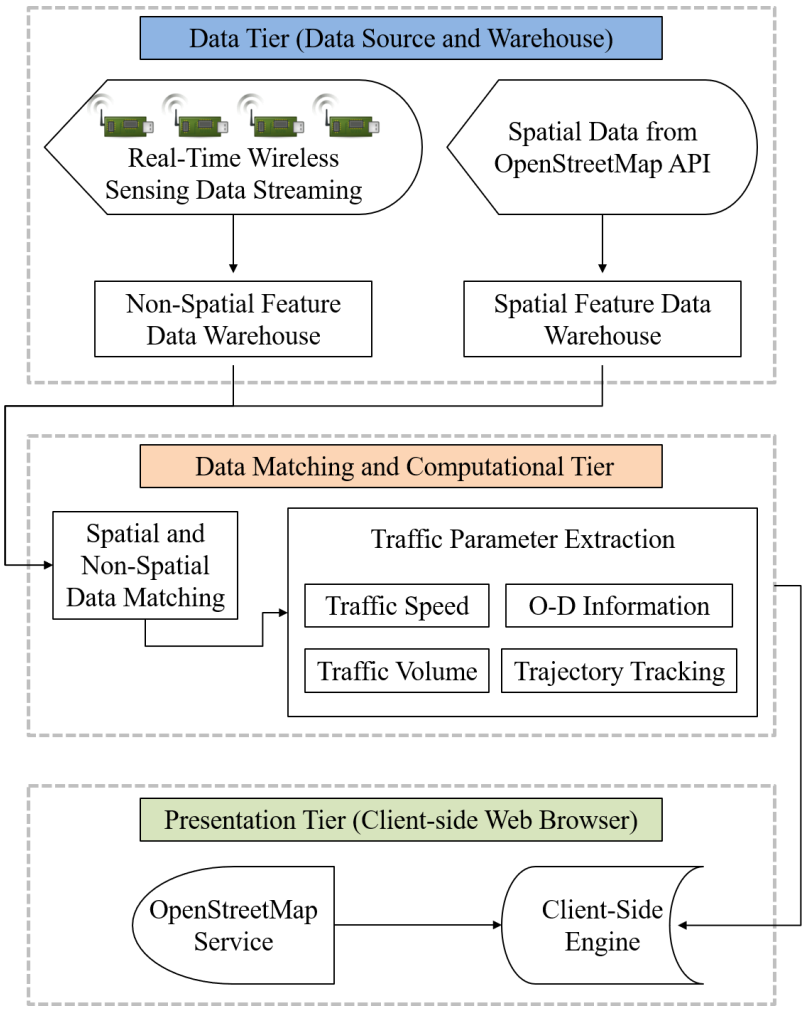


Figure 3-11 System Design

The data tier is responsible for the data reception service and data warehousing. The real-time wireless sensing data are remotely sent from customized wireless sensors by communication protocols. The data server is set in listening mode and data are inserted into a relational database

management system (RDBMS) in a real-time manner. The geospatial data are downloaded from the OSM application program interface and will automatically update when any renewal in the data source occurs.

The data matching and computational tier is used for spatial and non-spatial data matching and traffic parameter extraction. This tier will process the data based on client input and send the results to the client-side web server for data visualization. The client-side engine can parse the encoded geospatial data into a sequence of latitude and longitude pairs, and present these data on OSM map system.

The current chapter mainly presented the general methodology to address uncertainties while implementing passive Wi-Fi and BT sensing technology for traffic data collection. The following two chapters, Chapters 4 and 5, will focus on applying the proposed methodology in transit rider and non-motorized traffic detection.

Chapter 4. TRANSIT RIDERSHIP FLOW MONITORING

4.1 INTRODUCTION

Public transit ridership flow and O-D information are crucial for transit network planning, routes optimization, service quality improvements, and travel scheduling [9]–[12]. It is also an essential data input of Internet of Vehicles (IoV) in transit systems [113]. In this section, a novel transit ridership flow monitoring method based on passive Wi-Fi and BT sensing technology is proposed. As it is introduced in section 2.2.1, there are two uncertainties need to be solved for this application, those being traffic mode uncertainty and population uncertainty. For the traffic mode uncertainty, based on the assumption of the overlapping feature spaces of passengers and non-passengers, a Fuzzy C-Means (FCM) clustering algorithm is proposed for separating passenger and non-passenger MAC addresses in this research. FCM is one of the most popular fuzzy-based clustering algorithms which is suitable for separating the clusters with ambiguous boundaries [114]. 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. For the population uncertainty, considering the non-linearity among the dataset, a Random Forest regression (RF) model [115] is proposed for estimating the population ridership flow in this study, including the number of onboard, boarding and alighting passenger.

The focus of this section is to construct a novel three-step algorithm framework for addressing the uncertainties of wireless sensing technology. The target parameters include the number of onboard, boarding, and alighting passengers, and O-D information. The main contribution of this research can be summarized as follows:

1) A novel three-step data-driven algorithm framework is proposed for separating transit passengers and non-passengers and estimating population ridership flow based on passive Wi-Fi and BT sensing data.
2) 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 by comparing with other selected baseline models.
3) 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 section is organized as follows. Section 4.2 presents the proposed three-step data-driven algorithm. Section 4.3 describes the experimental design and the numerical results are presented in Section 4.4. This chapter is summarized by concluding the research findings.

4.2 PROPOSED METHODOLOGY

4.2.1 *Algorithm Framework*

The proposed algorithm framework is designed for mining real-time transit ridership flow using Wi-Fi and BT sensing data which is presented in Figure 4-1. 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 Fuzzy-C Means (FCM) clustering algorithm is employed to cluster the MAC addresses into the passenger and non-passenger clusters. In step three, 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.

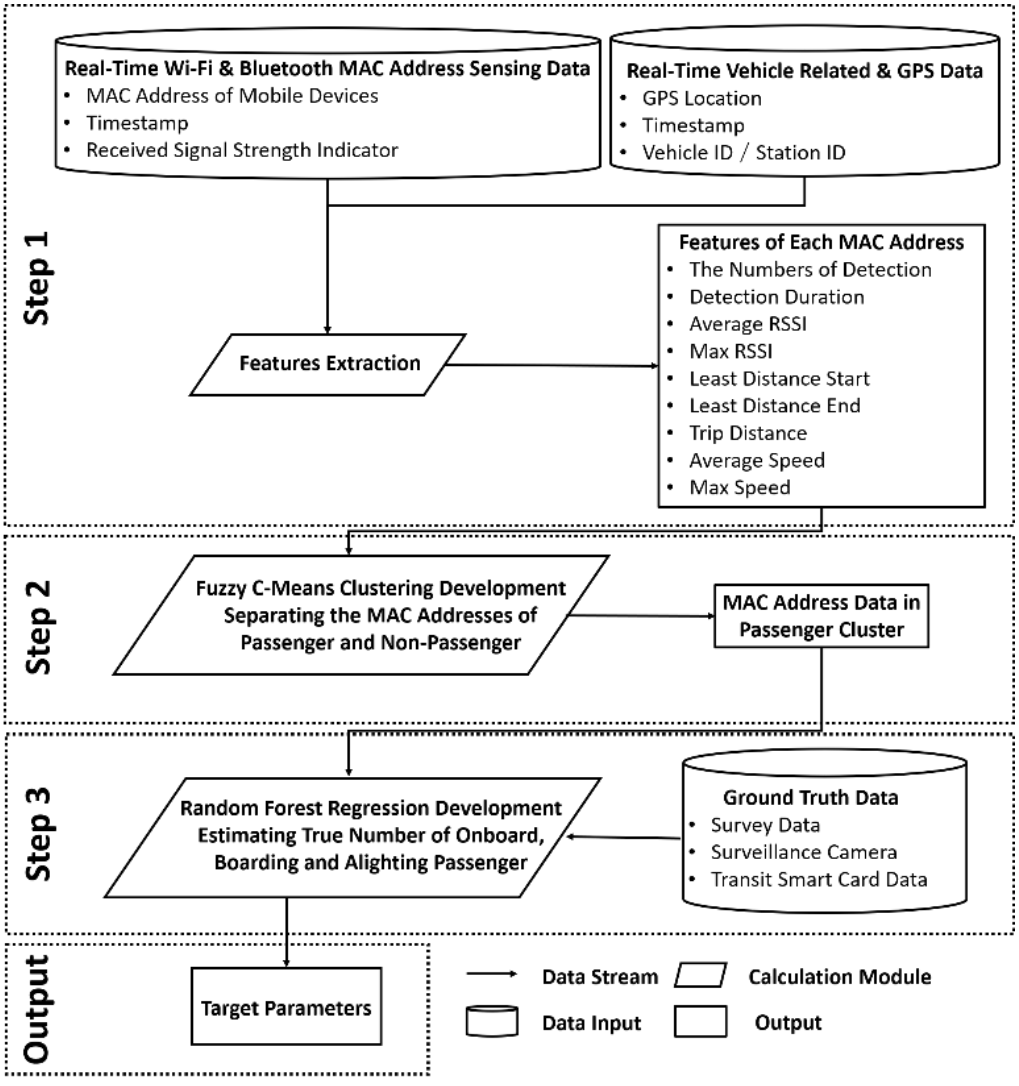


Figure 4-1 Algorithm Framework

4.2.2 *Feature Extraction*

Since the detection range is not exactly the inside space of transit vehicle, the MAC addresses come from the WBM devices outside the vehicle can be detected as well. In the following situations, the MAC address of non-passenger WBM devices can be detected by the on-board MAC address detectors.

- Fixed Wi-Fi or BT-enabled devices within the sensing range.
- WBM devices of the passengers standing at stations.
- WBM devices of the pedestrians or the bicyclists within the sensing range.
- WBM devices in other vehicles within the sensing range.

For the fixed Wi-Fi or BT-enabled devices and the WBM 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 few times and in a short time-period. For the WBM devices in other vehicle or of pedestrian and bicyclists, even they travel in parallel with transit vehicle, the MAC address features also would be different with 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, 9 features were extracted from the MAC address and GPS data which are presented in Table 4-1. 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 data with GPS location is transmitted from the

MAC address detector to the remote server for feature extractions. In this step, the output is the vectors of each MAC address and its features.

Table 4-1 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)

4.2.3 *Separating Passenger and Non-Passenger MAC Address*

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 [116]. Thus, fuzzy clustering algorithms usually are useful when the boundaries among clusters are ambiguous [117], which satisfies the characteristics of the overlapping feature spaces of passengers and non-passengers. Fuzzy-C Means (FCM) clustering is one of the most popular fuzzy clustering algorithms. It attempts to minimize the cost function J in Equation (4-1) 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 $\chi = (X_1, X_2, X_3, \cdots, X_N)$ denotes a set of N MAC

address to be partitioned into C clusters. $X_j = (x_1, x_2, x_3, \cdots, x_n)$ denotes n features of each MAC address. Then, the cost function J would be calculated as the following equation:

$$J = \sum_{j=1}^{N} \sum_{i=1}^{C} u_{ij}^{m} \left\| X_j - v_i \right\|^2 \qquad (4-1)$$

where m is the parameter for controlling the fuzzification, $u_{ij} \in [0,1]$ is the membership function of jth data point in cluster i, which represents the possibility that jth MAC address data point whether belongs to a passenger or not, thus, $\sum_{i=1}^{C} u_{ij} = 1 \ (j = 1,2, \cdots, N)$. v_i is the center of 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 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.

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{\left\| X_j - v_i \right\|}{\left\| X_j - v_k \right\|} \right)^{2/(m-1)}} \qquad (4-2)$$

$$v_i = \frac{\sum_{j=1}^{N} u_{ij}^{m} X_j}{\sum_{j=1}^{N} u_{ij}^{m}} \qquad (4-3)$$

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 Fuzzy C-Means clustering is presented in Algorithm 4-1. 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 hyper-parameter settings for achieving

optimal performance [118]–[120]. In this study, the parameters are set based on the results of previous studies. The fuzzification parameter m is set as 2 [121], the number of clusters C is set to 2 representing the clusters of passengers and non-passengers. The ε and L in the algorithm are set as 0.001 and 1000, respectively.

Algorithm 4-1: The Training Process of Fuzzy C-Means

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 (3-2), U^0 is a $C \times N$ matrix

Repeat:

Update $v_i^{(t)}$ using (4-3) based on U^{t-1}

Compute U^t using (4-2) based on $v_i^{(t)}$

Until: $\|U^t - U^{t-1}\| \leq \varepsilon$ or $t \geq L$

4.2.4 *Passenger Population Estimation*

In this study, Random Forest (RF) regression model was employed for the estimation task. RF regression is a widely used non-parametric machine learning regression algorithm. As shown in the previous study, RF regression can capture the non-linear relationship in the data set which result in a better goodness of fit than linear regression [122]–[124]. The general concept of RF is introduced by Breiman in 2001 [115]. In this study, the CART (Classification and Regression Trees) algorithm [125] 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 random forest, the pruning process will be ignored since the generalization error of a random forest will always converge.

For the RF regression model development in this study, 5 variables were selected as the regressors, including the day of week, the hour of day, the minute of hour, the dummy variable of whether the current stop is the last stop of the trip, and the number of passenger MAC addresses.

4.2.5 *Algorithm Evaluation*

4.2.5.1 Passenger and Non-Passenger Separation

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 dataset 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 [126]. 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 [127], which could further improve the performance of the GMM model. In addition, BGM can avoid overfitting by eliminating parameters using integration [128]. The model specification can be found in the references [126], [129]. 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 [130]. 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 [131]. Compactness and separation are two main criteria for evaluating cluster similarity. Compactness measures the intra-distance of each cluster and separation measures inter-distance [132], [133]. The following metrics are employed to measure compactness and separation, including Silhouette coefficient, Dunn’s index, Davies-Bouldin index, and Beta CV measurement.

Silhouette coefficient (SC) [134] evaluates the performance of clustering result based on the pairwise difference of inter and intra distances of clusters, which is simply expressed as equation below.

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

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 inter-cluster distances whilst minimizing intra-cluster distances [135]. The Dunn's validation index is calculated as

$$DU = \min_{1 \le i \le c} \left\{ \min_{\substack{1 \le i \le C \\ j \neq i}} \left\{ \frac{\delta(C_i, C_j)}{\max_{1 \le k \le c} \{\Delta(X_k)\}} \right\} \right\} \quad (4-5)$$

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

Davies-Bouldin (DB) index [136] is the ratio of the sum of intra-cluster distance to inter-cluster separation, which is expressed by

$$DB = \frac{1}{k} \sum_{i=1}^{k} \max_{i \neq j} \frac{\left(D(C_i) + D(C_j)\right)}{D(v_i, v_j)} \quad (4-6)$$

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

Beta CV Measure (Beta CV) [137] is a measurement of clustering validation based on the ratio of the mean intra-cluster distance to the mean inter-cluster distance which can be calculated as below [138]

$$Beta\ CV = \frac{Distance_{intra}/N_{intra}}{Distance_{inter}/N_{inter}} \quad (4-7)$$

where N_{intra} is the number of distinct intra-cluster edges, N_{inter} is the number of distinct inter-cluster edges.

4.2.5.2 Passenger Population Estimation

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:

$$MAE = \frac{\sum_{i=1}^{N}\left|\hat{Y}_i - Y_i\right|}{N} \quad (4-8)$$

$$MSE = \frac{\sum_{i=1}^{N}\left(\hat{Y}_i - Y_i\right)^2}{N} \quad (4-9)$$

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\frac{\left|\hat{Y}_i - Y_i\right|}{Y_i} \times 100\% \quad (4-10)$$

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.

4.2.5.3 Comparison with The Existing Filtering Methods

For the relevant existing studies, filtering method was employed for processing the Wi-Fi and BT sensing data to estimate the public transit ridership flow [50]–[52], [139]–[143]. 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. (2016) [50] developed a three-step filtering method for separating passengers and non-passengers. The MAC address which fits any following conditions would be considered as a non-passenger 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 seconds, 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 meters) for Wi-Fi and 300 ft (91 meters) 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. [51] defined a filtering method with four filters. If the features of a unique MAC address meet the following rules which is considered as a non-passenger MAC address: 1) detection duration is less than 3 minutes, 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 meters), 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 and stops nearby and a predefined threshold of the maximum sensor detection range of 200 ft (61 meters) are used to determine the boarding and alighting stops for each MAC address.

4.3 Experimental Design

The data used in this study were collected from 9 trips of 3 routes in Seattle. The detailed description of the study area and statistical summary of the data set are introduced in the following sections.

4.3.1 *Testing Fields and Data Collection*

The study area is three transit routes in the north of King County, including route 32, route 67 and route 372. Figure 4-2 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.

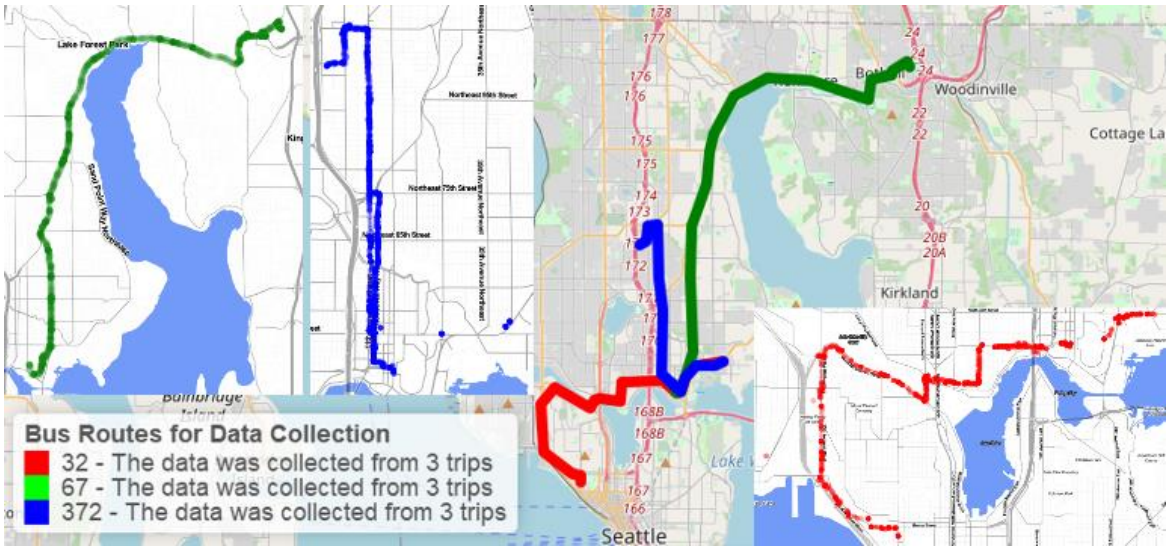


Figure 4-2 Study Area

The data were collected from three trips of each route by the customized MAC address detector. For each trip, the MAC address detector was carried by a volunteer seating in the middle of the vehicle. The sensing equipment 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. The customized MAC address detector is designed based on the method being introduced in section 3.2.1. Figure 4-3 shows the sensing equipment is used in this research.

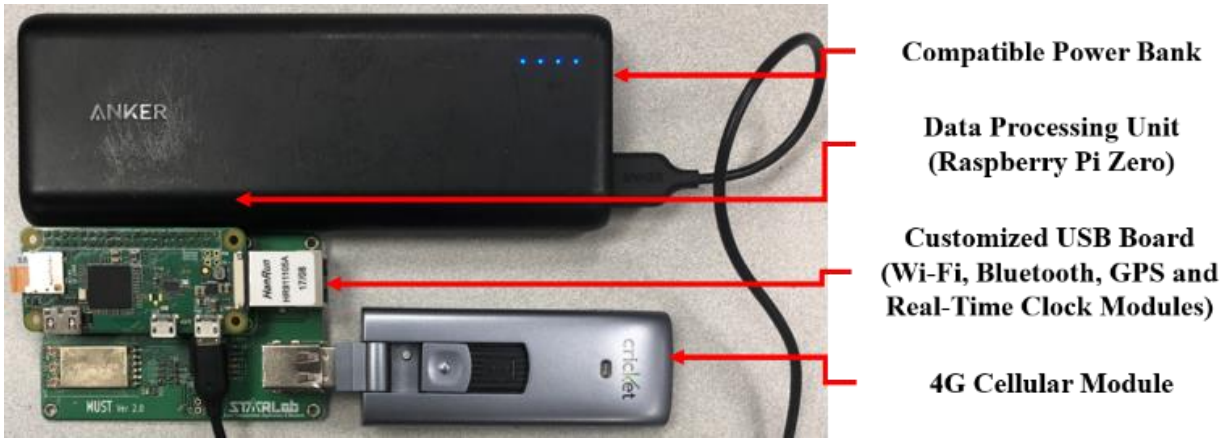


Figure 4-3 Customized MAC Address Detector

4.3.2 *Statistical Summary of the Dataset*

Table 4-2 shows the statistical summary of the dataset. There are 9 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 on-board passengers requesting for taking off. Only the stations where the vehicle stopped were counted as stops in the dataset. Besides the trip information, the amount of MAC address collected from each trip also are introduced. Based on the statistical summary, 17806 data points were collected, including 16027 Wi-Fi data points and 1779 BT data points. Totally, 5064 unique MAC addresses were detected, including 4859 via Wi-Fi network and 205 via BT network. Based on the dataset, averagely, one unique Wi-Fi MAC

address is collected out of 4 Wi-Fi data points and one unique BT MAC address is collected out of 10 BT data points.

Table 4-2 Statistical Summary of the Dataset

Routes No.	Trip Date	Trip Start Time	Trip End Time	The Number of Stops	Number of Data Points		Number of Unique MAC	
					Wi-Fi	BT	Wi-Fi	BT
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

4.4 Numerical Results

4.4.1 *Separating Passenger and Non-Passenger MAC Addresses*

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 non-passenger clusters. The metrics of each model are presented in Table 4-3. 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 4 metrics.

Table 4-3 Evaluation of Clustering Algorithms for Separating Passenger and Non-Passenger 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

Totally, 5064 unique MAC address were clustered by the FCM clustering algorithm into two clusters with 399 passenger MAC addresses and 4665 non-passenger MAC addresses. Based on the FCM clustering results, the statistical summary of each feature is presented in Table 4-4. The mean values of the detection times and the detection duration of passenger MAC address are much larger than those of non-passenger MAC address. The non-passenger MAC addresses have 1.26 average detection times and 4.24 seconds detection duration which is consistent with the assumption that non-passenger 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 non-passenger for all four numbers, which is reasonable that the signal strength of non-passenger's WBM device might be influenced by the bodyshell of the transit vehicle or the larger distance from the in-vehicle MAC address detector. The Least Distance Start and Least Distance End of passenger MAC addresses are about 200 meters which is smaller than those of non-passenger 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 non-passenger MAC addresses are more likely to be detected during the trip where the vehicle is far away from stations. However, since the non-passengers 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 non-passenger. Other three vehicle moving features of passenger MAC address, including trip distance, average speed, and maximum speed, have higher mean values than those of non-passengers for all four-number. The mean values of these three features of non-passengers' MAC addresses are close to zero, which indicates the vehicle almost halted during the time-period when the MAC addresses of non-passengers were detected. It is noticed that the maximum values of

average speed and the max speed of passenger are unreasonably high, which is caused by unstable GPS data.

Table 4-4 Statistical Summary of Passenger and Non-Passenger 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

4.4.2 *Estimating Population Number of Onboard, Boarding and Alighting Passenger*

After separating the passengers MAC address from the dataset, 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. To demonstrate the clustering results of the FCM, the total number of onboard, boarding and alighting passenger of each stop were also counted based on the BGM and GM clustering results.

Table 4-5 Evaluation of The Estimated Number of On-Board 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

Firstly, only the number of onboard passengers was estimated. The evaluation results are presented in Table 4-5. 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 Gaussian Mixture algorithm in RF regression case. The potential reason is that the passenger flow estimation based on Gaussian Mixture algorithm with RF regression might achieve more accurate results for the stations with 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 Gaussian Mixture algorithm in RF regression case is still higher than 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 on-board passengers of each stop based on the FCM clustering results and the ground truth are visualized in Figure 4-4.

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

non-linear 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.

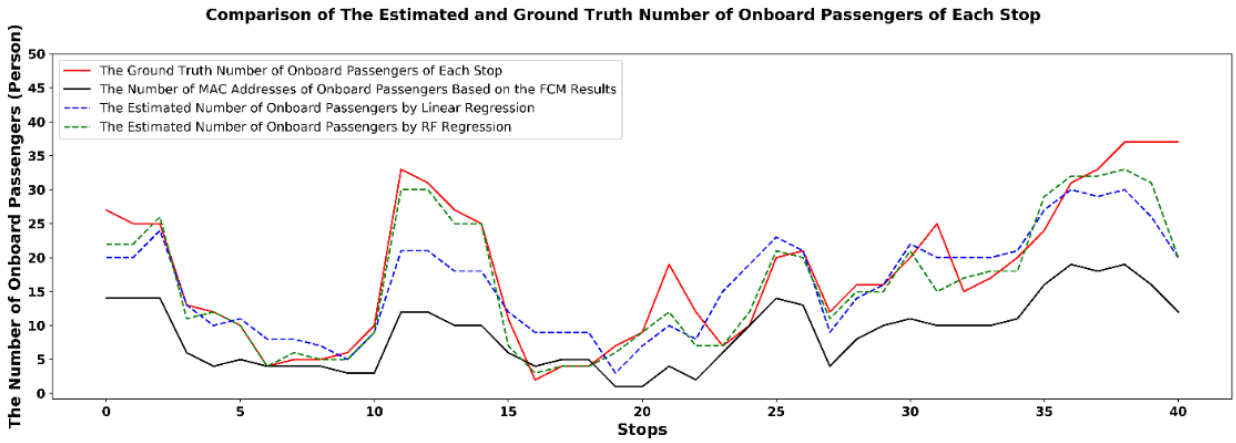


Figure 4-4 Comparison of The Estimated Number of Passengers and The Ground Truth

4.4.3 *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 4-6 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 non-passenger are hard to be well-separated by filters. By considering the overlapped feature spaces of passenger and non-passenger, the FCM clustering algorithm effectively separated the MAC addresses of passenger and non-

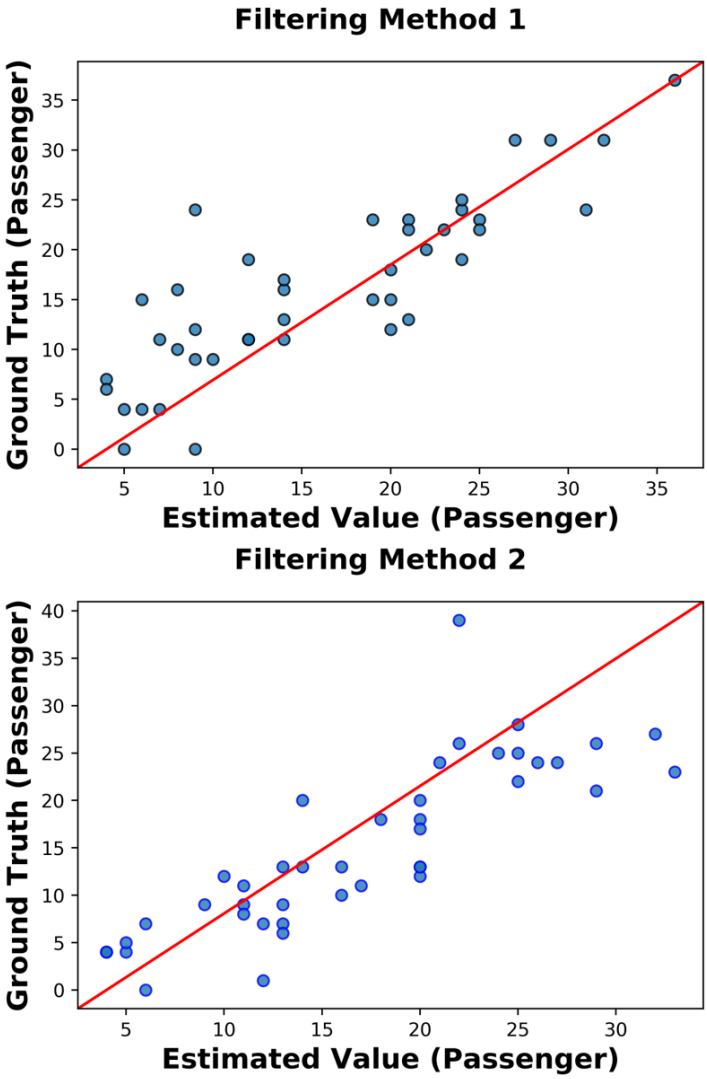
passenger. Furthermore, the RF regression model effectively estimated the population number of onboard passengers by capturing the non-linearity.

Table 4-6 Comparison of The Proposed Algorithm and The Existing Filtering Algorithms

Methods	Fuzzy C-means			Filtering Method 1 [50]			Filtering Method 2 [51]		
	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 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 Figure 4-5. 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 non-passenger 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 non-passenger cluster, e.g. the GPS location was recorded every 20 seconds 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 non-passenger MAC addresses to the passenger cluster since the distribution of signal strength of non-passenger MAC address is similar with the distribution of passenger MAC address. The rightmost scatter plot presents the estimation using the FCM results. The dots are more concentrated around the diagonal line than others. It is noticed 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 large value in the training data set.



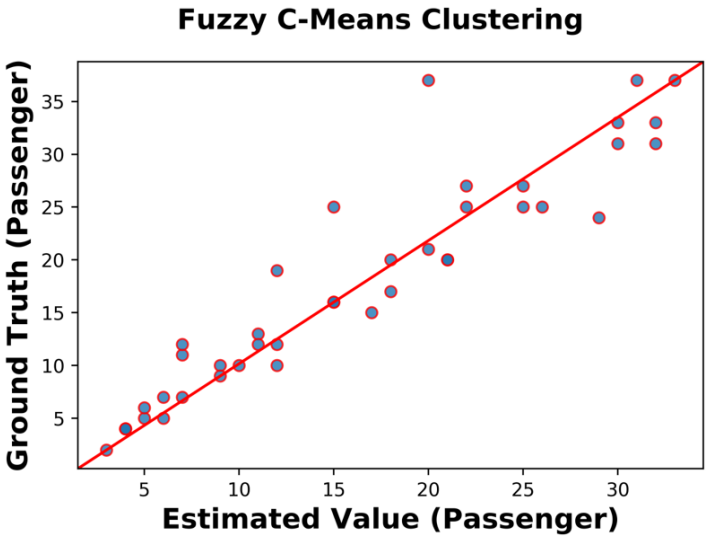


Figure 4-5 Comparison of Filtering Methods and The Proposed Method

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 4-7. 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 noticed 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 dataset.

Table 4-7 Evaluation of 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

4.4.4 *Estimating Real-Time Ridership Flow and O-D Information*

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. 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 Table 4-8 and 4-9. The selected trip was traveled on November 9th, 2018 from 19:38:58 to 20:05:48. Totally, the transit vehicle stopped at 15 stations during the trip.

Table 4-8 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 4-9 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 in-vehicle MAC address detector 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 that which parts of the trip have more travel demands and which stops are more popular for the traveler

Table 4-8 O-D Matrix of the Selected Trip

Alighting / Boarding	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 4-9 The 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

4.5 CONCLUDING REMARKS

In summary, this study proposed a three-step data-driven approach 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 data was collected by the customized MAC address detector from 9 trips of 3 transit routes in Seattle. Multiple metrics were calculated based on ground truth data and the estimates to provide the quantitative evidence for evaluating the estimation performance. According to the results, the proposed algorithm outperformed baseline models and the existing filtering methods in terms of estimation accuracy.

The finding of this study can help to provide real-time accurate transit ridership flow and O-D information for supporting the transit network plan and service optimization. In addition, transit passengers could get a better understanding of the operational status of the transit system for optimizing the travel plan. In this study, only the O-D information of transit partial passengers was achieved. The population O-D inference based on Wi-Fi and BT sensing data could be a valuable research direction for improving the comprehensiveness and accuracy of the whole system

Chapter 5. NON-MOTORIZED TRAFFIC MONITORING

5.1 INTRODUCTION

In this chapter, the methods to address the uncertainties for sensing non-motorized traffic is proposed. The research will be presented in two parts. The first part introduces the proposed method for real-time multi-modal traffic speed monitoring, which mainly deal with the traffic mode uncertainty, and localized spatial uncertainty. The second part shows the research about device-free wireless sensing technology development for pedestrian monitoring, which copes with the population uncertainty.

For real-time multi-modal traffic speed monitoring, the specific contributions include:

1) An algorithm is developed to correct the estimated traffic speed based on Received Signal Strength Indicator of Wi-Fi and BT signals. Ground truth speed measurements and the estimated traffic speed of 408 trips are compared to validate the performance. The accuracy of the proposed algorithm can be implied based on the comparison results.
2) Traffic mode identification algorithm is proposed based on a designed semi-supervised Possibilistic Fuzzy C-Means clustering algorithm. Multiple baseline algorithms are selected for the evaluation purpose. The evaluation results demonstrate the advantage of the algorithm in terms of accuracy.
3) A real-time multi-modal traffic speed estimation algorithm is established. The accuracy of the proposed algorithm is evaluated based on the comparison of estimated results and ground truth data. The evaluation results indicate the proposed algorithm is accurate for all three traffic modes, including walk, bike, and car modes.

For pedestrian monitoring based on device-free wireless sensing, the contributions include:

1) A set of Wi-Fi CSI signal denoising method is proposed based on Hampel Identifier, Linear Signal Interpolation, Kalman Filter and Wavelet Transform.
2) A pedestrian existence detection method is developed based on the level of fluctuations of the normalized average CSI amplitude in the time domain.
3) Pedestrian moving direction identification method is built based on Fresnel Zone theory.
4) Pedestrian moving speed estimation method is constructed by utilizing extracted features from Wi-Fi CSI signals.
5) The effectiveness of the proposed methods is demonstrated by conducting experiments in both indoor and outdoor environments. The impacts of Wi-Fi CSI sampling ratios and antennas are investigated as well.

The remainder of this section is organized as follows. Section 5.2 presents the proposed method for monitoring real-time multi-modal traffic speed. Section 5.3 describes the proposed device-free pedestrian sensing technology using Wi-Fi CSI.

5.2 Multi-Modal Traffic Speed Monitoring

For estimating multi-modal traffic speed based on passive Wi-Fi and BT sensing technology, the challenges mainly focus on traffic mode identification and addressing localized spatial uncertainty. This section presents the proposed method to address these two issues.

5.2.1 *Proposed Methodology*

5.2.1.1 Real-Time Multi-Modal Traffic Speed Monitoring

Algorithm 5-1 presents the proposed multi-modal traffic speed estimation algorithm. Firstly, the algorithm traverses each road segment of a road network to extract valid MAC trips within a specific time window $[t_0, t_0 + \Delta t]$, where t_0 is the start time of a time window, and Δt is a pre-defined time interval. A road segment is the road between two adjacent MAC address detectors, and a valid MAC trip is generated when a unique MAC is detected by two adjacent MAC address detectors in chronological order within a reasonable time range. After achieving all valid MAC trips for a road segment, the traffic speed of each trip will be corrected. The traffic speed correction algorithm is introduced in section 3.1.1 in details. Then, a vector of features which represents the characteristics of each MAC trip will be extracted.

Algorithm 5-1: Real-Time Multi-Modal Traffic Speed Estimation

Initialization: start time t_0, time interval Δt
for road segments i in $\{1, \dots, N\}$ **do**:
  Extract all MAC trips M within time interval $[t_0, t_0 + \Delta t]$
  for MAC trip j in $\{1, \dots, M\}$ **do**:
    Correct traffic speed based on RSSI using Equation (3-6)
    Extract features vector v_j of MAC Trip j
  end for
  Identify travel modes for all MAC trips M using Algorithm 5-2
end for
Output: average multi-modal traffic speed of all road segments for the time window $[t_0, t_0 + \Delta t]$

The extracted features are presented in Table 5-1. Basically, the features not only contain the travel time and speed attributes, but also include the movement features at two sensing stations of a road segment, e.g. detection times and duration. Then, all extracted features will be used as the input of the algorithm for identifying travel modes of each trip using Semi-Supervised Possibilistic Fuzzy

C-Means Clustering (PCM) which will be introduced in section 5.2.1.2. Finally, the average multi-modal traffic speed of a road segment within the time window $[t_0, t_0 + \Delta t]$ is the outcome of the proposed algorithm. The algorithm can be implemented in a real-time way by repeating Algorithm 5-1 for every Δt.

Table 5-1 Extracted Features

Features (Unit)	Unit	Definition
Start Detection Times	Times	The number of times that a unique MAC address is detected at the start sensing location
End Detection Times	Times	The number of times that a unique MAC address is detected at the start sensing location
Start Detection Duration	Seconds	The total amount of time for a unique mac to be detected at the start sensing location
End Detection Duration	Seconds	The total amount of time for a unique mac to be detected at the end sensing location
First Time Difference	Seconds	The detection time difference between the first detected data point at two sensing locations of a road segment
Last Time Difference	Seconds	The detection time difference between the last detected data point at two sensing point of a road segment
Original Speed	Meters/Second	The estimated speed calculating by Equation (4-1)
Corrected Speed	Meters/Second	The estimated speed calculating by Equation (4-6)

5.2.1.2 Travel Modes Identification

To monitor multi-modal traffic speed by passively sensing WBM devices, the traffic modes of valid MAC trips need to be identified based on the features of each trip. In this study, the traffic mode of each MAC trip is identified by the proposed semi-supervised Possibilistic Fuzzy C-Means (PCM) clustering algorithm. The proposed algorithm is introduced in the following sections.

5.2.1.2.1 Fuzzy C-Means Clustering

Fuzzy C-Means (FCM) clustering is a widely used fuzzy-based clustering algorithm [114]. Comparing to traditional hard clustering algorithms, e.g. K-Means clustering, FCM assigns a certain membership function of all clusters to each data point which allows the ambiguous

boundaries among the features of different clusters [117]. The objective function of FCM and the constraints for assigned membership functions of clusters are shown in Equation (5-1) and (5-2),

$$J_{FCM} = \sum_{j=1}^{N}\sum_{i=1}^{C} u_{ij}^{m} {d_{ij}}^{2} \qquad (5-1)$$

$$u_{ij} \in [0,1] \; for \; all \; i \; and \; j,$$
$$0 < \sum_{j=1}^{N} u_{ij} < N \; for \; all \; i, and$$
$$\sum_{i=1}^{C} u_{ij} = 1 \; for \; all \; j. \qquad (5-2)$$

where $u_{ij} \in [0,1]$ is the membership function of jth data point belonging to cluster i, ${d_{ij}}^{2}$ is the distance of jth data point to the center of ith cluster, m is the parameter for controlling the fuzzification, N is the total number of data points, and C is the total number of clusters. It is noticed that, the last constraints in Equation (5-2) restricts the memberships to lie on the hyperplane defined by $\sum_{i=1}^{C} u_{ij} = 1$. For the traffic modes identification, this constraint is too restrictive and lead to the memberships only represent relative numbers dependent on C rather than the real possibility of a data point belonging to clusters. Thus, Possibilistic Fuzzy C-Means (PCM) clustering algorithm was developed to improve the algorithm by releasing this constrain. The detailed introduction of PCM clustering is presented in the next section.

5.2.1.2.2 Possibilistic Fuzzy C-Means Clustering

PCM clustering was developed by releasing the constraint of $\sum_{i=1}^{C} u_{ij} = 1$ [144]. The constrains of PCM is shown in Equation (5-3),

$$u_{ij} \in [0,1] \; for \; all \; i \; and \; j,$$
$$0 < \sum_{j=1}^{N} u_{ij} \leq N \; for \; all \; i, and$$
$$\max_{i} u_{ij} > 0 \; for \; all \; j. \qquad (5-3)$$

where the memberships of each cluster are restricted by the constrain of $\max_i u_{ij} > 0$. In this case, the membership functions represent the degree of possibility of a data point belonging to clusters. The objective function of PCM is shown in Equation (5-4),

$$J_{PCM} = \sum_{i=1}^{C}\sum_{j=1}^{N} u_{ij}^{m} {d_{ij}}^{2} + \sum_{i=1}^{C} \eta_i \sum_{j=1}^{N} \left(1 - u_{ij}\right)^{m} \qquad (5-4)$$

where η_i are suitable positive numbers which determine the distance at which the values of the membership function of a data point in a cluster becomes 0.5. The selection of its value will be introduced later. The interpretation of other parameters is the same as those of the FCM clustering. The solution of the objective function can be achieved by Equation (5-5).

$$u_{ij} = \frac{1}{1 + \left(\frac{d_{ij}^2}{\eta_i}\right)^{\frac{1}{m-1}}} \qquad (5-5)$$

It is obvious that the membership function of PCM satisfies the constrains in Equation (4-9). In each iteration, the updated value of u_{ij} depends only on the distance between the center of ith cluster and jth data point. From the standpoint of possibility belonging to a cluster, the membership of a data point in a cluster should be determined solely by how far it is from the center of the cluster, and should not be coupled with its location with respect to other clusters. Then, the solution of the objective function of PCM clustering allows optimal membership solutions to lie in the entire unit hypercube rather than restricting them to the hyperplane given by $\sum_{i=1}^{C} u_{ij} = 1$. For the selection of the value of η_i, it should represent the possibility distribution for each cluster. The typical selection of η_i is shown in Equation (5-6),

$$\eta_i = K \frac{\sum_{j=1}^{N} u_{ij}^{m} d_{ij}^{2}}{\sum_{j=1}^{N} u_{ij}^{m}} \qquad (5-6)$$

where K is a constant number which is typically selected to be 1 [144]. Equation (5-6) makes η_i proportional to the average fuzzy intra-cluster distance of clusters. The initialized value of η_i depends on the initialization of u_{ij}^m and the center of clusters by randomly selection. However, random selection of the initialized parameters can generate unexpected errors to clustering accuracy [145]–[147]. In this study, a Semi-Supervised PCM clustering is proposed to mitigate the errors utilizing a small set of labelled data.

5.2.1.2.3 Semi-Supervised PCM Clustering

In Equation (5-5) and (5-6), η_i and u_{ij} are iteratively updated until the converging criteria are met. To utilizing the prior information of labelled data, the equation of η_i is designed as in Equation (5-7) for semi-supervised PCM clustering.

$$\eta_i = \frac{\sum_{x_j \in (labelled)}^{n_{labelled}} d_{ij}^2}{n_{labelled}} \qquad (5-7)$$

where $x_j \in (labelled)$ are the data points in the labelled data set and $n_{labelled}$ is the total number of data points in labelled data set. In this case, the memberships and u_{ij} and $\eta_i$can be initialized by the labelled data. The semi-supervised PCM clustering algorithm is trained by Algorithm 5-2.

Algorithm 5-2: Semi-Supervised PCM Clustering

Initialization:

The number of clusters C

The maximum number of iterations L

The fuzzification parameter m

Calculate the center of each cluster in labelled data $v_i^{(0)}$

Initialize η_i using Equation (5-7)

Initialize $U^0 \in R^{C\times N}, u_{ij}^0 = U^0(i,j) \in [0,1]$ using Equation (5-5)

Repeat:

Update U^t using (5-5), and Increment L

Until:$\|U^t - U^{t-1}\| \leq \varepsilon$ or $t \geq L$

5.2.1.3 Results Validation

To evaluate the performance of traffic speed correction based on RSSI and the final estimation of multi-modal traffic speed estimation, Mean Absolute Error (MAE), Mean Square Error (MSE) and Mean Absolute Percentage Error (MAPE) are used to compare the accuracy of corrected traffic speed and original speed. The following equations present the metrics formulation:

$$MAE = \frac{\sum_{i=1}^{N} \left| \hat{Y}_i - Y_i \right|}{N} \quad (5-8)$$

$$MSE = \frac{\sum_{i=1}^{N} \left(\hat{Y}_i - Y_i \right)^2}{N} \quad (5-9)$$

$$MAPE = \frac{1}{N} \sum_{i=1}^{N} \frac{\left| \hat{Y}_i - Y_i \right|}{Y_i} \times 100\% \quad (5-10)$$

where $\hat{Y}_i$ is the estimated average traffic speed for time-window i, Y_i is the ground truth value, and N is the number of time-windows in 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.

The performance of the proposed semi-supervised PCM clustering algorithm for identifying traffic modes is evaluated by comparing the accuracy of the selected unsupervised and semi-supervised clustering algorithms, including K-Means [148], Constrained K-Means [149], FCM clustering [114], semi-supervised FCM [150] and PCM clustering [144]. The evaluation metrics are presented in Equation (5-11),

$$Recall = \frac{TP}{(TP + FN)} \quad (5-11)$$

where TP is short for True Position which is the number of MAC trips belonging to traffic mode i with correctly assigning traffic mode i, and FN is short for False Negative which is the number of MAC trips belonging to traffic mode j with wrongly assigning traffic mode i.

5.2.2 *System Deployment and Data Collection*

Typically, sensors are installed at road intersections and get power supply from road-side cabinets. However, the metal shell of cabinets generates sever influence for Wi-Fi and BT signal communicating due to signal shielding. Thus, Wi-Fi and BT antenna should be extended to the outside of the cabinet with waterproof measures. The sensor installation is shown in Figure 5-1 (b). The Wi-Fi and BT antenna were put in a waterproof box which is attached to the side of the cabinet. The server of the prototype system is a general PC with data listening, managing, analyzing, and visualizing programs setting up.

(a)

(b)

(c)

Figure 5-1 System Deployment. (a) Customized MAC address detector, (b) Installation, and (c) Data Storage and Analysis Server

The data used in this study was collected by four Customized MAC address detector at Tongji University which locates in Shanghai City, China. The sensors' location is shown in Figure 5-2. Besides the Wi-Fi and BT MAC address data, the ground truth of traffic speed and traffic modes was also collected for the validation purpose. 4 hours ground truth of multi-modal traffic speed was collected by video camera from 16 p.m. to 18 p.m. on Jun 2nd and 3rd, 2019. The MAC address data collecting in the same time periods are used for the analysis which is shown in Table 5-2.

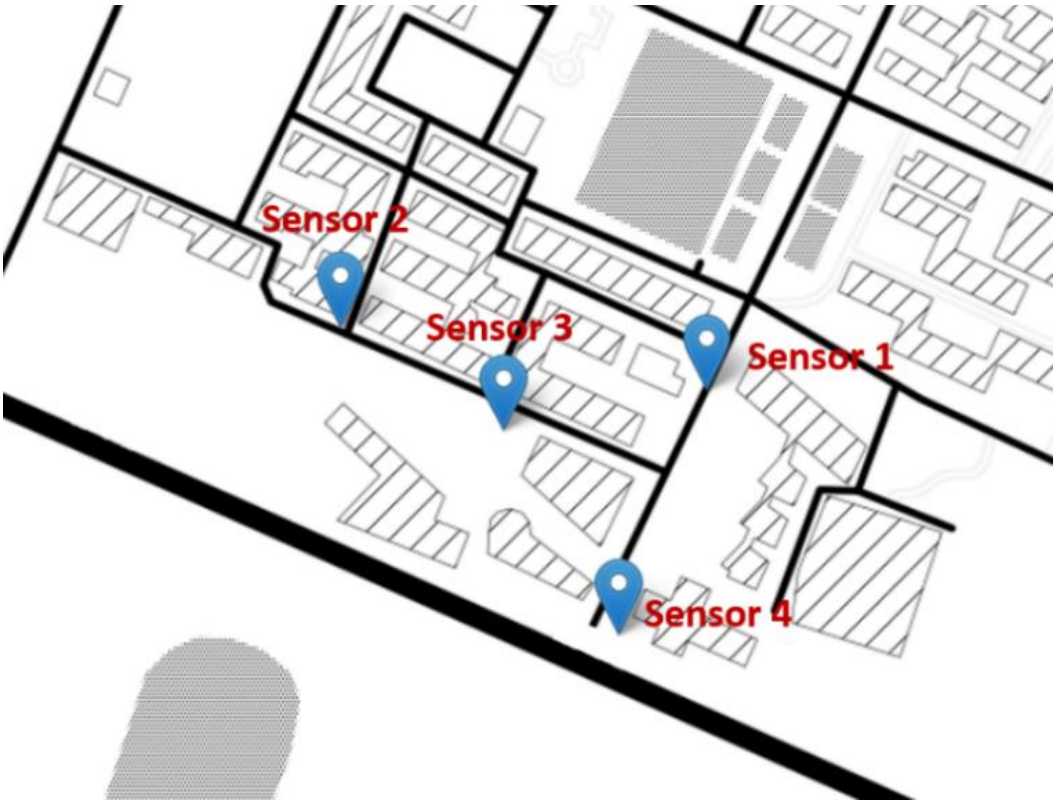


Figure 5-2 Study Site

In Table 5-2, the number of data points and the number of unique MAC address collecting by each sensor are presented. Totally, 106,926 data points and 26,596 unique MAC addresses were detected by four sensors within the 4 hours data collection period. Among them, 24,263 unique

MAC addresses were detected through Wi-Fi and 2,333 were collected via BT. The data volume of Sensor 4 is the most among four sensors and Sensor 2 obtained the least data volume.

Table 5-2 Statistics of the Data Collecting at 16 p.m. to 18 p.m. on Jun 2nd and 3rd, 2019

Sensing Points	Data Points					Unique MAC				
	1	2	3	4	Total	1	2	3	4	Total
Wi-Fi	22074	4136	6185	45820	78215	6596	1233	2193	14241	24263
BT	4771	3182	10485	10273	28711	602	395	422	914	2333
Total	26845	7318	16670	56093	106926	7198	1628	2615	15155	26596

In addition, the traffic mode labels of 408 MAC trips were collected by 10 volunteers. They traveled back and forth among four sensors by different traffic modes. At the meanwhile, the high-resolution GPS trajectories of their movements were collected by a phone app to calculate the ground truth traffic speed of each trip for evaluating the proposed algorithm of correcting traffic speed based on RSSI measurements. Table 5-3 shows the number of MAC trips with traffic mode labels for each mode.

Table 5-3 The Number of MAC Trips with Traffic Mode Labels

Traffic Modes	Car	Bike	Walk	Total
Wi-Fi	52	42	66	160
BT	70	108	70	248
Total	122	150	136	408

5.2.3 *Experimental Results*

5.2.3.1 Traffic Speed Correction Based on RSSI

To evaluate the proposed algorithm for correcting the estimated traffic speed based on RSSI measurements, the estimation accuracy of were calculated based on the estimated traffic speed before and after the correction. MSE, MAE and MAPE were utilized as the evaluation metrics. The evaluation results are presented in Table 5-4. According to the evaluation results, the proposed

algorithm for correcting traffic speed is highly effective for improving the estimation accuracy. The estimated traffic speed of all three traffic modes are closer to the ground truth after the correction based on RSSI measurements. All three metrics are decreased a lot comparing with those calculating by the original estimated speed. Among three traffic modes, walk mode achieved the most accurate estimation which the accuracy is about 95%, since the relative more detection points of each trip provides more information about the location of travelers within the detection range. The estimated speed of car mode and bike mode after correction also performed well which the accuracy is about 85%.

Table 5-4 Performance Evaluation of Correcting Traffic Speed by RSSI

Mode	Car		Bike		Walk	
	Original Speed	Corrected Speed by RSSI	Original Speed	Corrected Speed by RSSI	Original Speed	Corrected Speed by RSSI
MSE	7.1854	1.7194	1.3099	0.9208	0.0986	0.0121
MAE	2.4608	1.1672	0.8898	0.7160	0.2421	0.0863
MAPE	0.2909	0.1511	0.2005	0.1626	0.1869	0.0663

5.2.3.2 Travel Modes Identification

In this section, the evaluation result of traffic mode identification is presented. Totally, the data of 408 MAC trips were used for evaluating traffic mode identification algorithms. For semi-supervised clustering algorithms, the data was randomly split into labeled and unlabeled data by the ratio of 4:6 for algorithm training. The estimated labels of the unlabeled data were compared with true labels to calculate the evaluation metrics. For unsupervised clustering algorithms, the total data set was also randomly split into two parts with the same ratio of 4:6, and the evaluation metrics were calculated only based on the clustering results of the second part. The experiment was repeated five times. The presented result is the average performance of all five experiments.

Table 5-5 shows the evaluation result of the proposed algorithm and other baseline clustering algorithms. According to the results, the proposed semi-supervised PCM clustering algorithm outperformed all other baseline algorithms in terms of the highest accuracy of traffic mode identification for all three traffic modes. The average identification accuracy of walk mode and car mode are closer to 100%. For bike mode, there are about 17% trips were identified as car mode. The major reason is that the traffic speed of car mode is relative lower than usual condition due to the lower speed limits on the university campus. Seen from Figure 3-4 (b), the third quantile of the traffic speed of car mode is about 10 meters per second. In urban area, the average traffic speed usually is 14 meters per second or higher. Such lower traffic speed distribution of car mode made the features of car mode and bike mode similar.

Table 5-5 Performance Evaluation of Travel Mode Identification

K-Means		Predicted Walk	Predicted Bike	Predicted Car	Recall
Ground Truth	Walk	72.06%	26.47%	1.47%	72.06%
	Bike	0.00%	50.00%	50.00%	50.00%
	Car	0.00%	50.00%	50.00%	50.00%

COP K-Means		Predicted Walk	Predicted Bike	Predicted Car	Recall
Ground Truth	Walk	98.53%	1.47%	0.00%	98.53%
	Bike	0.00%	75.00%	25.00%	75.00%
	Car	0.00%	6.56%	93.44%	93.44%

FCM		Predicted Walk	Predicted Bike	Predicted Car	Recall
Ground Truth	Walk	85.29%	13.24%	1.47%	85.29%
	Bike	0.00%	39.33%	60.67%	39.33%
	Car	0.00%	1.64%	98.36%	98.36%

Semi-FCM		Predicted Walk	Predicted Bike	Predicted Car	Recall
Ground Truth	Walk	98.53%	1.47%	0.00%	98.53%
	Bike	0.00%	78.67%	21.33%	78.67%
	Car	0.00%	1.64%	98.36%	98.36%

PCM		Predicted Walk	Predicted Bike	Predicted Car	Recall
Ground Truth	Walk	85.29%	13.24%	1.47%	85.29%
	Bike	0.00%	63.62%	36.38%	63.62%
	Car	0.00%	0.82%	99.18%	99.18%

Semi-PCM		Predicted Walk	Predicted Bike	Predicted Car	Recall
Ground Truth	Walk	99.81%	0.19%	0.00%	99.81%
	Bike	0.00%	82.67%	17.33%	82.67%
	Car	0.00%	0.82%	99.18%	99.18%

For other clustering algorithms, K-Means, FCM clustering and PCM clustering performed with the accuracy in ascending order. The main reason is that the mechanism of FCM clustering

algorithm allows the features space of clusters to be partially overlapped, and PCM clustering improves membership function by reflecting the essential of “possibility”. It is noticed that all clustering algorithms trained by semi-supervised learning performed better than they were trained by unsupervised learning strategy. As it is discussed in the section of methodology, the better performance is attributed to avoiding random initialization based on the prior information in the labeled data set.

5.2.3.3 Multi-Modal Traffic Speed Estimation

After the traffic mode of each MAC trip is identified, the average traffic speed of each road segment within a pre-defined time-window can be calculated. 15 minutes was selected for evaluating the estimated multi-modal traffic speed. The 4-hour time-period from 16 p.m. to 18 p.m. on Jun 2nd and 3rd, 2019, was divided into 16 time-windows. The ground truth of multi-modal traffic speed was calculated based on the video recording. The trips in video data were extracted by manually identification. In this study, multi-modal traffic speed was estimated for the road segments between two adjacent MAC address detectors. The road segments with sensing points in the middle were not considered. Actually, if the traffic modes can be identified accurately based on the data collecting by two sensing points, adding more sensing points in the middle of road segments will definitely make the estimated accuracy higher [74]. In the 16 time-windows, not every time window had valid trips of every traffic mode for each road segment. Thus, the evaluation results are calculated only based on the time windows with valid measurements of both the ground truth and estimated multi-modal traffic speed. Figure 5-3 shows the comparison of the ground truth and the estimated multi-modal traffic speed for the time windows with 15 minutes interval. In the figures, the solid blue lines present the estimated traffic speed of three modes and the dashed red lines show the ground truth. During the 4-hour time-period, the traffic speed of

three modes were fluctuated. The estimated multi-modal traffic speed is highly close to the ground truth and reflect the flotation very well. The evaluation metrics are shown in Table 5-6.

Table 5-6 Performance Evaluation of Estimated Multi-Modal Traffic Speed

Mode	Car		Bike		Walk	
	Original Speed	Corrected Speed by RSSI	Original Speed	Corrected Speed by RSSI	Original Speed	Corrected Speed by RSSI
MSE	2.2898	1.5235	1.0612	0.6867	0.1383	0.0943
MAE	1.2768	1.0507	0.8688	0.7094	0.3258	0.2537
MAPE	0.2001	0.1594	0.1926	0.1606	0.1947	0.1519

In Table 5-6, the evaluation metrics calculated by the estimated speed before and after correcting based on RSSI measurements are presented. According to the evaluation results, traffic speed correcting based on RSSI improved the estimation accuracy a lot for all three traffic modes. The values of the evaluation metrics calculated by the corrected multi-modal speed is highly reduced comparing with those calculated by the original multi-modal traffic speed. The overall traffic speed estimation accuracy is around 85% for all three traffic modes.

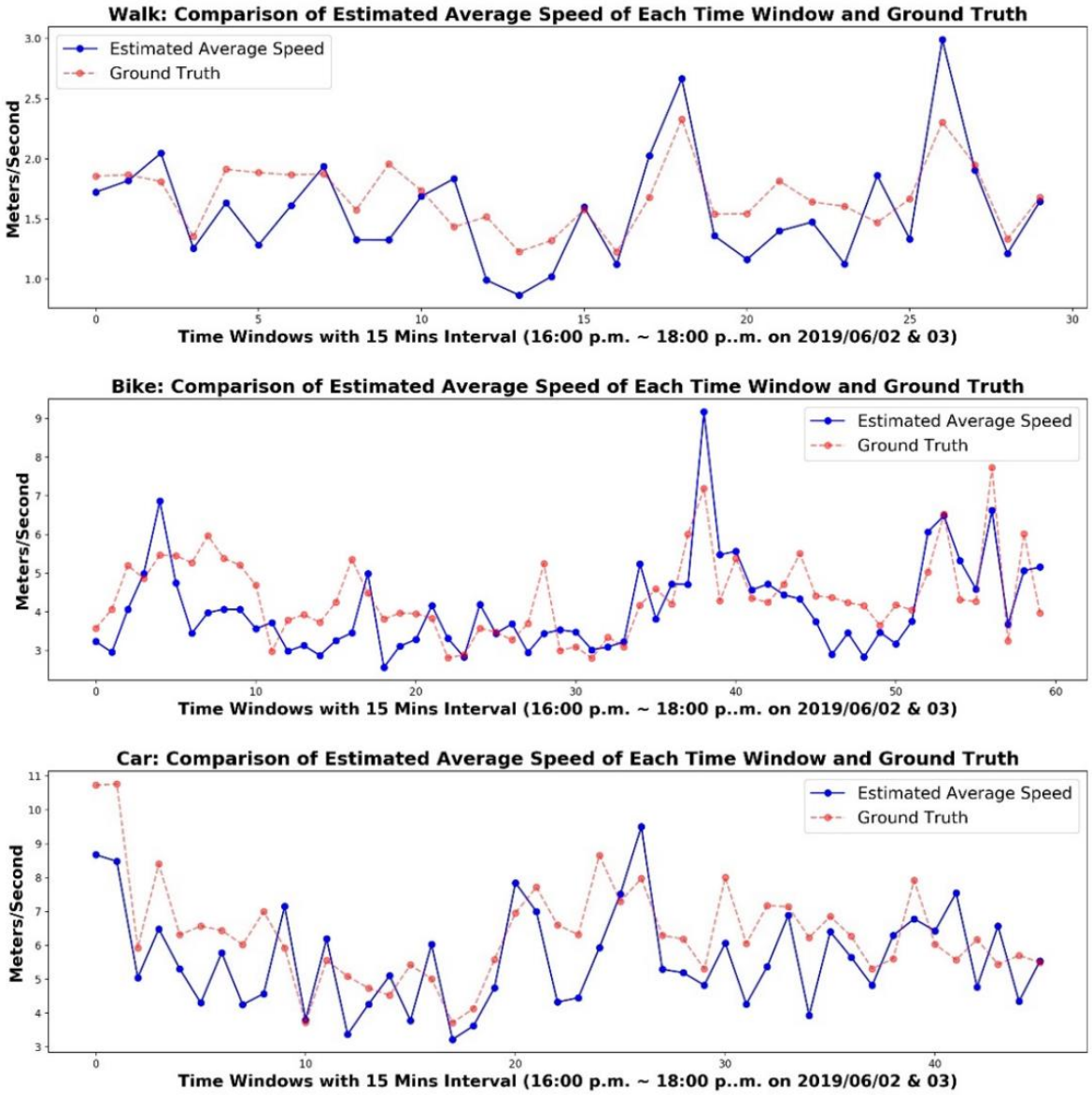


Figure 5-3 Comparison of Estimated Multi-Modal Traffic Speed and Ground Truth

It is noticed that some of time windows have no valid MAC trips were detected, therefore not every time windows have estimated traffic speed values. Basically, two factors highly influence the detection rate therefore affecting the feasibility of the implementations, those being smartphone penetration rate and MAC address randomization. From the perspective of multi-modal traffic speed monitoring based on passive Wi-Fi and BT sensing technology, Figure 5-4 can tell how much the feasibility of the proposed method for monitoring traffic speed for the time-window with

different time interval. Figure 5-4 shows the proportion of the time-windows had valid MAC trips been detected for three traffic modes. The proportion was calculated for the time-window with different pre-defined time interval. According to the figure, when the time interval is 5 minutes, only 20% time-windows have valid MAC trips detected for walk mode, and about 60% time-windows have valid MAC trips by car and bike modes. While the time interval is increasing, the proportion of time-windows with valid MAC trips gradually increased as well. When the time interval over 18 minutes, over 80% time-windows contain valid MAC trips by walk, and over 90% time-windows contain valid MAC trips by car and bike modes. Based on the results, it indicates that the proposed multi-model traffic speed monitoring system is more effective for estimating average multi-modal traffic speed for the time-window with longer time interval.

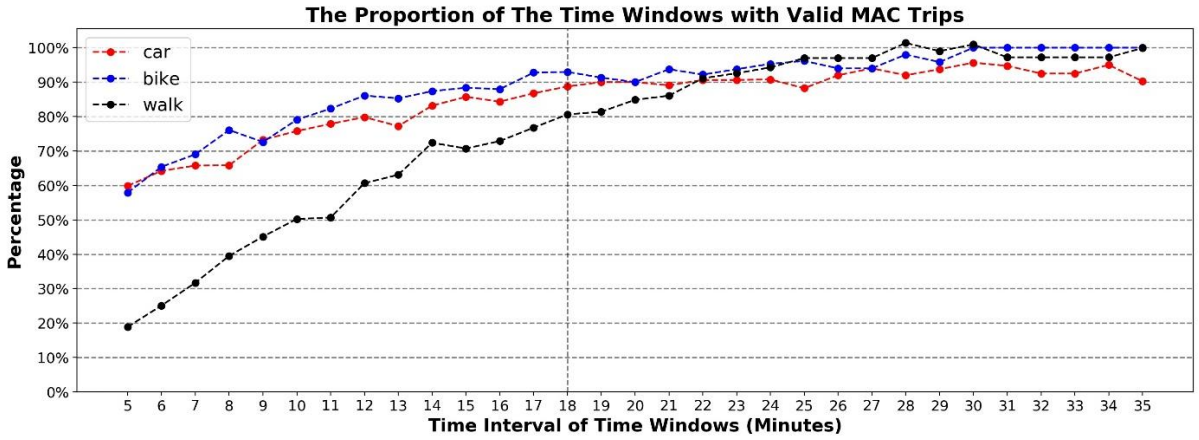


Figure 5-4 Ratio of Time Slots with Detected Traffic Status to Time Slots with GT Traffic Status and with Different Time Interval

5.2.4 *Data Visualization*

5.2.4.1 Study Area

There are fifty-five customized MAC address detectors were installed in the study site. Figure 5-5 shows the sensors' location. The visualization platform is developed based on the proposed

method in section 3.2.3. In this section, only the traffic data in walk mode is selected to visualize as an example.

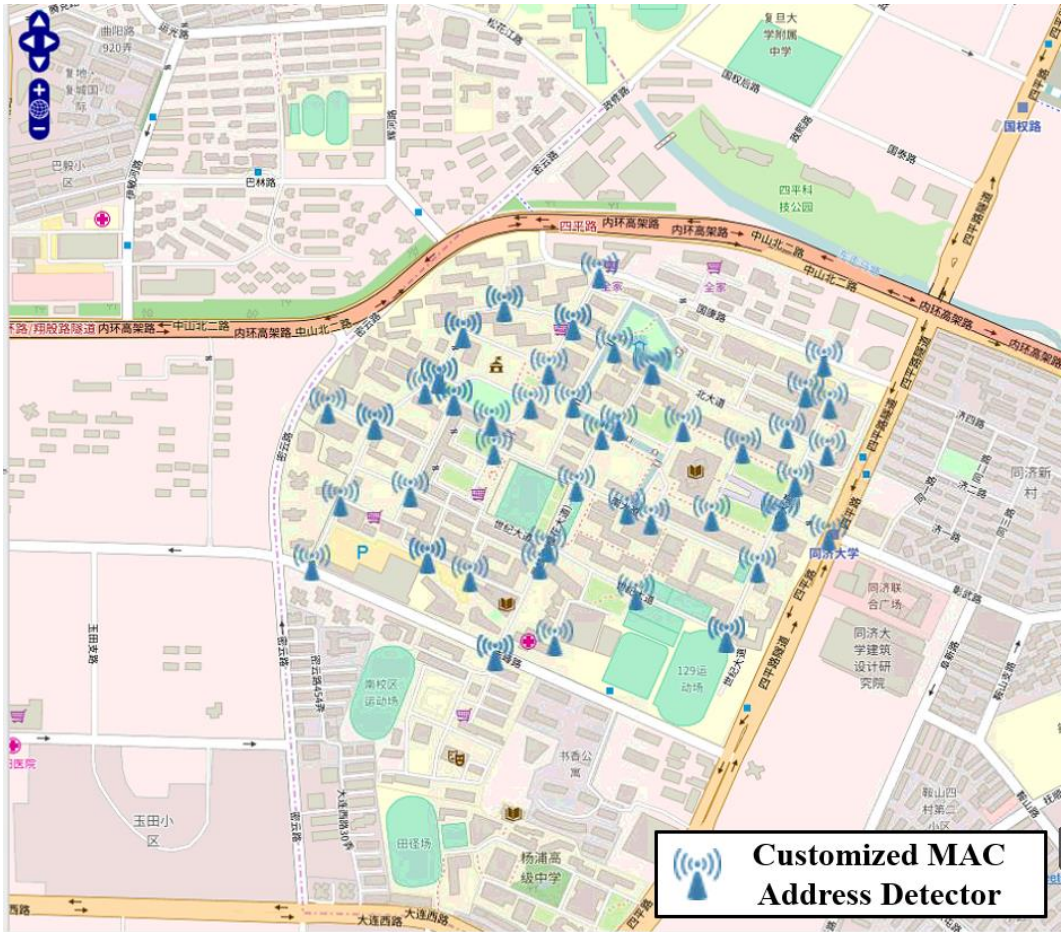


Figure 5-5 Study Area

5.2.4.2 Pedestrian Traffic Speed

The traffic speed of pedestrian is shown in this section. The traffic mode and traffic speed are calculated using the proposed method. If there is more than one path between two sensors, the shortest path algorithm was used to calculate the traffic speed. The estimation results of pedestrian traffic speed for both peak hours and non-peak hours were shown in Figure 5-6. As shown in the figure, the majority area shared the similar traffic speed for both peak hours and non-peak hours. The main reason could be the most paths were not influenced by the increased volume in peak

hours. Some differences of pedestrian traffic speed on several paths were identified by the comparison between non-peak and peak hours. The difference could be caused by pedestrian flow, as congested pedestrian flow during peak hours could result in longer travel time on the path.

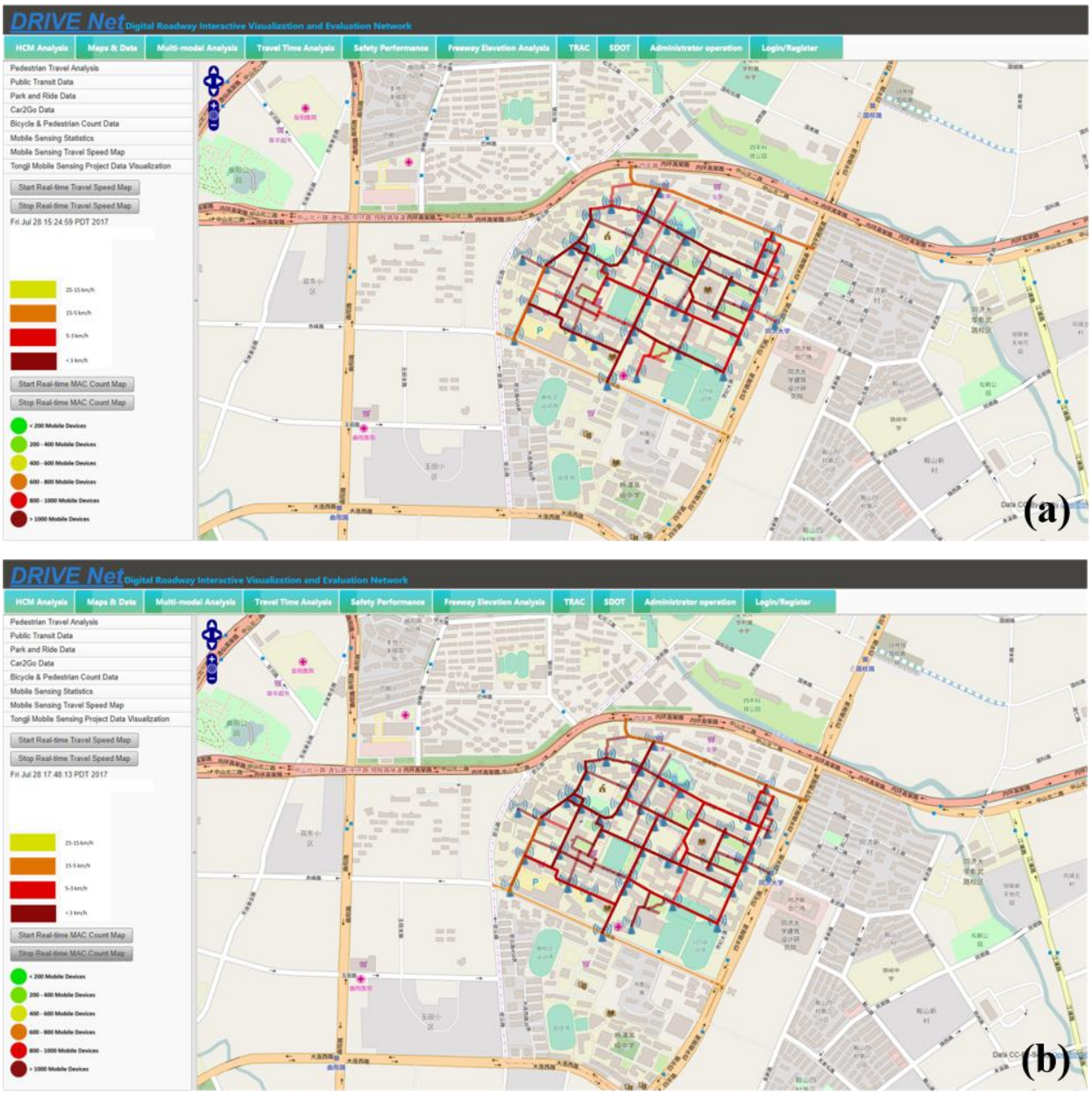


Figure 5-6 Real-Time Pedestrian Traffic Speed Estimation. (a) Non-Peak Hours, and (b) Peak Hours

5.2.4.3 Pedestrian Traffic Volume

Figure 5-7 shows the real-time pedestrian traffic volume estimation for peak hours and non-peak hours, respectively. As shown in the figure, there are several features can be summarized based on the real-time pedestrian traffic volume monitoring: (1) the traffic volume at each detection point during non-peak hours is relative small than the traffic volume in peak hours; (2) the traffic volume spatial distribution is more homogeneous during the non-peak hours than the traffic volume during the peak hours; (3) during the peak hours, traffic volume dramatically increased at the main gate of the campus, food center, teaching halls and student dormitory and the traffic volume at other area increase a little or keep stable.

The potential reason for the first feature is that students, teachers and campus staff were at work or in the class during non-peak hours, so the WBM devices they carried were either connected with access point or out of the detection range. For the second and the third features, the main reason is the student and staff had more travel demand which origin is teaching halls or destination is main gate of the campus or food center.

Even wireless sensing data is only a sample of the whole population since not all travelers have WBM device, the measurements still conformed to the reality and the true traffic volume trend. Thus, the real-time traffic volume estimation function could provide useful and valuable information for traffic monitoring and management.

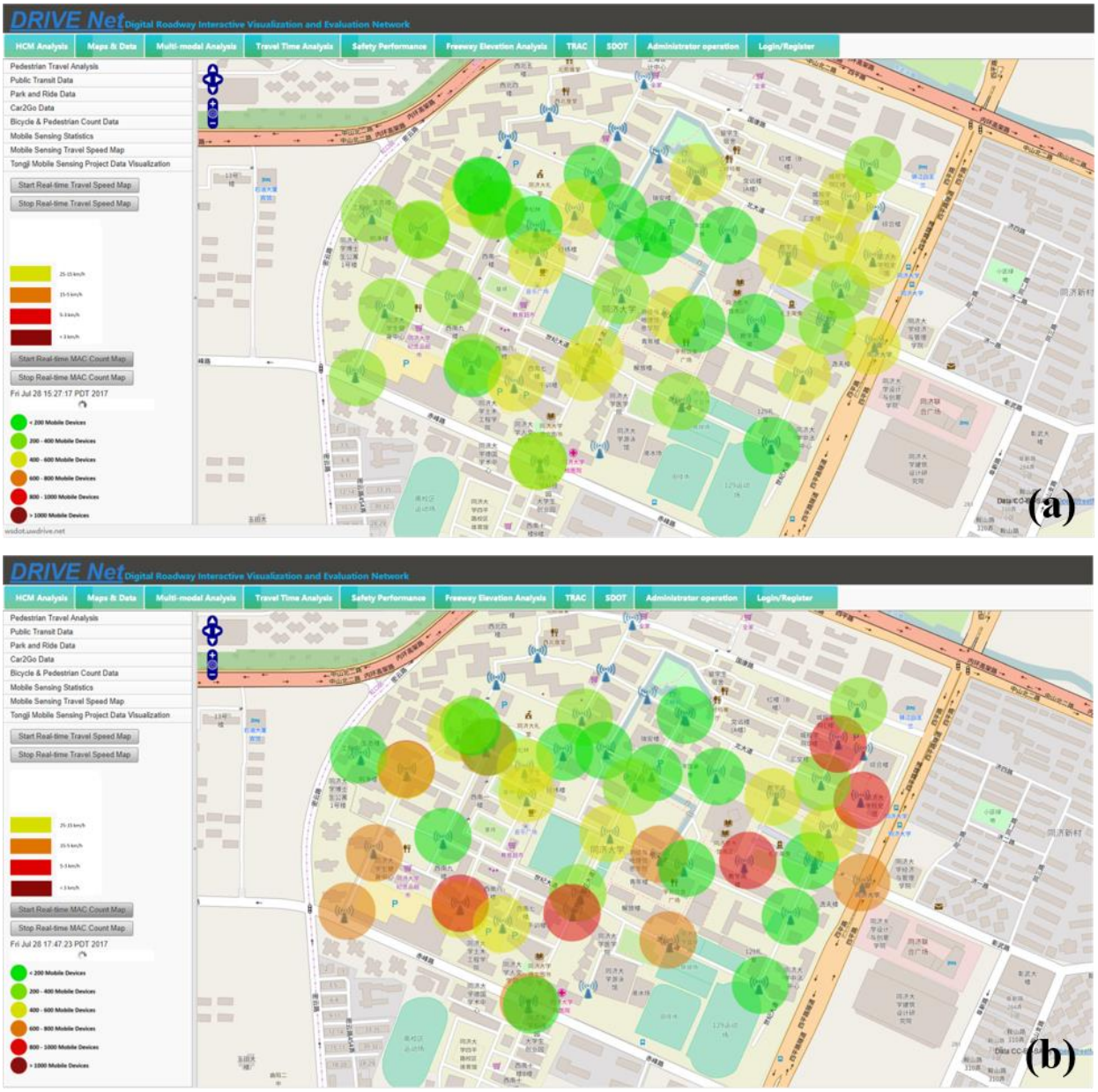


Figure 5-7 Real-Time Pedestrian Traffic Volume Estimation. (a) Non-Peak Hours, and (b) Peak Hours

5.2.5 *Concluding Remarks*

This study proposes a real-time multi-modal traffic speed monitoring system based on passive Wi-Fi and BT sensing technology. An algorithm is established to correct the estimated traffic speed based on RSSI measurements. The traffic mode of each trip is identified by the proposed semi-

supervised PCM clustering algorithm. The performance of the proposed system was evaluated based on the ground truth data. The evaluation results indicate the effectiveness and accuracy of the proposed system. By considering the impact of MAC address randomization function on MAC address detection rate, two datasets were collected in the study site and Seattle to support an analysis about randomized MAC address impacts on the feasibility of implementing the proposed system. The analysis results show that the proposed system is feasible by given the current condition of MAC address randomization implementation, and the proposed system is more effective for estimating average multi-modal traffic speed for the time-window with longer time interval. The future research direction includes urban mobility recognition based on the multi-modal traffic status estimating by Wi-Fi and BT sensing data.

Besides, based on the proposed system design, this study developed a real-time analysis and visualization platform for multi-modal traffic monitoring based on the wireless sensing data. A web-based data-driven platform for visualizing, modeling, and analyzing mobile sensing data for pedestrian flow analysis was developed in this study. This platform has ability to visualize the real-time travel time, traffic volume and trip generation and attraction by tying the geospatial data with wireless sensing data in relational database management system. In addition, we design a three tiers system for processing real-time wireless sensing data which includes data tier, data matching and computational tie and presentation tier. Finally, the analysis results were found and discussed. In the future, more endeavors should be made to enhance both the depth and width of the proposed work in this study. In this study the shortest path was used as routing rule for selecting route between two mobile sensors. However, an optimal itinerary in terms of the distance may not guarantee the truthful travel time. Thus, how to identify the most truthful route between any two mobile sensors is another research direction to improve analysis accuracy and effectiveness.

Finally, by integrating cloud computing technology into this system, the computing ability will hugely increase for more complex computational function development.

5.3 Device-Free Wireless Sensing for Pedestrian Detection

The passive Wi-Fi and BT sensing only can detect partial non-motorized travelers, which is the aforementioned population uncertainty. For some implementation scenarios, e.g., smart pedestrian signal control systems, precise pedestrian detection is critical. Thus, this section presents a proposed device-free wireless sensing technology aiming to address the population uncertainty for pedestrian detection.

5.3.1 *Proposed Methodology*

5.3.1.1 Wi-Fi CSI Signal Pre-Processing

Due to the environmental influential factors, weak signal strength, and the impact of non-LoS links, the Wi-Fi CSI base signal would have issues about generating outlier, losing packets, unsmooth waveform, and noise signal. To this end, Hampel identifier, Linear Interpolation, Kalman Filter and Wavelet Transform are deployed for addressing these issues.

5.3.1.1.1 Outlier Filtering Using Hampel Identifier

The raw CSI amplitudes and phases contain outliers that are generated by internal state transitions, e.g. transmission power and rate adaptations, and thermal noises in the devices, which can introduce variations to the base signal that is not caused by pedestrian movements. In order to remove the unexpected outlier, Hampel Identifier [151] is deployed in this research. Basically, for each value x of the base signals, the CSI base signal is computed as the median of a window which is consisted of x and m neighboring points on each side. The Median Absolute Deviation (MAD) is calculated as the standard deviation of x about its window. The value of x is replaced by the

median if the value of x is obviously different from the median by more than a predefined number of MAD. Namely, the Hampel Identifier filters the discrete values as outliers outside the interval $[\mu - \gamma * \sigma, \mu + \gamma * \sigma]$ where μ and σ depict the median and MAD respectively and γ is dependent on the application which is set to be 3 as default. In our research, the parameters are set as $\gamma = 3, m = 8$ and $MAD = 2$. Figure 5-8 presents an example of outlier filtering result. Figure 5-8 (a) shows the Wi-Fi CSI signal before filtering outliers where the black circles point out the outliers. In Figure 5-8 (b), by implementing the Hampel Identifier, the outliers are filtered without losing any main component of the Wi-Fi CSI base signal.

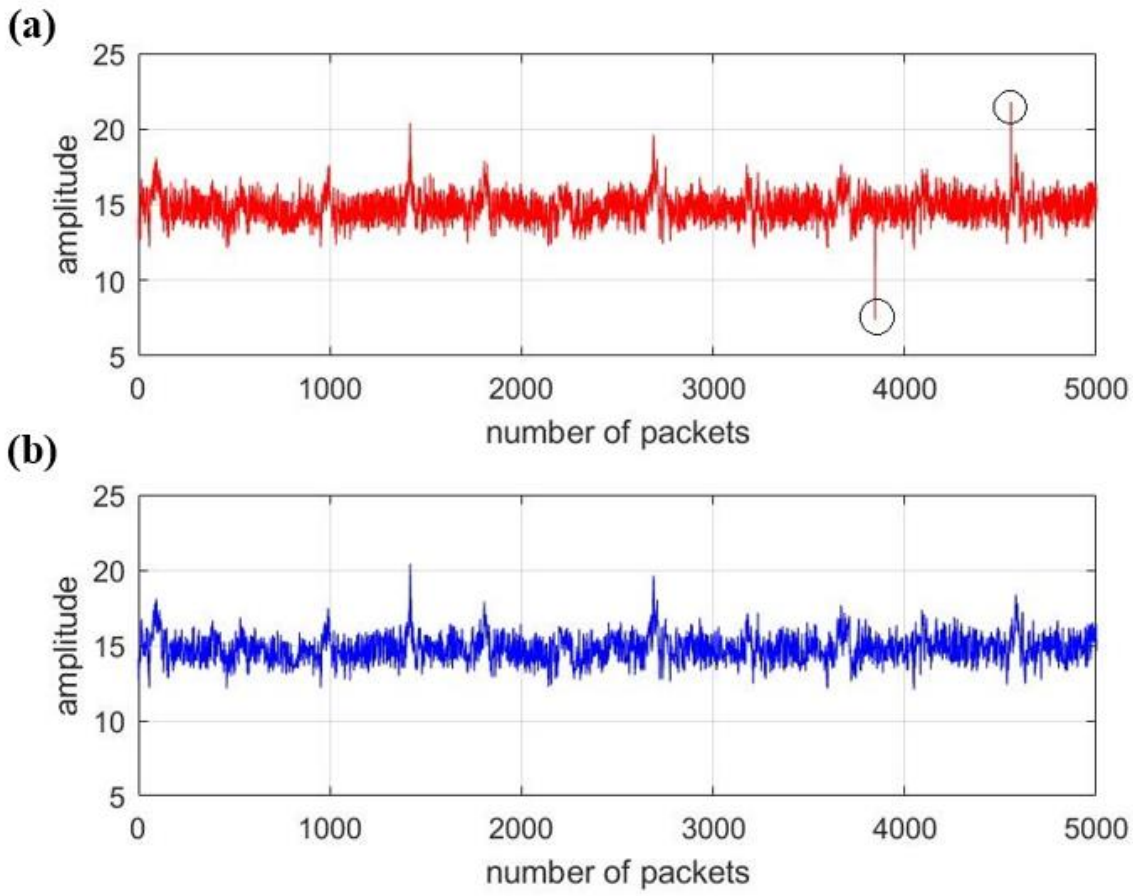


Figure 5-8 Example of Outlier Filtering Result. (a) Before Outlier Filtering. (b) After Outlier Filtering

5.3.1.1.2 Linear Signal Interpolation

For obtaining a fixed sampling ratio of Wi-Fi CSI amid packet losses due to weak signals in through-wall and non-LoS links. The losing packet result in incomplete and uneven time domain data. As the packet loss rate is random, Linear interpolation is deployed in this study for imputing the losing values. Usually, the incremental time interval caused by losing packet is integral multiples of a sampling cycle. Thus, the linear data interpolation method is conducted when the time interval is larger than a standard sampling cycle. Firstly, the time interval between each packet is measured to check whether the time interval is larger than a standard sampling cycle. Then, the data is interpolated piecewise [152]. Figure 5-9 presents an example of losing packet interpolation result. Since the packet losing rate is random, the time intervals between each adjacent packet are variates (see Figure 5-9 (a)). As shown in Figure 5-9 (c), the blue circles point out the location where the losing packet is interpolated by linear interpolation, and the Wi-Fi CSI signal in time domain has uniform time interval between each adjacent packet after interpolating losing packet. It is noticed that there is a cluster of packets were lost which are indicated by the dense blue circles in the right part of Figure 5-9 (c). The potential reason could be the electromagnetic wave generated by other device in surrounding areas, e.g. BT and Wi-Fi devices. As mentioned before, multiple factors can trigger packet to loss, which usually is random without any regularity. This is the reason why uniform interpolation is not suitable for this implementation.

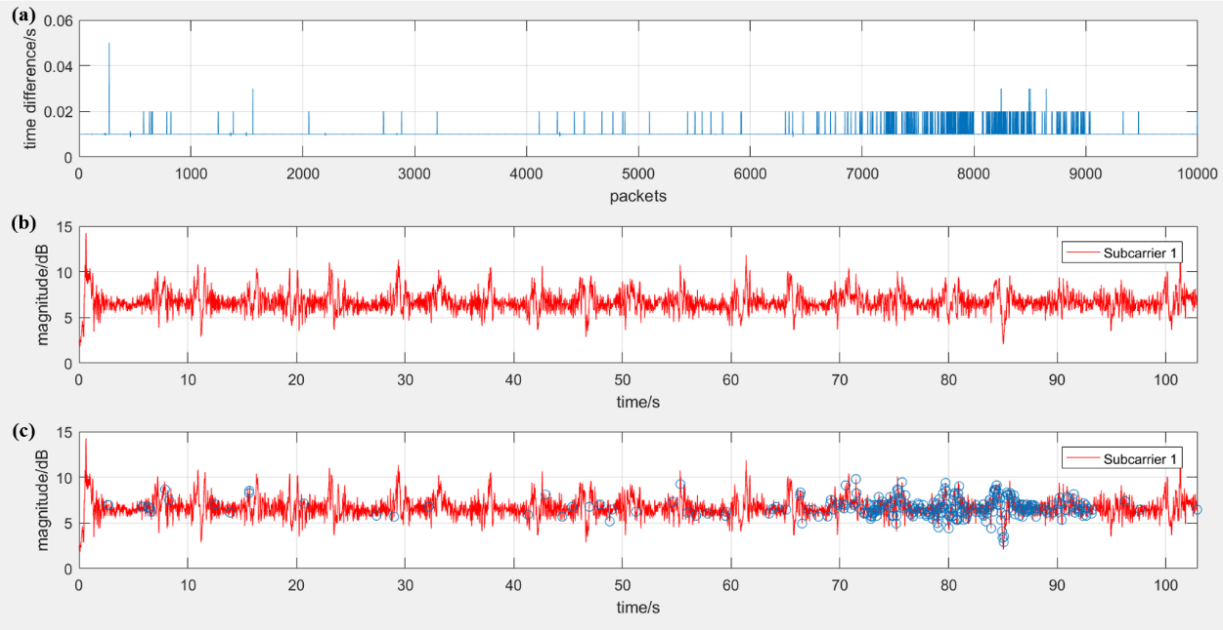


Figure 5-9 Example of Data Interpolation Result. (a) Time Interval between Each Adjacent Packet. (b) Corresponding Wi-Fi CSI Signal in Time Domain before Interpolating Losing Packets. (c) Wi-Fi CSI Signal in Time Domain after Interpolating Losing Packets.

5.3.1.1.3 Signal Smoothing Using Kalman Filter

After interpolating the losing packet, the Kalman Filter is deployed in this research for smoothing the Wi-Fi CSI base signal [153]. The Wi-Fi CSI signal contains noisy signals which are randomly produced by multiple influential factors, e.g. environmental factors, and sampling ratio. The Kalman Filter can smooth the signal by filtering the random noise out and can filter partial high-frequency component which is not helpful for sensing pedestrian movement.

5.3.1.1.4 Signal Denoising Using Wavelet Transform

Using wavelet analysis to denoise the data consist of two steps. First, discrete wavelet analysis (DWT) is used to decompose the data of each subcarrier to levels of details. Assume the sampling rate of the original data is f Hz, then, the i^{th} level of detail in the result of decomposition contains the frequency component with frequency between $f \times 2^{i-1}$Hz to $f \times 2^{i}$Hz, which is used to select

the appropriate levels of details representing the desired frequency band of the original signal. For example, if only the i^{th} level of detail of the data is used in the recomposing process, the components which have frequencies above $f \times 2^i$Hz or below $f \times 2^{i-1}$Hz would be suppressed while the $f \times 2^{i-1}$Hz to $f \times 2^i$Hz band components remains largely intact. Then, the wavelet recomposing is used to recompose the selected details. The results are the final result of data preprocessing.

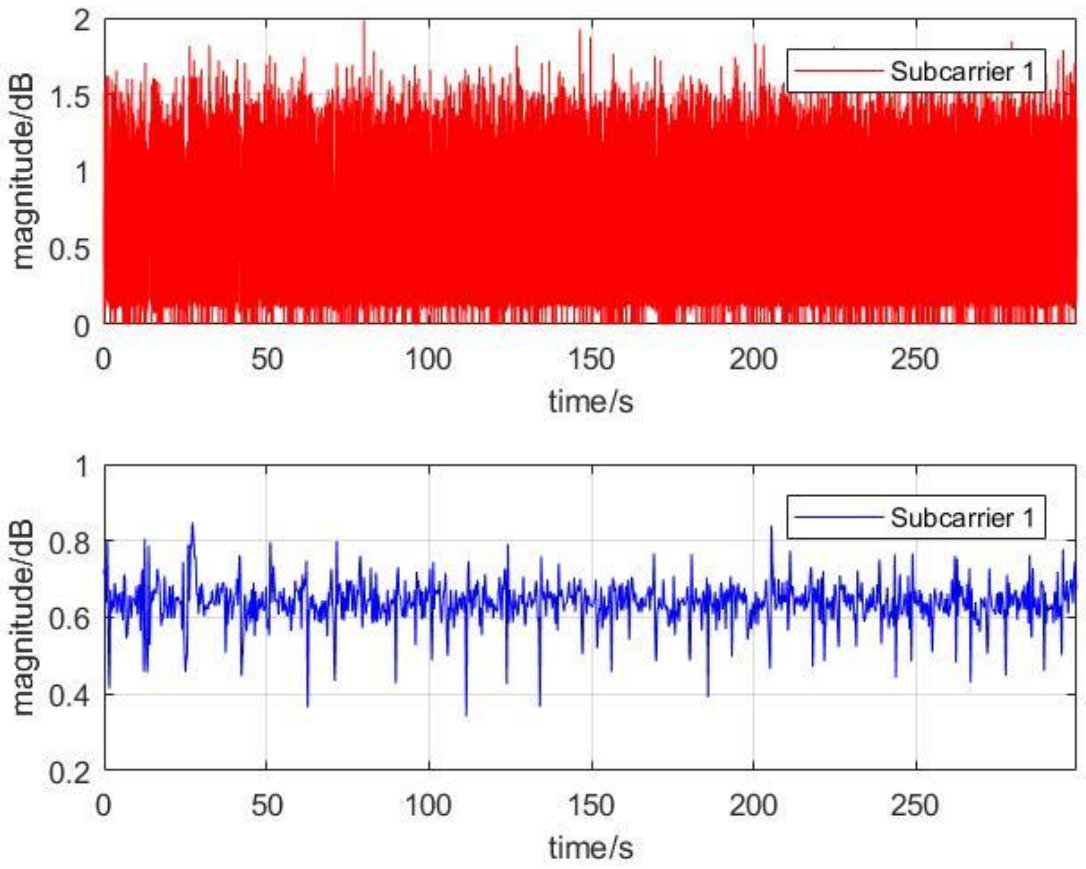


Figure 5-10 Effect of Wavelet Denoising on Time Domain

In our experiment, 2.4GHz directional antenna is utilized as the signal transmitter. Daubechies 4 wavelet [154] is employed as the mother wavelet and recomposed the signal based on the 8^{th} level of detail. The result corresponds to the combination of frequency components on 0 Hz to

3Hz. This level of detail depicted the less reflected part of the signal received by the receiver and retains the amplitude features of pedestrian existence well. It is a concise representation of the pedestrian movements that is beneficial for further detection of pedestrian existence. The following figures shows the effect of wavelet decomposition and recomposing's effect on the time and frequency domain of the original signal.

On time domain, as seen from Figure 5-10, the effect of wavelet denoising is highly desirable. The environmental noises in the original data of subcarrier 1 is successfully filtered, leaving significant features to distinguish human movements. After the wavelet denoising, the fluctuations of the amplitude of subcarrier 1 when there were no pedestrians became much lower than the fluctuations of the amplitude of subcarrier 1 when there were pedestrians. On frequency domain, as shown in Figure 5-11, components larger than 5Hz shrinked significantly while components lower than 5Hz remained intact.

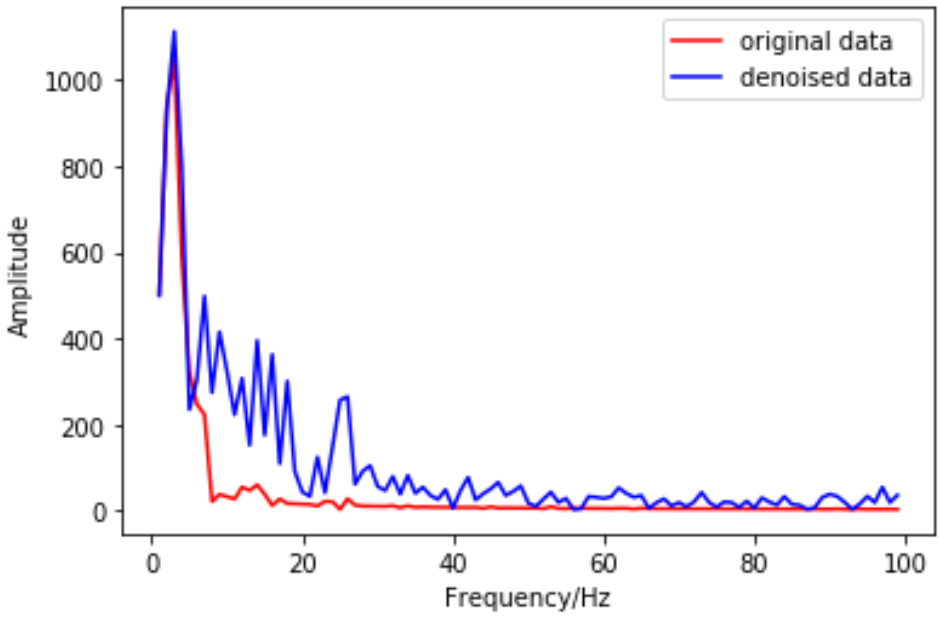


Figure 5-11 Effect of Wavelet Analysis on Frequency Domain

5.3.1.2 Pedestrian Existence Detection

After processing the Wi-Fi CSI signal of 30 subcarriers, the normalized average amplitude in time domain of 30 subcarriers is used for the pedestrian existence detection. Firstly, a parameter is calculated to represent the level of fluctuation of normalized average CSI amplitude at each timestamp which is denoted as $diff$. A time window is used for calculating the parameter. For the Wi-Fi in 2.4GHz frequency band, any object with the dimension size of about 6 centimeters can produce fluctuation. Combining the consideration of that the walking speed of pedestrian normally range in 0.3 ~ 2 meters per second, the window with 0.2 second is selected as the window size. To avoid undesired errors generation, the adjacent sliding time windows is set to have 50% overlap. Then, the $diff$ parameter is calculated by Equation (5-12).

$$diff = \max\big(x(i{:}\, i+T)\big) - \min\big(x(i{:}\, i+T)\big) \qquad (5-12)$$

where $x(i)$ is the normalized average CSI amplitude at timestamp i, $T = 0.2\ second/sampleing\ cycle$. When the sampling ratio equals 100Hz, then $T = 20$. After measuring the level of fluctuation of CSI amplitude at each timestamp, a predefined threshold is used to determine whether a pedestrian is passing or not. The predefined threshold is calculated in Equation (5-13).

$$threshold = \mu - p * \sigma \qquad (5-13)$$

where μ and σ is the mean and standard deviation of $diff$ over the experiment time periods, p is a hyper-parameter which equals 1 as default. Based on the observation, most perturbed CSI signal has higher value of $diff$, and the $diff$ is close zero for the stable CSI signal. Thus, if there are k successive $diff$s larger than the predefined threshold, we consider there is a pedestrian existed. Since the time interval between two adjacent $diff$s is 0.1 second and the time cost for passing the LoS with normal walking speed (1 meter per second) is 1 ~3 second, k is set as 5 in this research.

5.3.1.3 Pedestrian Moving Direction Recognition

There is a series of concentric ellipsoidal regions of alternating reinforced strength and weakened strength of a wave's propagation is called Fresnel zone [155]. When a reflective surface (e.g. pedestrian) is along a radio propagation path, the reflected radio waves by the surfaces may have phase difference with the CSI signal which does not impact by the reflective surface. Then the phase difference among the reflected signal and the signal directly travel to receiver can be utilized for pedestrian moving direction recognition. Basically, when pedestrian walks inward a Fresnel zone, the subcarrier with longer wavelength (lower frequency) will be perturbed first. Then, the waveforms of two subcarriers have time delay Δt caused by difference of each initial Fresnel phase ρ. The pedestrian moving direction can be identified based on the phase difference of the waveforms from two different subcarriers. For any two subcarriers of wavelength λ_1 and λ_2, we have the phase difference:

$$\Delta\rho = 2\pi(d_1 - d_0)\left(\frac{1}{\lambda_1} - \frac{1}{\lambda_2}\right) = 2\pi(d_1 - d_0)(f_1 - f_2)/c \qquad (5-14)$$

$$\Delta\rho = 2\pi(d_1 - d_0)\Delta f/c \qquad (5-15)$$

where d_1 is the reflected path and d_0 is the path for directly travelling to receiver, f_1 and f_2 are frequencies of two subcarriers and c is light speed in air.

The Fresnel direction can be identified by measuring the phase difference of two waveforms from different subcarriers. According to the previous study[156], positive Fresnel direction is defined as outwards the Fresnel zone, and negative in the opposite direction. As analyzed in previous study, any object with the dimension size of 3 centimeters can produce fluctuation of Wi-Fi CSI in Wi-Fi 5GHz frequency band. Then the CSI power roughly has (0.3, 2)/0.03 times fluctuation per second. A 0.1 second window is suitable in this case. For the Wi-Fi in 2.4GHz

frequency band, any object with the dimension size of about 6 centimeters can produce fluctuation. The sliding time window size is set as 0.2 second as well.

According to the distribution of waveforms, the time difference between two subcarriers is closer if the order number of the subcarriers is close. But if the order number of two subcarriers is adjacent, the time difference is ignorable to identify the phase difference. If two subcarriers are too far away from each other (e.g. subcarrier 1 and 20), the correlation would be rare and the phase difference between them can be more than one cycle. Thus, the phase difference is calculated using subcarrier pairs from 1 to 15 with the interval of 6 subcarriers, e.g. 1 and 7, 2 and 8, 3 and 9, and 4 and 10.

5.3.1.4 Pedestrian Speed Estimation

5.3.1.4.1 Feature Selection

To measure the speed of the pedestrian, the time duration that a pedestrian crossing the detection range need to be detected. Measuring the speed of pedestrian involves the following equation:

$$\mathrm{v} = \frac{s}{t} \qquad (5-16)$$

In the equation, s is the detection range of the Wi-Fi CSI sensor and t is the time of a pedestrian crossing the detection range. Therefore, the accuracy of the speed measurement is mostly reliant on the measurement of time t. To achieve this, the detection of pedestrian existence requires accuracy of a few milliseconds. Due to the remain of noise and unavoidable small fluctuations of the original data, the required accuracy could not be achieved by simply analyzing the unprocessed temporal features of the denoised data. Other features are required for analysis.

Correlation features:

The occurrence of a pedestrian significantly changes the relationship between the subcarriers in each 30-subcarrier group in each antenna pair. This relationship is illustrated in Figure 5-12 and Figure 5-13:

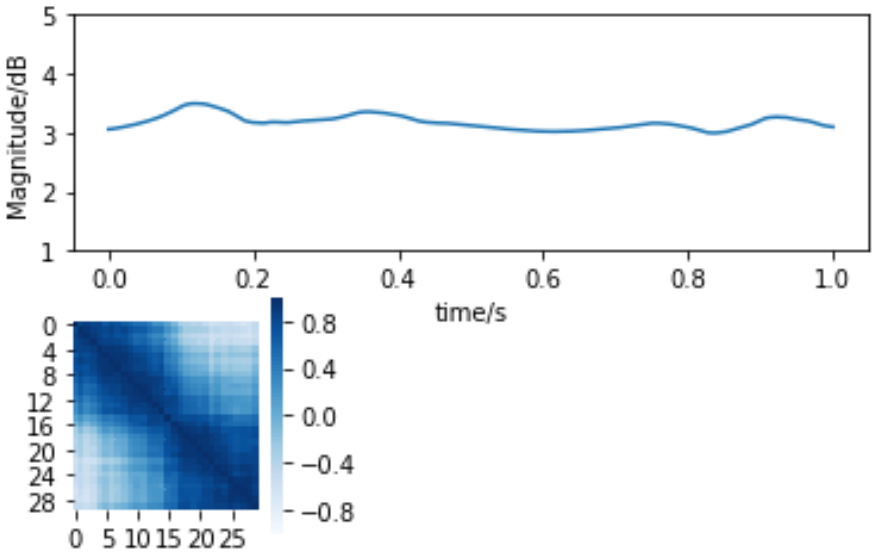


Figure 5-12 Amplitude of Subcarrier 1 and The Pearson Correlation Matrix of 30 Subcarriers of Antenna 1 in Static Environment

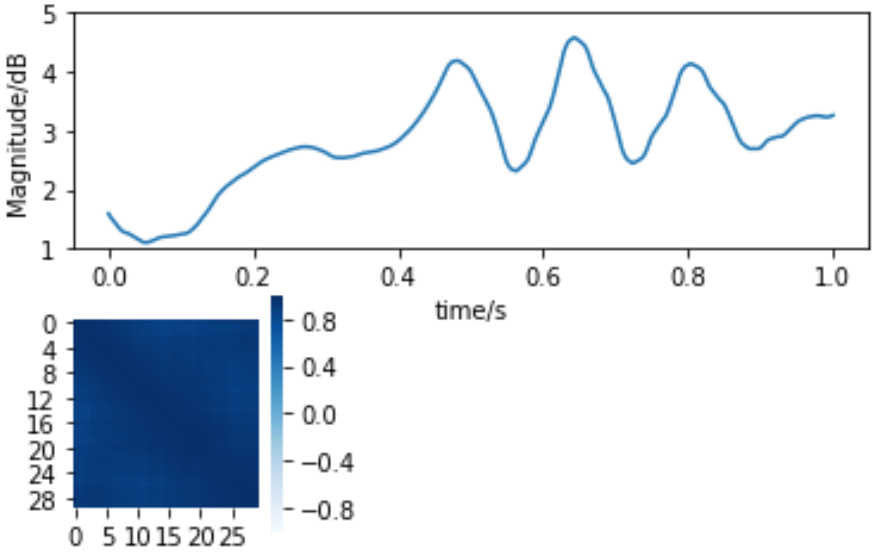


Figure 5-13 Amplitude of Subcarrier 1 and The Pearson Correlation Matrix of 30 Subcarriers of Antenna 1 During Detected Pedestrian Movement

The figures show the original amplitude data and the Pearson correlation matrix between the amplitude of subcarriers received by antenna 2 in two 1-second time-windows, one without pedestrian existence and the other with. The figure indicates that when the environment near the LOS is static without the invasion of pedestrian, the correlation between subcarriers is rather evenly distributed between -1 and 1, showing various kind and intensity of correlation. However, when a pedestrian attempt to move across the line of sight, the amplitudes of the subcarriers become highly synchronized with high positive correlation between each pair of subcarriers. Thus, correlation between subcarriers become an important feature in distinguishing static and disturbed environment.

Covariance features:

To shorten the number of features in describing the correlation, we seek a concise feature to represent the overall correlation situation of the set of subcarriers. For a matrix containing the correlation information, its eigen vectors and eigen values are good representation of itself. Another highly efficacious feature in pedestrian existence detection is the maximum eigen value of the covariance matrix of the group of 30 subcarriers in a certain antenna pair during a small temporal window. The following figure depicts the variation of the original amplitude data and the maximum eigen value of the covariance matrix in a time window of 0.125 seconds before and during the crossing of a pedestrian.

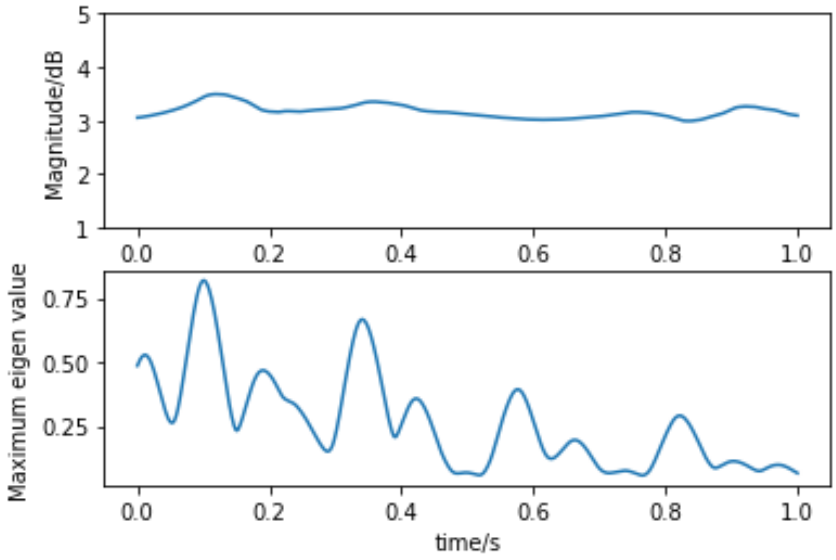


Figure 5-14 Amplitude of Subcarrier 1 and The Maximum Eigen Value of Covariance Matrix of 30 Subcarriers of Antenna 1 in Static Environment

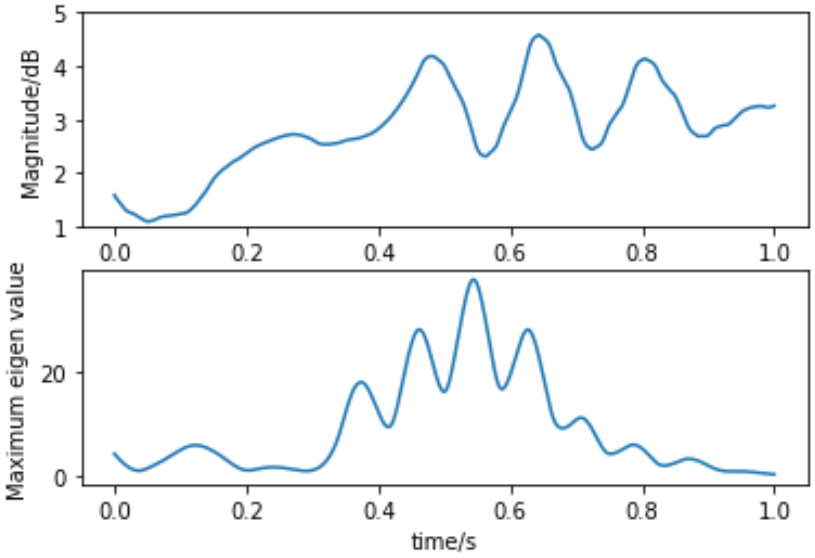


Figure 5-15 Amplitude of Subcarrier 1 and The Maximum Eigen Value of Covariance Matrix of 30 Subcarriers During Detected Pedestrian Movement

It can be observed that in Figure 5-14, the maximum eigen value during the pedestrian crossing event is significantly higher than the maximum eigen value in static environment. This indicates that the maximum eigen values of covariance matrixes successfully retain the significant

difference appeared in the overall correlation status between pedestrian existence and non-existence while reducing the shape and numbers of features. This makes it an effective and less costly feature in further analysis.

Comparing to the original amplitude data, the features regarding the correlation features are more sensitive and recognizable when a pedestrian occurs near the light of sight. Using these correlation features, we can detect the existence of pedestrians accurately.

Utilizing cycle in measurement:

According to the discussions in the previous section, a simple way to pinpoint the start and end time of pedestrian occurrence would be adopting hard threshold to the maximum eigen value time series, as the difference of the maximum eigen value of covariance matrix between static environment and pedestrian movement is very significant. In this case, it can tell that the start point of pedestrian occurrence is the point when the maximum eigen value exceeds a certain threshold. However, the threshold is highly related to the environment and could vary from time to time and situation and situation. Even for the same antenna pair, the rise of the maximum eigen value varies in different pedestrian crossing events. Figure 5-16 and Figure 5-17 depict this phenomenon.

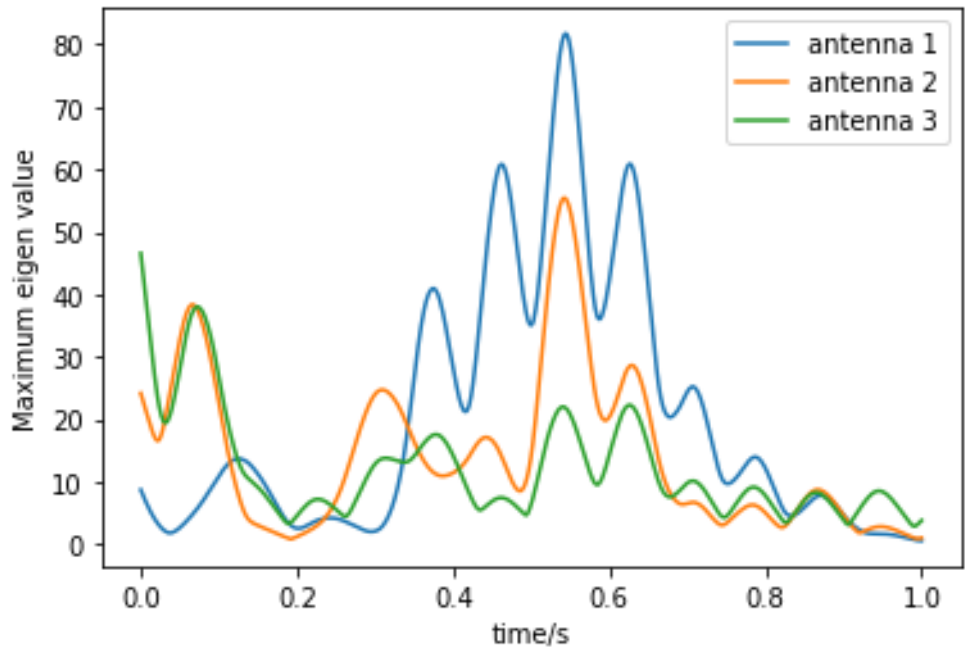


Figure 5-16 The Difference of Maximum Eigen Value of Covariance Matrix for Different Antennas in The Same Pedestrian Crossing Event

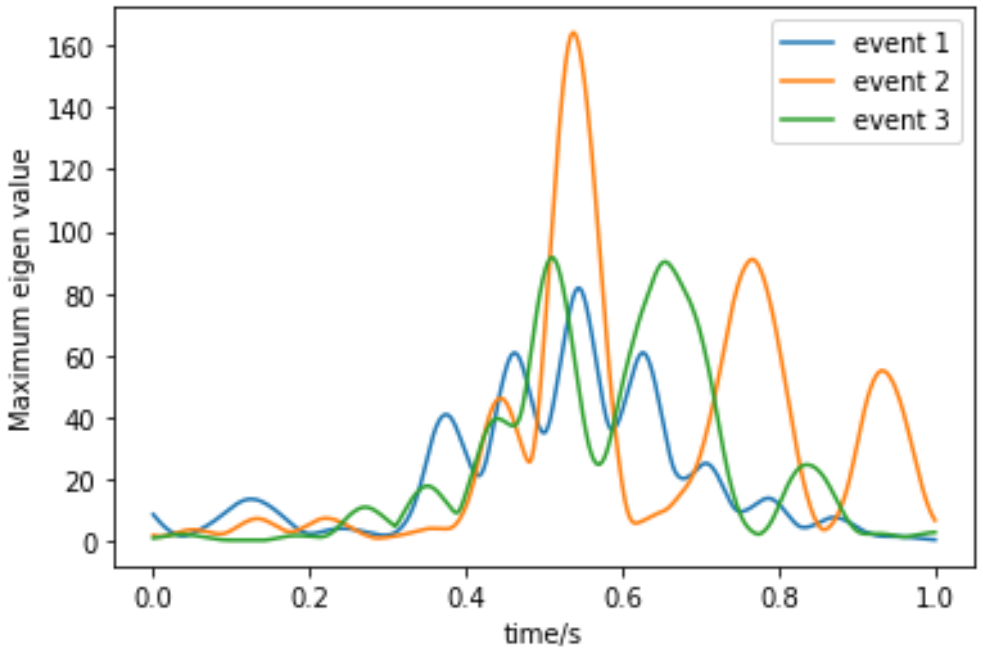


Figure 5-17 The Difference of Maximum Eigen Value of Covariance Matrix for Different Pedestrian Crossing Event of The Same Antenna, The Pedestrian Remains as The Same

Furthermore, the maximum eigen value also has high fluctuation during the movement of the pedestrian. This issue could be observed in Figure 5-15, 5-16 and 5-17. At some point inside the cycle, the maximum eigen value might grow lower than the threshold. This would hurt the continuity of the detection and possibly wrongfully recognize the movement of a single pedestrian as the movement of several independent pedestrians. To make the detection more stable and accurate, the difference and the fluctuations of the maximum eigen value is considered in an improved pinpoint method.

To further investigate the sampling mechanism, we investigated the fluctuations of the maximum eigen value of the covariance matrix in a sliding time window of 0.125 seconds. The results are illustrated in Figure 5-18 to 5-21:

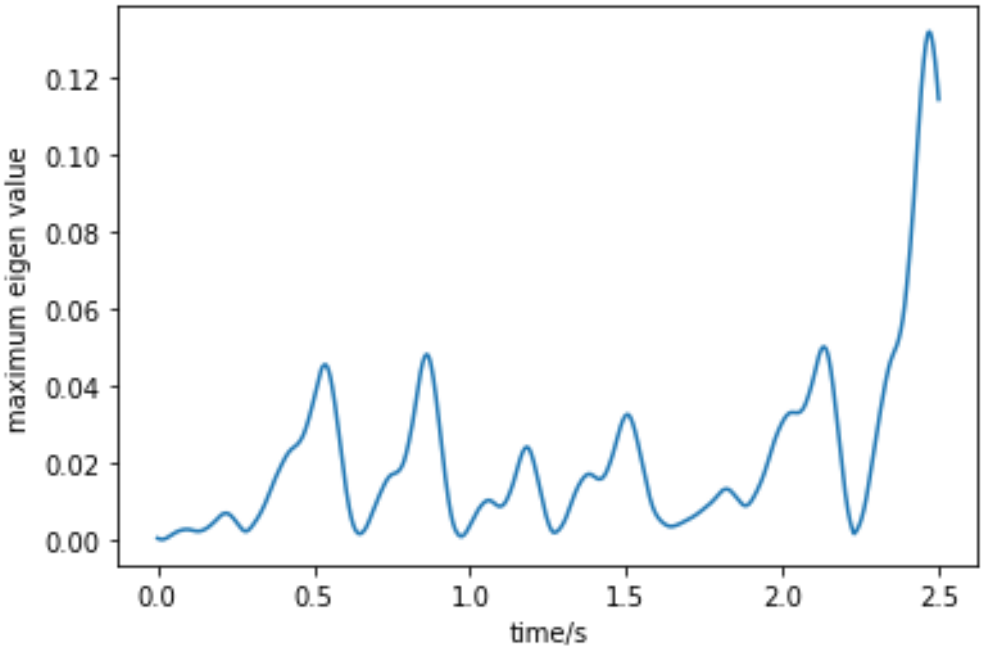


Figure 5-18 Intrinsic Cycle of The Maximum Eigen Value of Covariance Matrix of 30 Subcarriers Received by Antenna 1

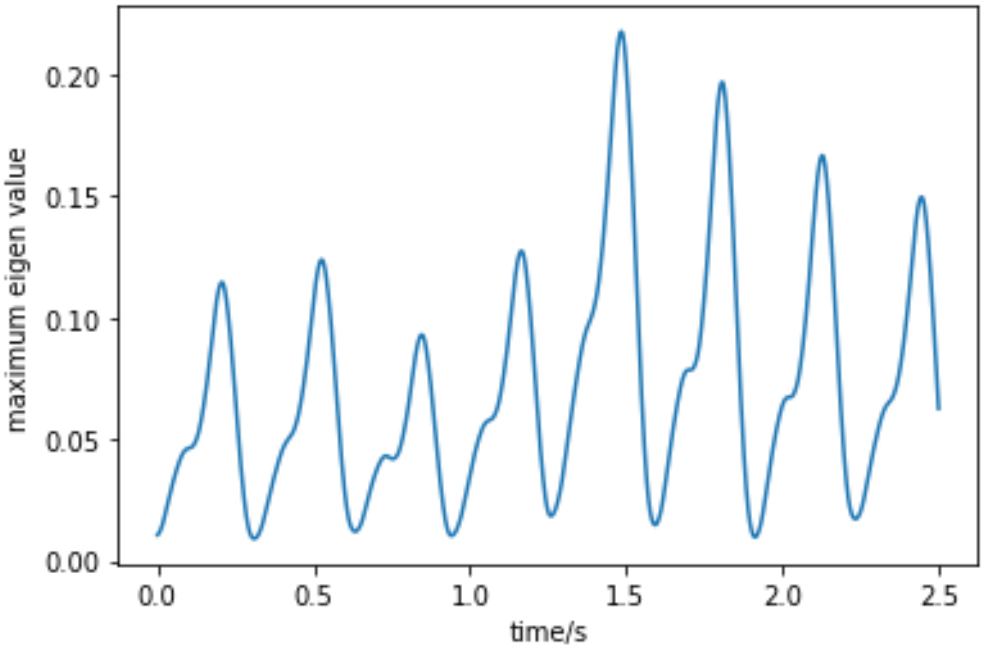


Figure 5-19 Intrinsic Cycle of The Maximum Eigen Value of Covariance Matrix of 30 Subcarriers Received by Antenna 2 in Static Environment

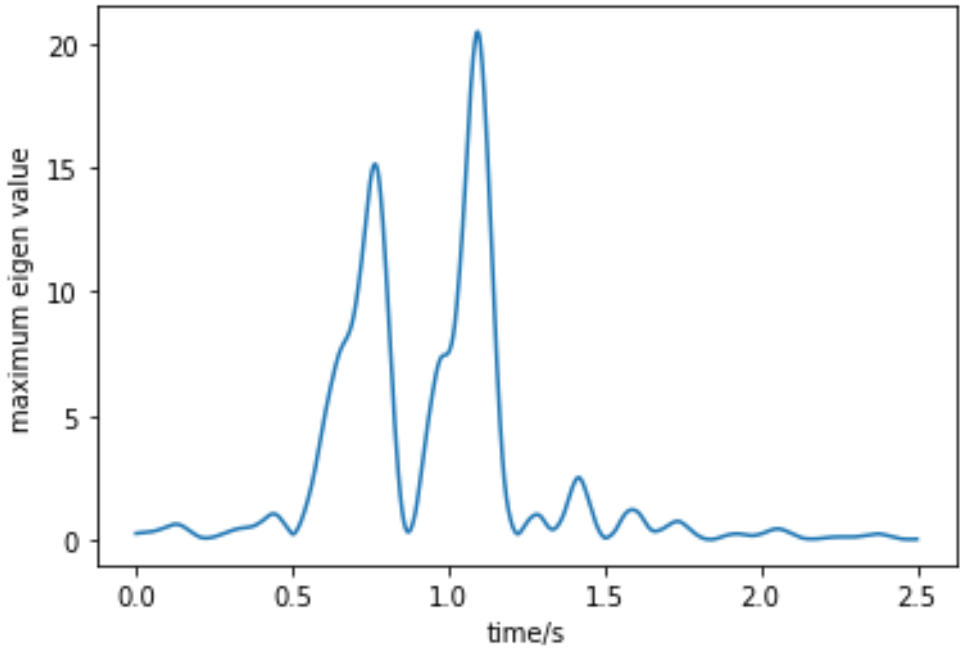


Figure 5-20 Intrinsic Cycle of The Maximum Eigen Value of Covariance Matrix of 30 Subcarriers Received by Antenna 2 Around Pedestrian Movement

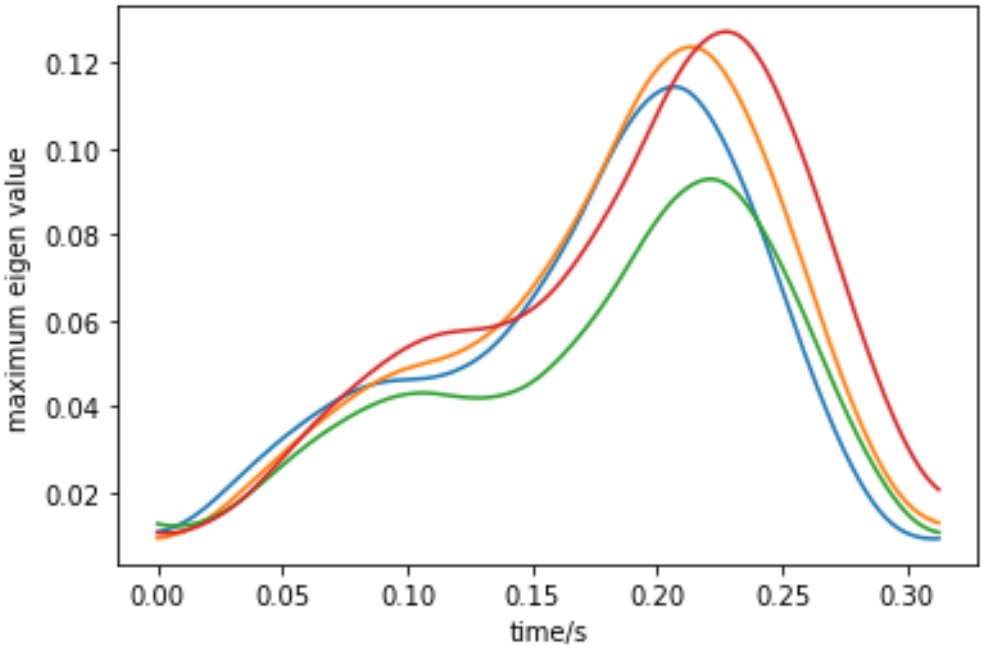


Figure 5-21 Comparison of Various Cycles of The Maximum Eigen Value of Covariance Matrix of 30 Subcarriers Received by Antenna 2

Figure 5-18 indicated that apart from the significant difference of maximum eigen value between static environment and pedestrian movement, the fluctuation of the maximum eigen value also has a significant intrinsic cycle around 0.3125s. This cycle is relatively consistent in all the antenna pairs, as shown in Figure 5-19.

The source of this intrinsic cycle is possibly an intrinsic cycle of the Wi-Fi CSI sensor from the transmission side or the receiving side. Utilizing this intrinsic cycle, the data can be smoothened and a new target of pinpointing the pedestrian crossing time emerged.

5.3.1.4.2 Speed Estimation

As it is shown in Equation 5-16, s is fixed in the calculation which is the detection range of the CSI sensor1.2 meters, the focus of velocity measurement is to accurately determine t in the detection. t is the interval between the start and end of pedestrian crossing event.

To pinpoint the start and end of pedestrian crossing event, the main target of the detection is to determine the time when the relative magnitude of the maximum eigen value in the intrinsic cycle mentioned in the previous section shifts from the low magnitude of static environment to the relative high magnitude of pedestrian interrupted environment. To detect this shift of pattern, two methods are developed for the comparison purpose.

Approach 1: Moving Average

Based on the existence of the intrinsic cycle, calculation of moving average has been adopted to smoothen the data and solve the problems caused by fluctuations of the maximum eigen value. First, the maximum eigen value time series is generated from the data of antenna 1. Denote x_i as the data point of maximum eigen value time series at the i^{th} sampling point. It is generated from the data matrix:

$$A_i = \begin{bmatrix} a_{1,i} & a_{1,i+1} & \cdots & a_{1,i+99} \\ a_{2,i} & a_{2,i+1} & \cdots & a_{2,i+99} \\ \cdots & \cdots & \cdots & \cdots \\ a_{30,i} & a_{30,i+1} & \cdots & a_{30,i+99} \end{bmatrix} \quad (5-17)$$

In matrix A_i, $a_{i,j}$ is the amplitude data of subcarrier i at sampling point j. The covariance matrix M of the 30 subcarriers is calculated at each sampling point:

$$M_i = cov(A_i) \quad (5-18)$$

Then, x_i is defined as the maximum eigen value of M_i. For a point x_i in the maximum eigen value time series, the smoothened data $x_i{}'$ is calculated by median of 250 data points $x_i, x_{i+1}, \ldots x_{i+249}$:

$$x_i' = \frac{x_i, + x_{i+1} + \cdots + x_{i+249}}{250} \quad (5-19)$$

The moving average time series $\{x_i'\}$ is generated for further analysis. In the calculation, a sliding window size of 0.125 second is adopted and the window size of the moving average is 250 data points. The effect of this smoothen method is depicted in Figure 5-22.

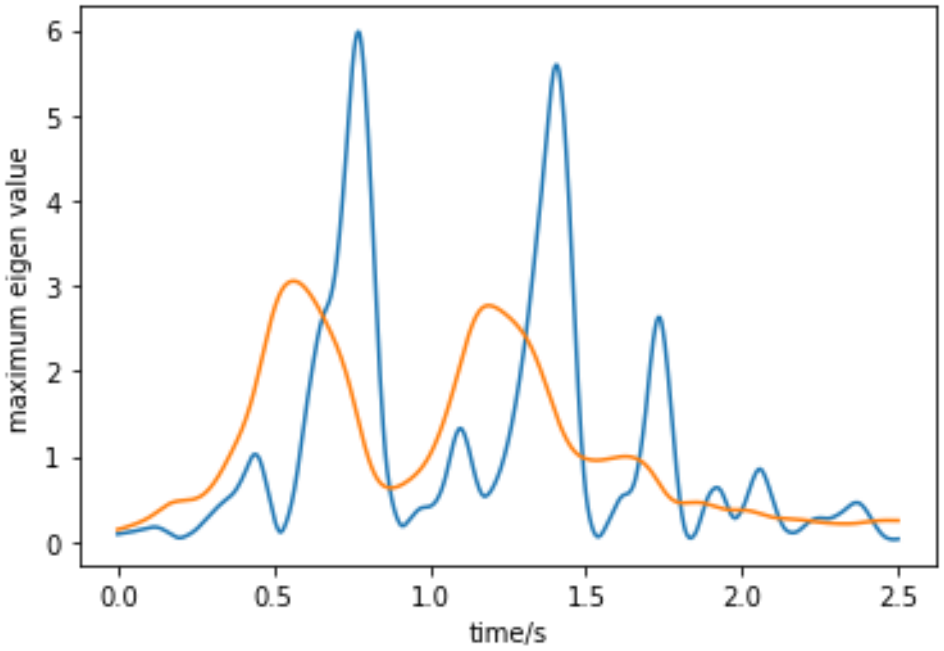


Figure 5-22 The Effect of Moving Average Smoothing

After the data have been smoothened, fluctuation of the maximum eigen value disappeared, thus enabling the application of thresholding in pinpointing the start and end time of pedestrian existence. After this, fixed threshold is adopted to determine the existence of pedestrian. The selection of the threshold is a trade-off factor of sensitivity and micro-accuracy. If the threshold is too high, the sensitivity of the detection would be too low, thus enlarging the error at the start and end. Otherwise, the sensitivity would be too high, leading to false detection. After some experiments, we found 0.9 to be the appropriate threshold in this environment. For the smoothened data x_i', if x_i' is larger than 0.9, it indicates that there is pedestrian moving near the LOS. After determining the start and end time of each pedestrian walking sample, the speed of pedestrian is calculated based on the size of detection zone and the detected time of pedestrian movement.

Approach 2: Neural Network

Approach 1 used the maximum eigen value of the covariance matrix of a group of subcarriers in a sliding time window as the feature of pedestrian movement. Essentially, this feature aims to describe the correlation between various subcarriers in the time window. However, the correlation matrix with shape of 30×30 is the original description of the correlation of the subcarriers. Compressing the 30×30 matrix to a single maximum eigen value would unavoidably cause the loss of some correlation features, thus potentially increasing the error and decreasing the accuracy. Also, as shown in Figure 5-16 and 5-17, the amplitude of the maximum eigen value varies significantly in different samples. Adopting hard threshold is not universal and has the danger of missing pedestrians. In this approach, we aim to use the whole correlation information from the correlation matrixes we obtained for each time to identify the occurrence of the pedestrian.

Assume M_i is the correlation matrix of a group of subcarriers at sampling point i. In this approach, the correlation matrix is calculated from the same original data matrix as equation 5-7, and M_i is calculated as:

$$M_i = corr(A_i) \qquad (5-20)$$

Based on the existence of the intrinsic cycle, an array of correlation matrix $M_i, M_{i+1}, \dots, M_{i+249}$ would be the full representation of the correlation feature of the intrinsic cycle time i is in. However, stacking all 250 correlation matrixes as the feature map would be costly and is hard to deploy. In the experiments, a sampling of the features in the cycle has been deployed to shorten the feature expression. 10 temporal evenly distributed point are selected in the cycle and their corresponding correlation matrixes are stacked as the feature map of time point i.

The final expression of the feature map of time point i is:

$$[M_i, M_{i+24}, M_{i+49}, \dots, M_{i+249}] \qquad (5-21)$$

The ground truth of the first one third of the 42 normal pedestrian walking event and the non-walking intervals between then have been manually labeled as training data. ResNet 2.0 [157] has been used as the training network. The first one third of the samples are used as training data and others are used for verification.

5.3.1.5 Results Validation

For validating the detection results, two evaluation metrics are deployed for measuring the level of accuracy of the detection results which are Accuracy and False Positive (FP) Rate. They are calculated using the following equations,

$$Accuracy = \frac{Y_{total} - Y_{miss}}{Y_{total}} \times 100\% \quad (5-22)$$

$$False\ Positive\ Rate = \frac{T_{FP}}{T_{FP} + T_{TN}} \times 100\% \quad (5-23)$$

where Y_{total} is the total number of times of pedestrian passing, Y_{miss} is the total number of times of missing detection, T_{FP} is the total time duration of detecting pedestrian existence when there is actually no pedestrian exists, T_{TN} is the total time duration of detecting pedestrian presence when there is actually any pedestrian exists, and $T_{FP} + T_{TN}$ is the total time duration of no pedestrian existence during the experiment. Since the window size is set as 0.2 second, when there is no pedestrian exists, the time duration of detecting pedestrian existence is calculated by multiplying the number of detections of pedestrian passing by the window size.

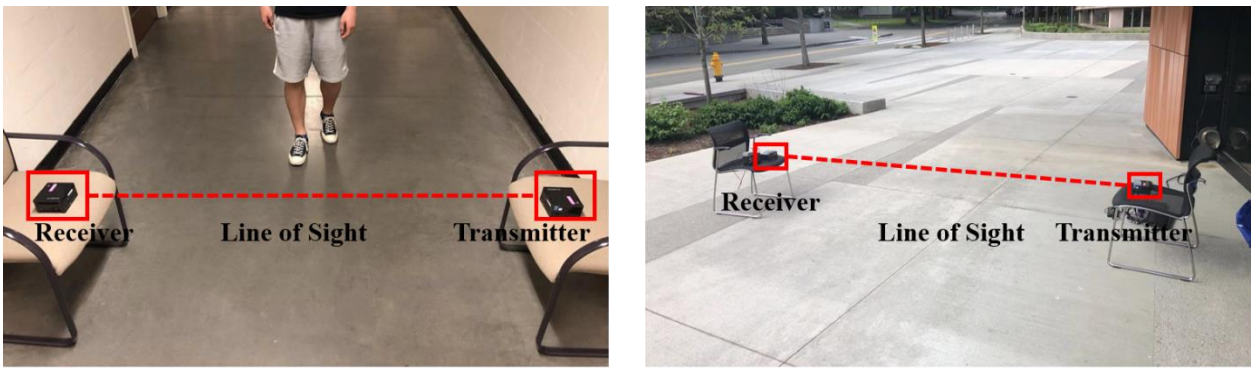
5.3.2 *Experimental Design*

In this paper, the pedestrian existence and moving direction will be sensed based on Wi-Fi CSI in both indoor and outdoor environments. The Wi-Fi CSI data is collected by an open-source software tool that runs on the GIGABYTE 4570R computer with the Intel Wi-Fi Link 5300 wireless NIC

chipset [158]. The Wi-Fi CSI sensor can capture CSI for 30 groups of subcarriers, spread evenly among the 56 subcarriers of a 20 MHz channel or the 114 carriers in a 40 MHz channel. Three 4DBi omnidirectional antennas are connected with the Wi-Fi chipset of receiver and transmitter respectively. The detailed information about the study sites and experimental design are introduced in the following sections.

5.3.2.1 Study Site

As shown in Figure 5-23, two study sites were selected for conducting the designed experiments in both indoor and outdoor environments. In both environments, the Wi-Fi packet receiver and transmitter were put on chairs with about 50 centimeters from the ground, and the distance between the receiver and the transmitter is set as 3.5 meters. The transmitter keeps sending Wi-Fi packets with a specific sampling ratio during the experiments, and the receiver was set as a listener mode. Multiple volunteers walk through the Line of Sight (LoS) back and forth. In addition, there is a video camera recording the volunteers' movements for validating the results of Wi-Fi CSI detection. Detailed information about data collection is introduced in the next section.



(a) **(b)**

Figure 5-23 Study Sites. (a) Indoor Environment, and (b) Outdoor Environment

5.3.2.2 Data Collection

The data collection work was conducted in both indoor and outdoor study sites. Totally, we have 4 volunteers for collecting data. All volunteers are male with similar height and weight. They walked through LoS one person at a time with the normal walking speed of about 1 meter per second. We count passing the LoS once is one time of pedestrian existence. The total data collection duration is one hour for each site. In addition, as claimed in previous studies, the sampling ratio is an influential factor for the detection accuracy [159]. To explore the influence of sampling ratio in this application scenario, we repeat the data collection work in three different sampling ratios which are 100Hz, 500Hz, and 800Hz. The description of the dataset is presented in Table 5-7.

Table 5-7 Description of Datasets

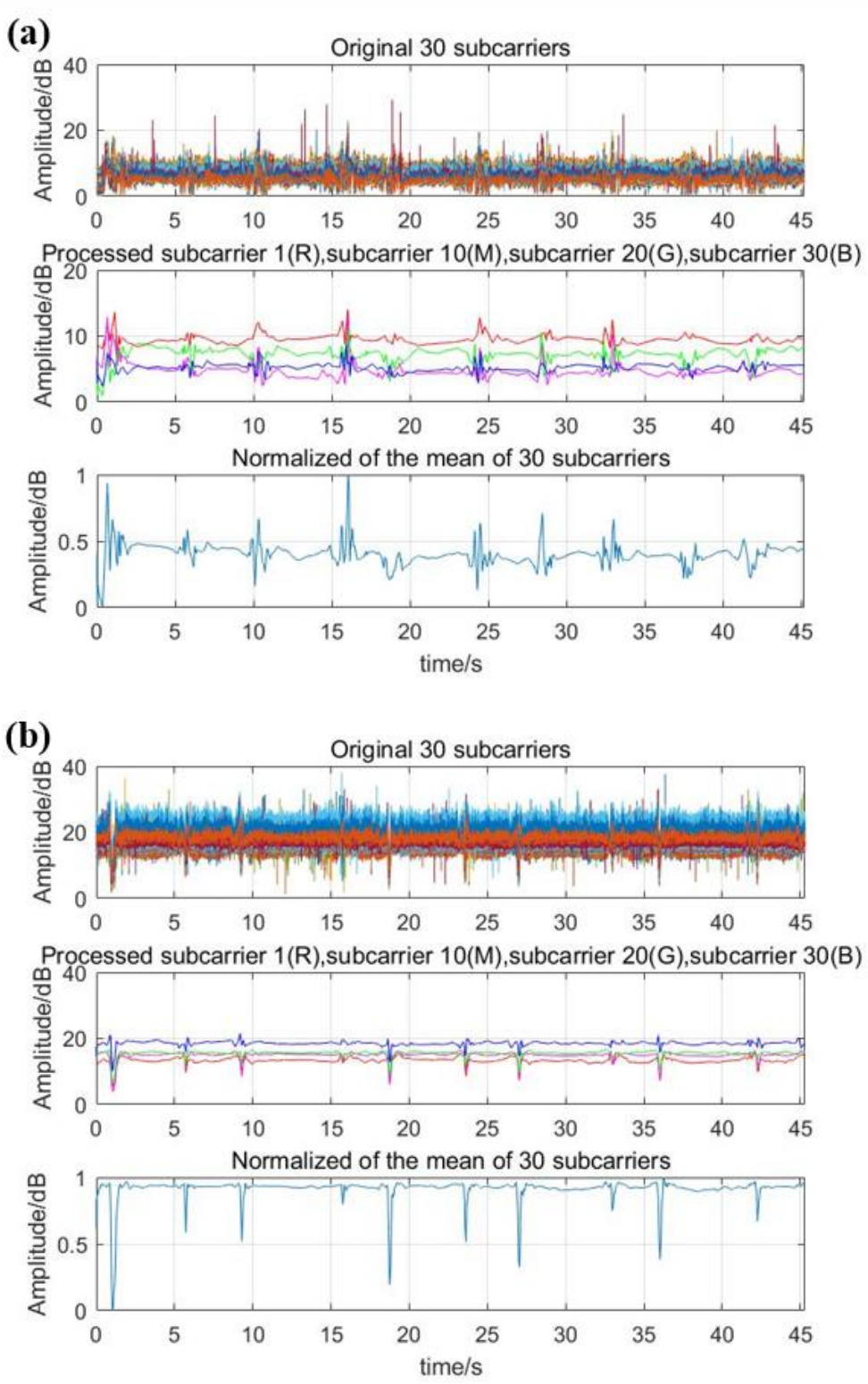
Sampling Ratio	Environment	# Pedestrian	Direction 1	Direction 2
100Hz	Indoor	130	65	65
	Outdoor	154	77	77
500Hz	Indoor	49	24	25
	Outdoor	48	24	24
800Hz	Indoor	24	12	12
	Outdoor	27	14	13

5.3.3 *Numerical Results*

5.3.3.1 Wi-Fi CSI Signal Pre-Processing

Figure 5-24 shows the Wi-Fi CSI signal pre-processing results using the CSI data collected by three sampling ratios in the indoor environment. The presented results were randomly selected which cover 50 seconds of the experimental time periods for all three sampling ratios. The subplots in Figure 5-24, from the top to the bottom, the original Wi-Fi CSI signal of 30 subcarriers, the waveforms of subcarriers 1, 10, 20 and 03 after signal pre-processing, and the normalized average amplitude of 30 subcarriers are presented respectively. For the base signal of original subcarriers,

since it contains lots of random noise and the signal is in higher frequency, the fluctuation caused by pedestrian existence cannot be observed directly. In addition, due to the higher frequency in the signal collected in 500Hz and 800Hz sampling ratio, the signal trend has more fluctuation which is uniformly distributed in the time domain than the signal collected in100Hz. After signal pre-processing, the waveforms of each subcarrier are smooth (see the mid-subplots of Figure 5-24 (a), (b) and (c)). For the CSI signal collected in100Hz, the perturbation caused by pedestrian existence is obvious, and all subcarriers share a similar fluctuation trend. However, for the CSI signal collected with 500Hz, on account of the signal perturbation caused by walking movement is around 40Hz [156], the signal in 500Hz would be also sensitive to the pedestrian gait and gesture. Thus, the signal waveforms of the CSI signal collected in 500Hz has more minor fluctuations (see the mid-subplots in Figure 5-24 (b)). For the waveform of CSI signal collected with 800Hz, since it can be more sensitive to the influence of the electromagnetic wave generated by other devices in surrounding areas. Thus, the waveform of the CSI signal collected in 800Hz is the most disorderly (see Figure 5-24 (c)). Finally, the normalized average amplitude of all 30 subcarriers in the time domain is shown in the bottom-subplots of Figure 5-24. It is obvious that the fluctuation caused by pedestrian existence can be observed in the data collected in 100Hz (bottom-subplots in Figure 5-24 (a)). However, since the CSI signal collected with 500Hz and 800Hz can capture the movements of other objects or the influence of the electromagnetic wave generated by other devices in the building, the waveforms contain plenty of disorderly fluctuations.


(a)
Original 30 subcarriers
Amplitude/dB
Processed subcarrier 1(R),subcarrier 10(M),subcarrier 20(G),subcarrier 30(B)
Amplitude/dB
Normalized of the mean of 30 subcarriers
Amplitude/dB
time/s
(b)
Original 30 subcarriers
Amplitude/dB
Processed subcarrier 1(R),subcarrier 10(M),subcarrier 20(G),subcarrier 30(B)
Amplitude/dB
Normalized of the mean of 30 subcarriers
Amplitude/dB
time/s

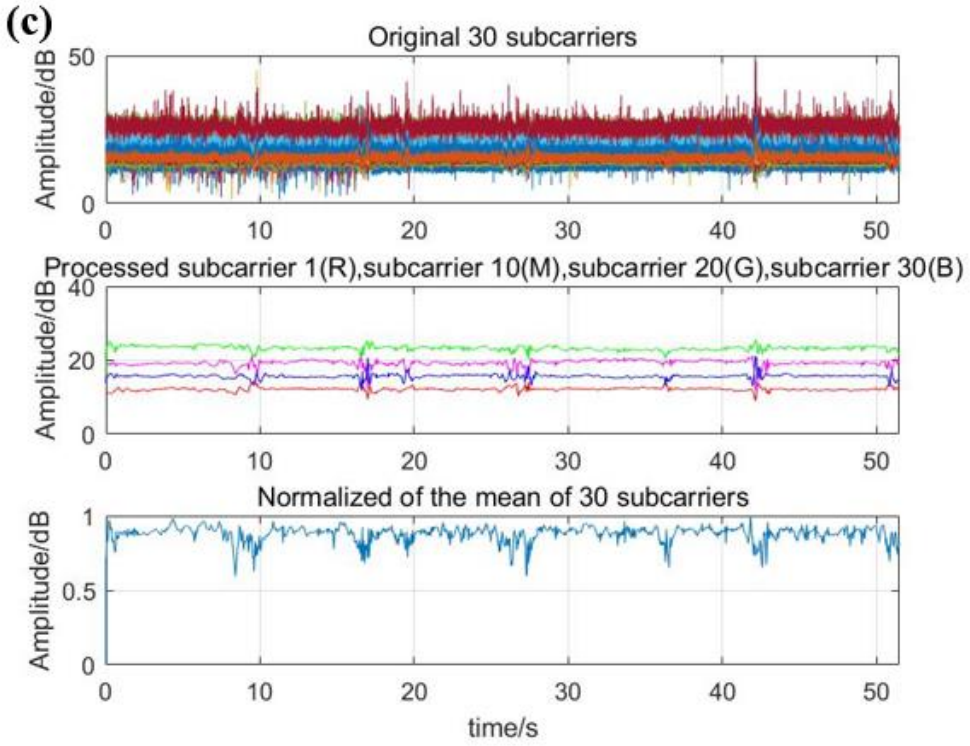
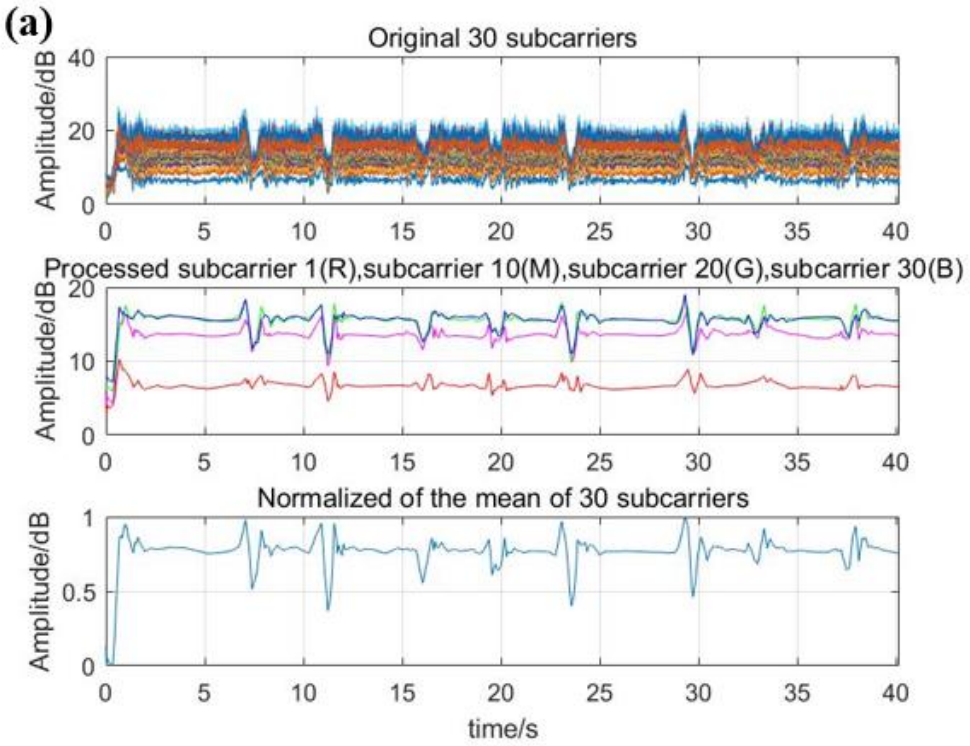


Figure 5-24 Wi-Fi Signal Pre-Processing Results in Indoor Environment. (a) 100Hz Sampling Ratio, (b) 500Hz Sampling Ratio, (c) 800Hz Sampling Ratio.



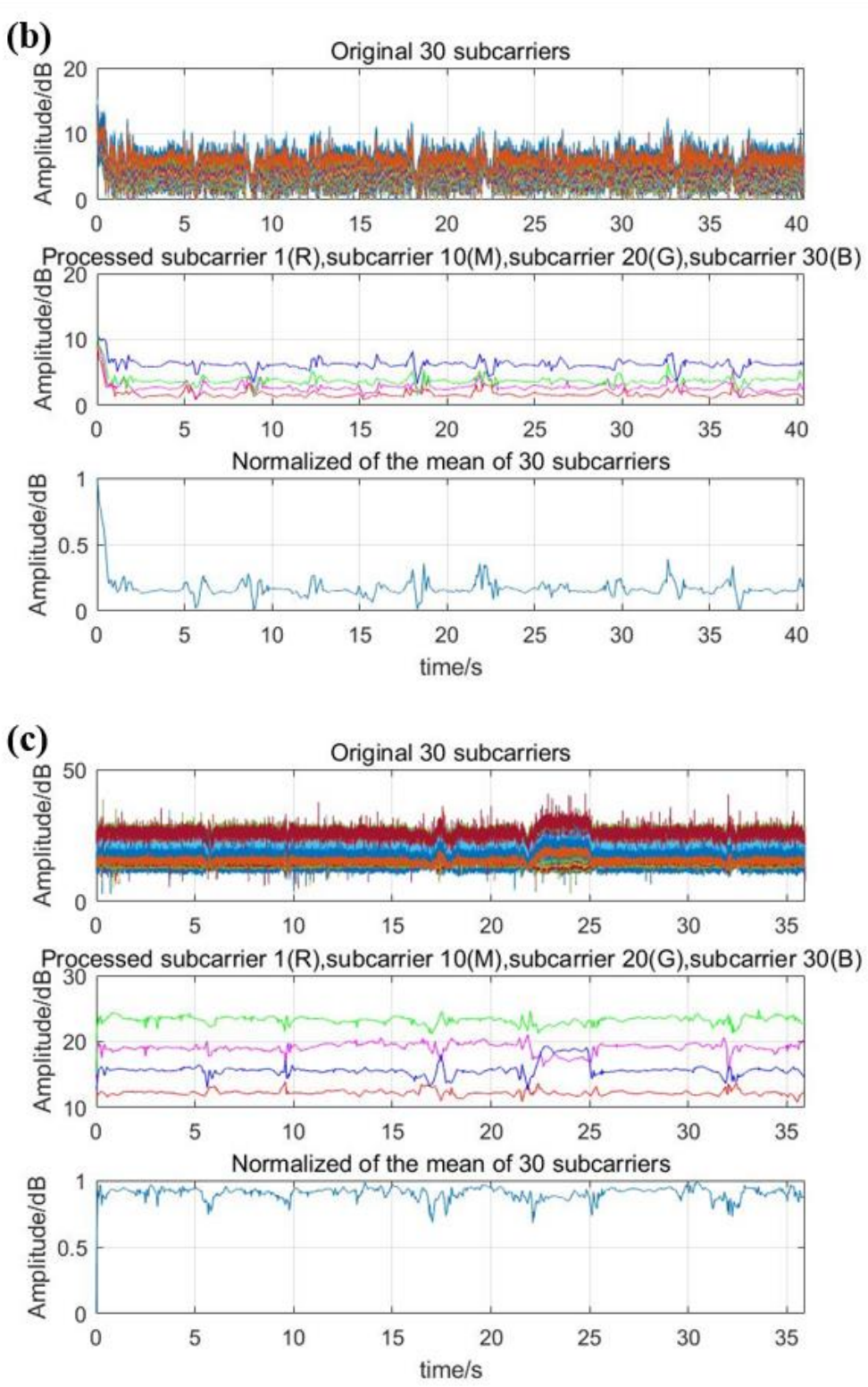


Figure 5-25 Wi-Fi Signal Pre-Processing Results in Outdoor Environment. (a) 100Hz Sampling Ratio, (b) 500Hz Sampling Ratio, (c) 800Hz Sampling Ratio.

Figure 5-25 shows the signal pre-processing results based on the data in the outdoor environment. It is similar with the signal pre-processing results based on the data collected in indoor environment the signal perturbation caused by pedestrian existence is obvious in the signal collected in 100Hz (see Figure 5-25 (a)) after signal pre-processing, and the waveforms of the signal data collected in higher sampling ratios have more minor disorderly fluctuations and the level of disorder increased as the sampling ratio increases (see Figure 5-25 (b) and (c)).

5.3.3.2 Pedestrian Existence Detection

For the pedestrian existence detection, the normalized average CSI amplitude in the time domain was used to identify the pedestrian existence based on the proposed algorithm. Table 5-8 shows the detection results for both indoor and outdoor environments for three different sampling ratios.

Table 5-8 Validation Results of Pedestrian Existence Detection

Sampling Ratio	Environment	# Pedestrian	# Correct Detection	Accuracy	FP Rate
100Hz	Indoor	130	129	99.23%	0.26%
	Outdoor	154	147	95.45%	1.53%
500Hz	Indoor	49	46	93.88%	2.07%
	Outdoor	48	45	93.75%	2.11%
800Hz	Indoor	24	13	54.17%	15.47%
	Outdoor	27	10	37.04%	21.26%

According to the results, the detection results based on the data collected in the 100Hz sampling ratio in the indoor environment achieved the best accuracy and the lowest FP rate. The accuracy is close to 100% and the FP rate is close to zero. As the sampling ratio increased from 100Hz to 800Hz, the accuracy and FP rate became worse. Since the 800Hz sampling ratio is too high, the Wi-Fi CSI signal is not only impacted by pedestrian existence but also is influenced by the other electromagnetic wave and the movements of minor objects in the surrounding area. Thus, the accuracy of the detection results dropped a lot comparing to the results calculated by 100Hz data and 500Hz data. Besides, since the environmental factors are more unstable in the outdoor

environment, the results of the outdoor environment were worse than the results of the indoor environment for the data of all three sampling ratios. But the detection accuracy in the outdoor environment is still at an acceptable level.

5.3.3.3 Pedestrian Moving Direction Recognition

Once a pedestrian existence event is detected, the proposed algorithm was utilized for recognizing the moving direction of the pedestrian. During the data detection procedure, only one person passing the LoS at a time. Thus, the moving direction is unidirectional. As the data collected with the 100Hz is the most effective in this implementation for pedestrian existence detection, so we only recognized the moving direction of the pedestrian for the data collected with the 100Hz sampling ratio. Table 5-9 shows the results of the moving direction recognition. For the indoor environment, the accuracy is 100% and 96.92% for both two directions. For the outdoor environment, since the stability of the Wi-Fi CSI signal is worse than in the indoor environment, the accuracy dropped a little comparing to the results of the indoor environment. However, the level of accuracy is still acceptable with 92.21% and 93.51% of two directions, respectively.

Table 5-9 Validation Results of Moving Direction Recognition using The Dataset in 100Hz

Environment	Direction 1			Direction 2		
	# Pedestrian	# Correct Detection	Accuracy	# Pedestrian	# Correct Detection	Accuracy
Indoor	65	65	100.00%	65	63	96.92%
Outdoor	77	71	92.21%	77	72	93.51%

5.3.3.4 Pedestrian Speed Estimation

The proposed Approach 1 and 2 were both conducted for the comparison purpose. These approaches would accurately measure the overall time of a pedestrian walking by the Wi-Fi CSI sensor and across its sensing zone.

5.3.3.4.1 Approach 1 Moving Average

A representative sample of detection results of approach 1 is shown in the following figure.

In Figure 5-26, the blue line is the original amplitude data of subcarrier 1 of antenna 2, the orange line is the detection result of approach 1, value of 1.0 indicates the existence of pedestrian. Comparing to pinpointing the start and end of pedestrian existence from the original signal, usage of correlation features and neural network identification can generate a sharper edge of detection in the orange line and could explicitly show the time of pedestrian existence.

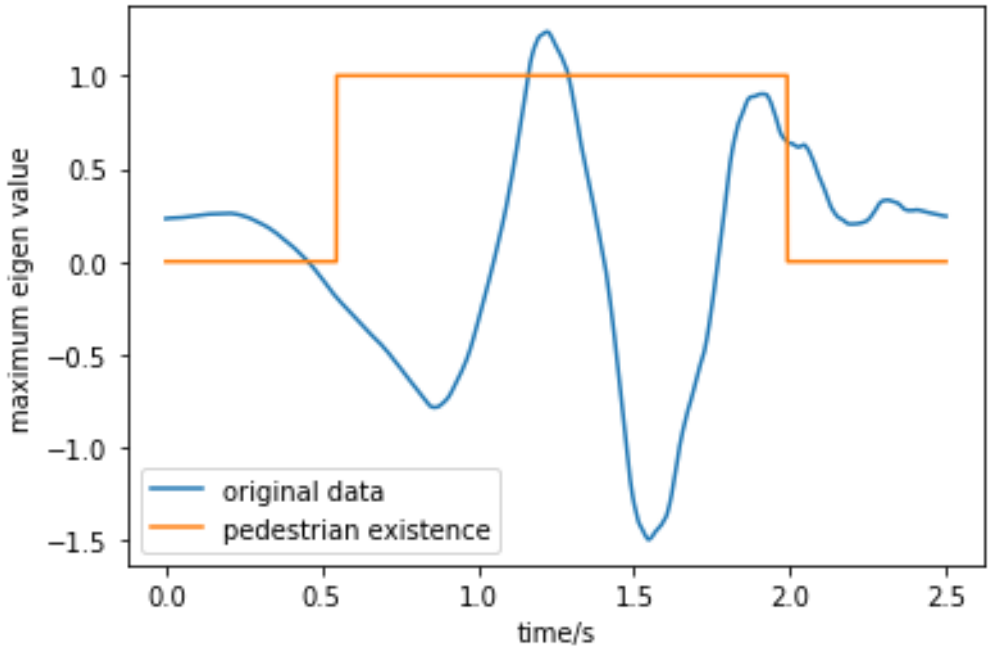


Figure 5-26 A Sample Detection Result of A Pedestrian in The Experiment by Approach 1

For the 42 samples of normal walking samples, the distribution of measured pedestrian crossing time by fixed threshold approach is illustrated in the following figure:

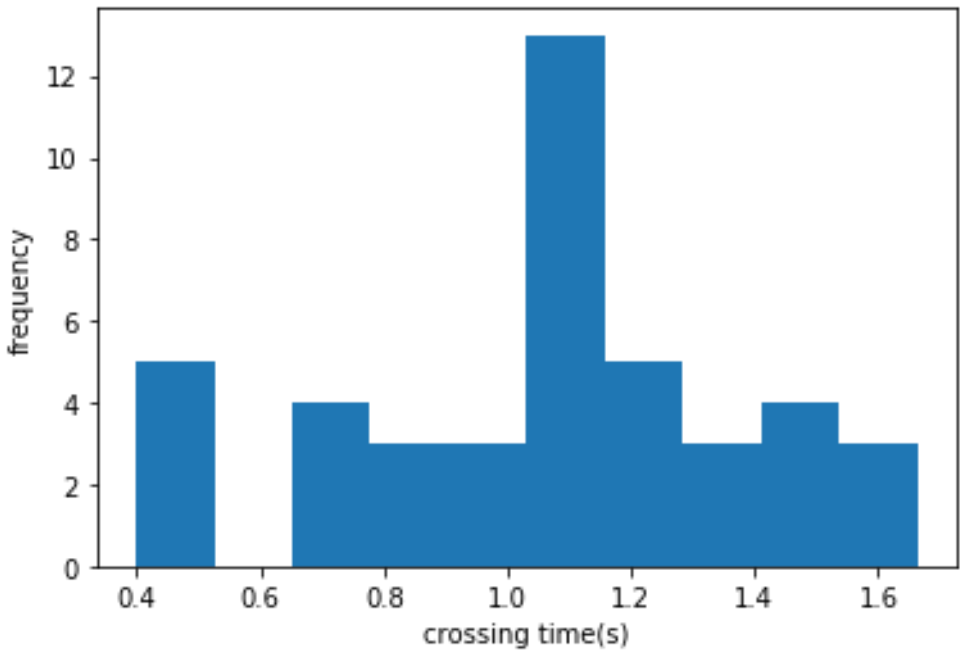


Figure 5-27 Distribution of Detected Pedestrian Walking Time Measured by The Fixed Threshold Approach

The distribution of calculated pedestrian crossing speed by fixed threshold approach is shown in the following figure.

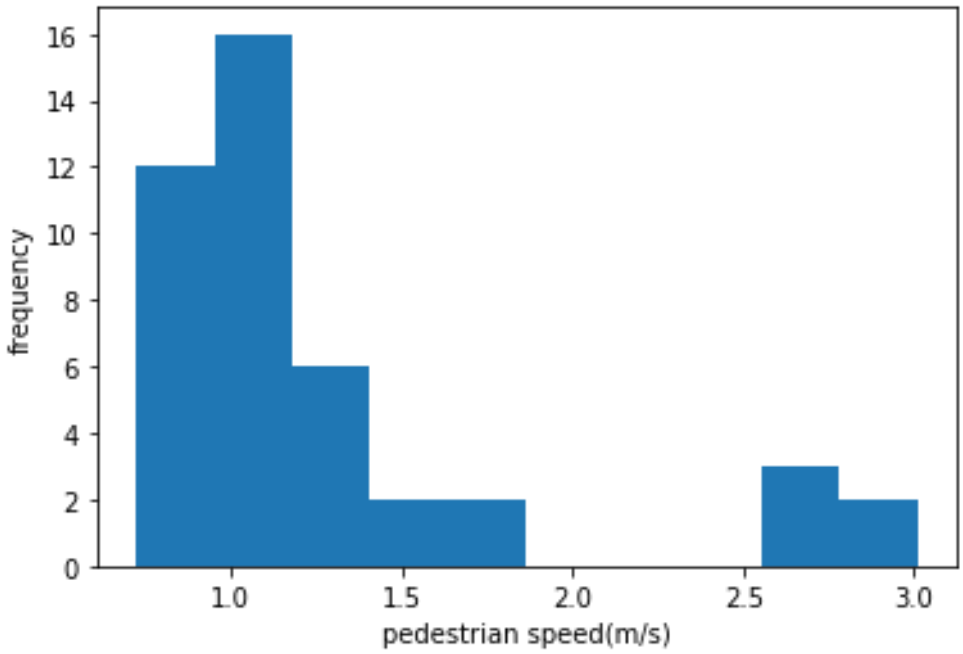


Figure 5-28 Distribution of Detected Pedestrian Speed Measured by The Fixed Threshold Approach

The mean of speed measured by this approach is 1.29m/s and the standard variation is 0.58m/s. From Figure 5-28, we can see that the main reason for the high variance is that for a few samples, the measurement gave result that is abnormally high. Eliminating them gives a result of 1.13m/s in mean speed and 0.34m/s in standard deviation.

5.3.3.4.2 Approach 2 Neural Network

A sample of pedestrian crossing time detected by Resnet is illustrated in Figure 5-29:

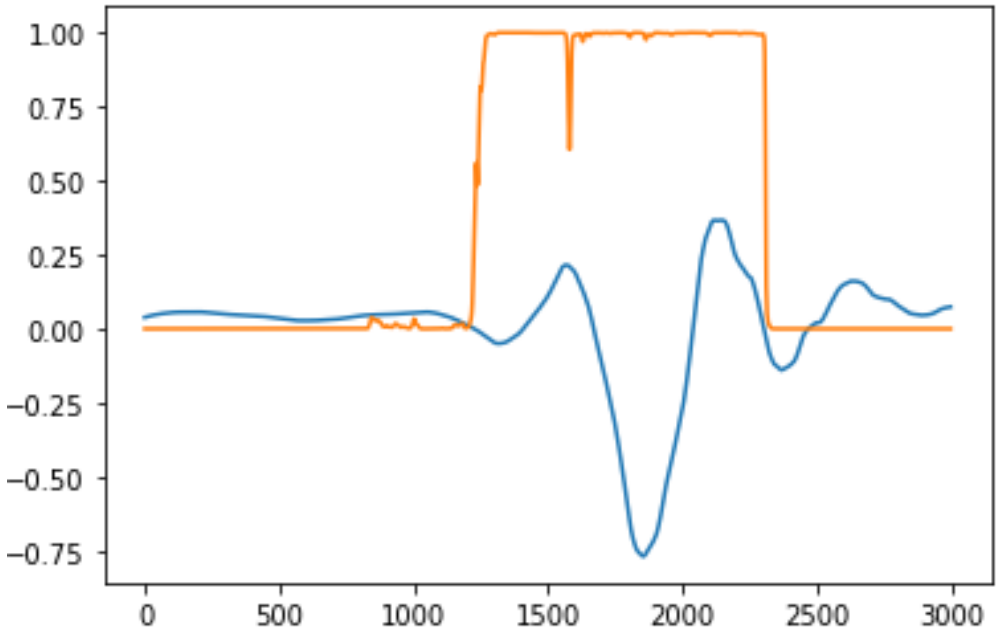


Figure 5-29 A Sample Detection Result of a Pedestrian in The Experiment by Approach 2

It is similar with Figure 5-26, the blue line is the original amplitude data of subcarrier 1 of antenna 2, the orange line is the detection result, namely the output of the neural network, of approach 2. A higher output of the neural network indicates higher probability of pedestrian existence. Comparing to the original data and the detection result of approach 1, this neural network approach could generate a sharp detection edge similar with approach 1. Moreover, this

approach does not have hard threshold which naturally generates sharp edge of detection, but rather indicates a sharp transition of subcarrier correlation between pedestrian existence and non-existence.

For the 42 samples of normal walking samples, the distribution of measured pedestrian crossing time by Resnet is illustrated in the following figure:

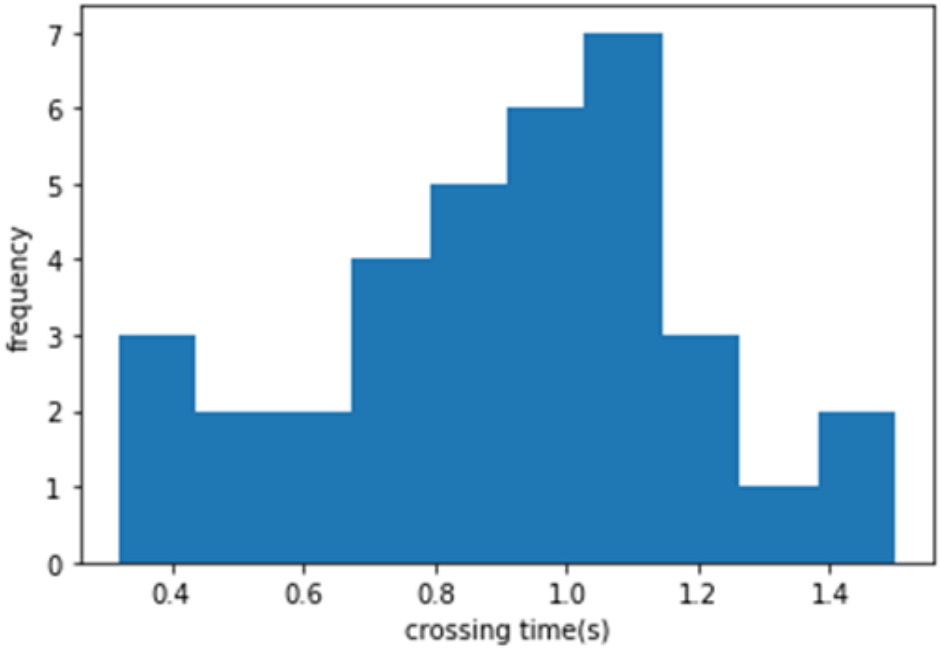


Figure 5-30 Distribution of Detected Pedestrian Walking Time Measured by The Neural Network Approach

The distribution of calculated pedestrian crossing speed by ResNet is shown in the following figure.

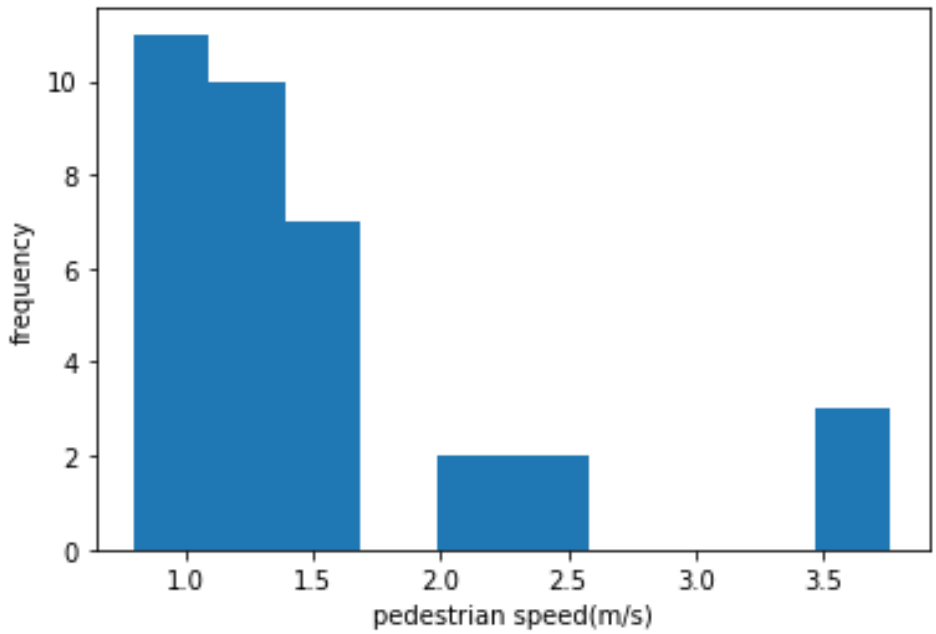


Figure 5-31 Distribution of Detected Pedestrian Speed Measured by The Neural Network Approach

The mean of speed measured by this approach is 1.27m/s and the standard variation is 0.46m/s. From Figure 5-31, we can see that the main reason for the high variance is that for a few samples, the measurement gave result that is abnormally high. Eliminating them gives a result of 1.06m/s in mean speed and 0.21m/s in standard deviation.

5.3.3.4.3 Discussion of detection result of approach 1 and 2

As illustrated in Figure 5-27 and 5-28, the most of crossing time measured by approach 1 is between 0.7 to 1.6s, with most samples around 1.1s. For speed measurement, most samples have a measured speed between 0.8m/s to 1.5m/s, with a few samples being the outlier at speed of 2.5m/s to 3.0m/s. Comparing to the usual pedestrian walking speed around 1.2m/s, the outlying measurement outcomes is significantly higher and should be considered erroneous. However, the outlier samples only occupy few percentages of the samples. Removing the outliers, the

measurement result has a median of 1.13m/s, which is very close to the empirical pedestrian walking speed.

For approach 2, as shown in Figure 5-30 and 5-31, the most of crossing time measured by approach 1 is between 0.6 to 1.2s, with most samples around 0.8s to 1.1s. For speed measurement, most samples have a measured speed between 0.8m/s to 1.5m/s. There are also some outlying results in this approach and their number is slightly higher than the number in approach 1, but both proportions are low. Removing the outliers, the measurement result has a median of 1.06m/s, which is very close to the empirical pedestrian walking speed. Comparing to approach 1, the velocity measurement result is relatively lower, and the result is slightly more unstable reflected by the number of outliers.

Overall, the two proposed approaches are efficient in measuring pedestrian velocity and could measure pedestrian velocity reliably in most circumstances.

5.3.3.4.4 The discussion about factors influencing speed measurement

In general, multiple factors could influence the result of speed measurement. The actual walking space for pedestrian is not a line without width in the experiment. In the samples, the pedestrian could be in the middle or relatively closer to the left or right side of the lane. Pedestrian walking in different lines could have different actual sensing range since it crosses the detection zone in different manners. This could impact the received signal since the obstacle for the signal and the length of sensing range are different in different samples. Another factor that could influence the result is the gait situation caught by the sensor. Since the sensing window is around 1s, pedestrians in various samples could be in varied part of his or hers gait cycle. This difference in walking manner could also affect the result of speed measurement.

5.3.4 *Concluding Remarks*

In summary, this research demonstrated the feasibility and the reliability for sensing pedestrian existence, recognizing pedestrian moving direction and estimating pedestrian moving speed using the Wi-Fi CSI signal. The experiment data was collected using three different sampling ratios in both indoor and outdoor environments. Hampel Identifier, Linear Interpolation, Kalman Filter and Wavelet Transform were employed to pre-process the Wi-Fi CSI raw signal. The pedestrian existence detection relied on the level of fluctuations of the normalized average CSI amplitude in the time domain. The pedestrian moving direction is recognized based on the Fresnel zone theory. The results indicated the proposed detection method based on the Wi-Fi CSI signal with the 100Hz sampling ratio achieved the best accuracy in pedestrian existence detection which is 99.23% in the indoor environment and 95.45% in the outdoor environment. For pedestrian moving direction recognition, the results of the outdoor environment are worse than those of indoor environments due to the more environmental influence. But the detection in the outdoor environment still achieved an acceptable accuracy with 92.21% and 93.51% of two directions, respectively. For pedestrian speed estimation, the proposed methods are effective to estimate accuracy pedestrian moving speed. The future research will continue in overlapped pedestrian identification, and pedestrian, bicyclists, and wheelchair classification.

Chapter 6. DISCUSSIONS

The proposed wireless sensing methods and technologies provide a real-time way to detect some traffic parameters of transit rider and non-motorized traffic which are hard to achieve accurate values by the existing sensing methods and technologies, e.g. O-D information and traffic speed. Traffic Data collection is the first step for improving and understanding traffic mobility. How the novel traffic data sources can benefit improvements in transportation systems still need more efforts in data mining and modeling. Thus, the first sub-section of this chapter introduces some potential implementation scenarios that may lack effective solutions based on the existing data sources. In addition, besides the aforementioned three uncertainties, other issues, e.g. randomized MAC address, been generating impacts on traffic sensing methods based on passive Wi-Fi and BT sensing technology. Hence, the second sub-section presents an evaluation of the impacts of MAC address randomization.

6.1 Potential Application of The Proposed Wireless Sensing Methods and Technologies

Traffic Congestion Spatial Causal Factors Identification Based on Wireless Sensing Data

Unlike traditional traffic sensing methods, wireless sensing not only can extract traffic parameters of roadway networks, e.g. traffic volume, travel time and traffic speed, but also the O-D information of a sample of travelers can be achieved. The O-D information can tell the spatial distribution of the locations where mainly generates and attracts traffic mobility. The basic traffic parameters only can help to observe the traffic status of road networks. By combining O-D information and basic traffic parameters, the traffic flow can be observed in detail. When traffic congestions occur, the origins of the traffic which mainly contribute to traffic congestions can be

explored, and even the magnitude of the contribution to traffic congestions can be quantified. Once the hotspots for generating traffic congestions are determined, transportation managers can build control strategy for traffic optimization. Thus, developing an algorithm to identify the spatial causal factors of traffic congestions can be an important future research topic.

Assisting Autonomous Driving by Cooperating with On-Vehicle Sensing System

Autonomous driving relies on in-vehicle sensing systems that assist drivers or fully control vehicles while driving and parking [160]. By far, environment perception and understanding remains to a critical challenge for autonomous driving [161]. Specifically, precise 3-D detection of vehicles and pedestrians [162], and understanding road users' intentions [163], are needed for roadway traffic decision making and parking assistance. Current advanced driving systems mainly rely on in-vehicle Light Detection and Ranging (Lidar) and cameras for environmental perception [164], which work well under certain circumstances, but has the following limitations:

1) relying on only in-vehicle sensors makes the perception system be limited in both spatial (limited horizontal and vertical fields of views and suffering from occlusions) and temporal (only observing other traffic participants for a limited amount of time and hard to fully understand their intentions) dimensions [165];
2) Vehicle lidar sensors are expensive, require a high communication bandwidth for transmitting the generated huge amount of point cloud data and suffer from degraded performances in cases of rain, snow, fog, and dust [166];
3) In-vehicle cameras are suitable for 2-D vehicle and pedestrian detections [167], [168] but have poor performances when used for 3-D object detections [169].

By having the proposed wireless sensing methods and technologies for non-motorized traffic monitoring, the affluent real-time data being collected by the road-side wireless sensing equipment

can be used as complementary data resources to broaden the views of current driving perception systems in both spatial and temporal dimensions. Multiple challenges (such as occlusion issues and low precisions of monocular 3-D detections) can be solved by fusing the real-time data of in-vehicle sensors and road-side wireless sensing devices instead of solely relying on in-vehicle sensing technologies.

Improving the Understanding of the Impacts of Built Environment on Non-Motorized Traffic Demand

The impacts of built environment on traffic demand is quite important for transportation planning purpose. Previous studies demonstrated the non-motorized traffic demand is highly corrected with built environment of urban areas [170]–[172]. However, due to the lack of non-motorized traffic data, the impacts are hard to be quantified. By having the proposed wireless sensing methods, the non-motorized traffic data can help with the improvements of the understanding of the built environment impacts on non-motorized traffic demand, therefore future improving non-motorized transportation infrastructures.

6.2 Evaluation of Randomized MAC Address Impacts

In 2014, Apple Inc. initialized a new function called MAC address randomization in iOS 8 operation system [173]. After that, Android and Windows systems started to employ MAC address randomization function. Basically, randomized MAC address is not a global unique identifier, and re-identification based on MAC address is no longer feasible if all MAC addresses are randomized. It generates considerable impacts on device-based wireless sensing for traffic parameter extraction. However, for the randomized MAC address, there are no existing literatures discussing or demonstrating its impacts on the traffic sensing methods based on passive Wi-Fi and BT-based

sensing technology. In this section, the impact of MAC address randomization is discussed based on a long-term data set collecting in Shanghai City and Seattle.

A MAC address is either globally unique or locally assigned. A global MAC address is the hardware MAC address of each WBM device, which is a globally unique identifier assigned to its network interface controller by the manufacturer. While, a locally assigned MAC address is assigned temporarily to override the global address, which is either randomized by the operating system for protecting privacy or generated for special local communication services [174]. The type of a given MAC address can be identified by the Universal/Local (U/L) bit in it (see Figure 6-1), which is the second-least significant bit of the first octet of the address. When the U/L bit is set as 1, the MAC address is locally assigned, otherwise the MAC address is globally unique [175].

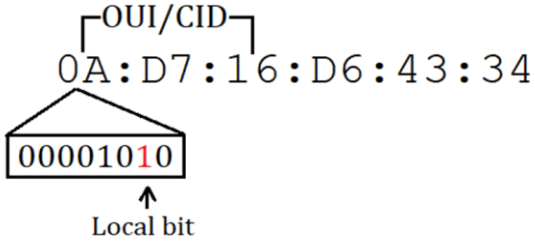


Figure 6-1 Universal/Local bit in a MAC address

Figure 6-2 presents a comparison of global MAC address rate in two places. In Figure 6-2 (a), the long-term monthly unique MAC address detection rate and global MAC address rate of the data set collecting in Shanghai City is presented. 50 hours data which were randomly selected between 8 a.m. to 20 p.m. are considered as the representatives of one month. The solid blue line displays the average hourly detection rate of unique MAC address by one MAC address detector. According to the results, from November 2017 to January 2020, the detection rate was relative stable ranging from 400 to 500. After January 2020, it was rapidly dropped under 100 due to the quarantine policy. The cyan line displays the globally unique MAC address rate of 50 hours data.

Seen from the figure, from November 2017 to August 2019, the rate gradually decreased from 35% to 18%, and then, the rate is relative stable ranging from 15% to 20%. The interpretation can be that some of the cellphone manufacturers gradually enhanced the privacy protection function of cellphones by lunching new embedded MAC address randomization mechanism before August 2019. After that, there are probably no manufacturers newly initialize function for randomizing MAC address, therefore the global MAC address rate remains stable.

Figure 6-2 (b) shows the globally unique MAC address rate of the data collected in Seattle. Five MAC address detectors were installed along SR-522 for data collection from October 17th to November 3rd, 2019. In this figure, three statistics are displayed, including the number of unique MAC address collecting by each MAC address detector which is shown in orange bars, the number of global MAC address collecting by each MAC address detector which is shown in blue bars, and the global MAC address rate of the data collecting by each MAC address detector which is shown in dashed red line. Basically, the number of unique MAC address collecting by each MAC address detector is various. Whereas, the global MAC address rate of each MAC address detector is relative consistent through all MAC address detectors. Besides, it is noticed that the global MAC address rate of the data collecting in Seattle is almost tripled the global MAC address rate of the data collecting in Shanghai City. The reason could be the lower frequency of people changing cellphones in Seattle. Thus, the MAC address randomization function is still inactivated for the most of cellphones are in use.

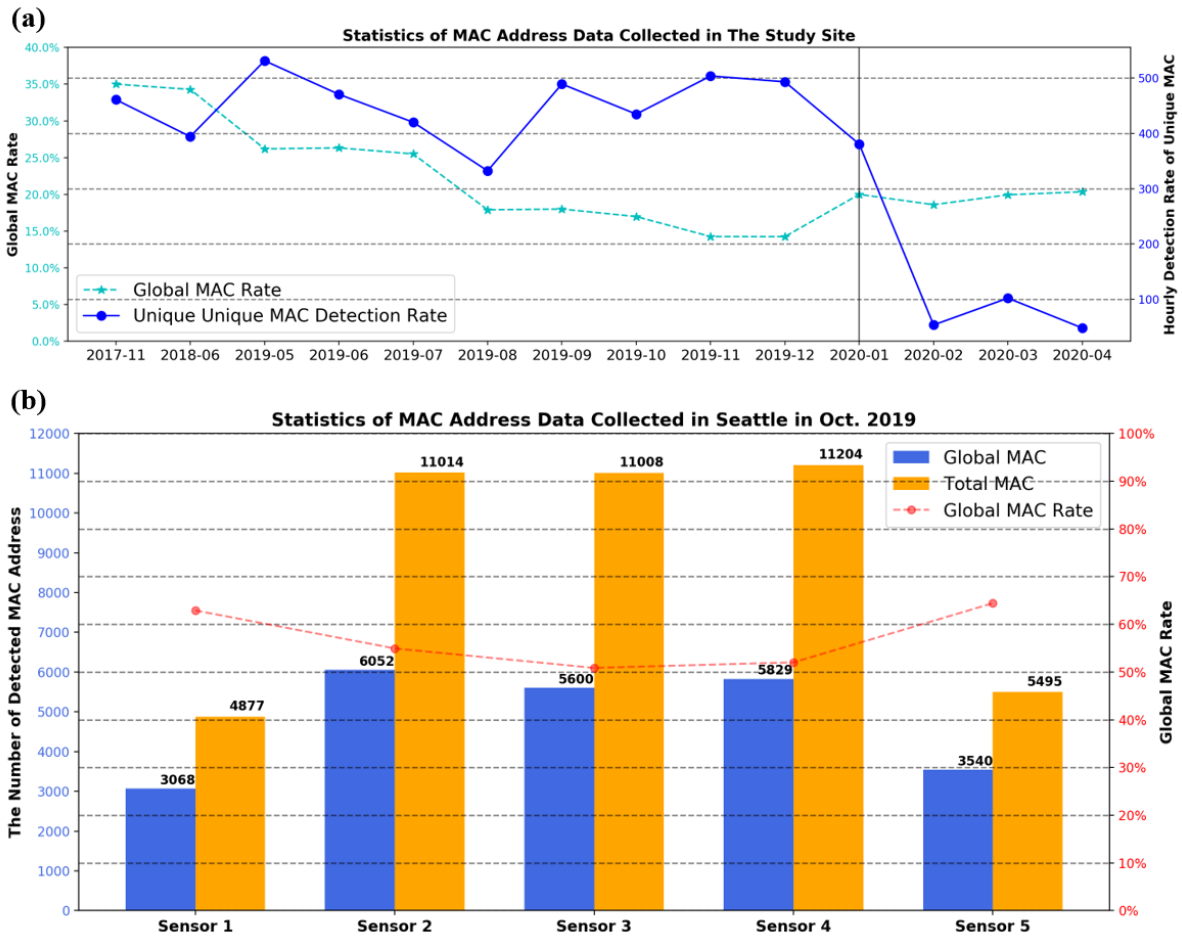
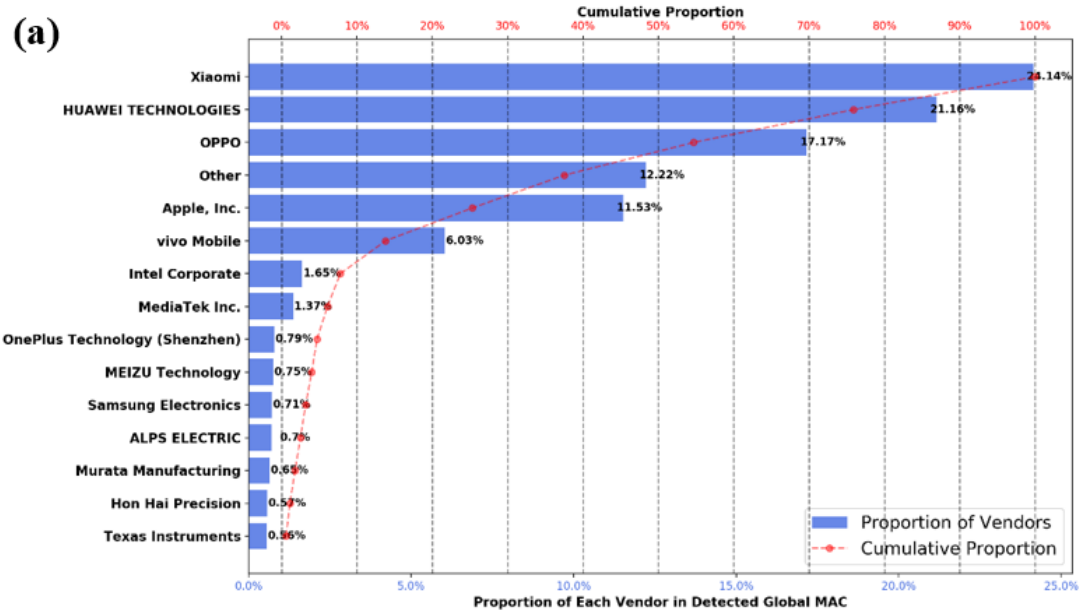


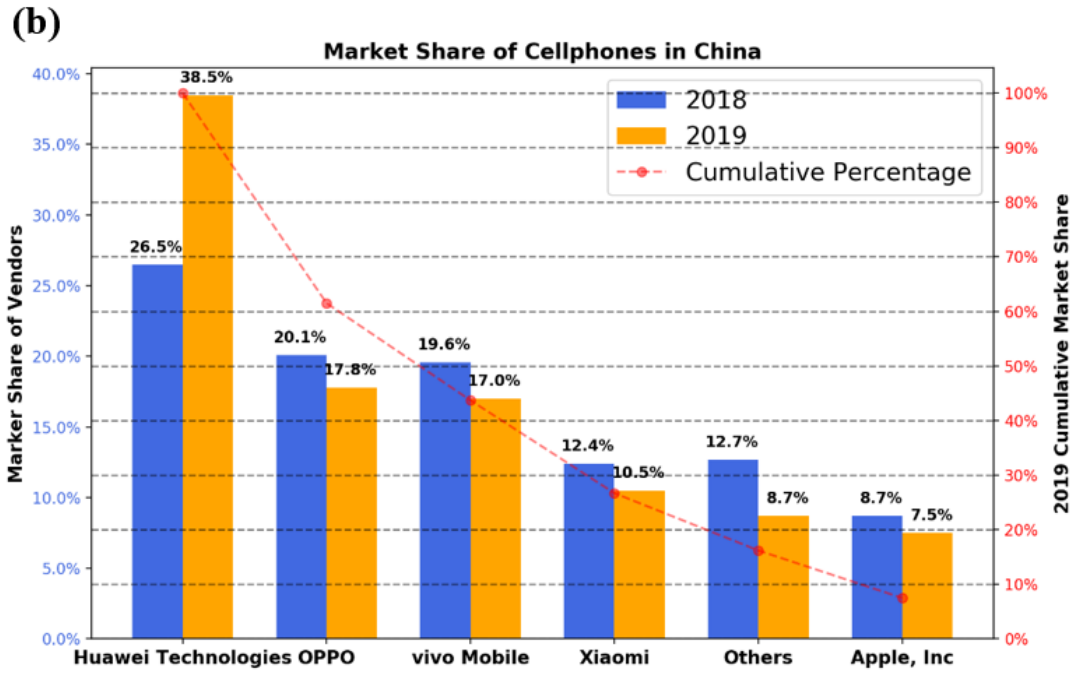
Figure 6-2 Global MAC Address Rate. (a) Statistics of The Data Collected in Shanghai City, and (b) Statistics of The Data Collected in Seattle

Besides the global MAC rate, the vendor distribution of global MAC address probably can tell the status of MAC address randomization implementation. Figure 6-3 (a) and (c) show the vendor distribution of global MAC address collecting in Shanghai City and Seattle. The vendors occupying less than 0.5% global MAC address were integrated into the "Other" category. For the global MAC address collecting in the study site, there are six vendors, including "Other" category, dominate the global MAC address generation, which are Xiaomi, HUAWEI, OPPO, Apple, Inc, and vivo. The proportion of rest of the vendors is less than 2%. In the data collecting in Seattle, there are 11 vendors generated over 2% of global MAC address. Among them, except for the

"Other" category, 4 of them are the manufacturers of automotive technology, e.g. ALPS Electrics, eSSys, Garmin International, and Pioneer Corporation. Comparing the vendor distribution of global MAC address with the market share of cellphones in China and the US which is shown in Figure 6-3 (b) and (d), it is consistent that, in China, the top six vendors in the rank of vendor distribution of global MAC address shared over 90% percentage cellphone market in China in 2019. However, in the US, the ranking order in two ranks are quite different. Levono achieved the top one order in the rank of vendor distribution of global MAC address that 24.41% global MAC addresses of Levono's WBM device were detected. For the vendors, e.g. Apple, Samsung, and LG Electronics, they shared over 80% cellphone market in 2018 and 2019, however, only about 15% global MAC addresses of these vendors were detected.

In summary of the analysis about global MAC address rate in two data set, the major manufacturers have not implemented MAC address randomization according to the data collecting in Shanghai City, but some of minor vendors have implemented. The current global MAC address rate is stable around the level of 18% based on the long-term data analysis. The average hourly global MAC address detection rate is about 100 for one MAC address detector. Thus, the level of data quantity still can support the sensing system to capture enough valid MAC trips for the multi-modal traffic speed estimation. In the dataset collecting in the US, even some major cellphone manufacturers have implemented MAC address randomization due to the privacy concerns, other cellphone and automotive devices still can provide considerable amount of global MAC address data, the average global MAC address rate is over 50 %, to support the implementation of the proposed system.


(a)
Cumulative Proportion
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
Xiaomi
HUAWEI TECHNOLOGIES
OPPO
Other
Apple, Inc.
vivo Mobile
Intel Corporate
MediaTek Inc.
OnePlus Technology (Shenzhen)
MEIZU Technology
Samsung Electronics
ALPS ELECTRIC
Murata Manufacturing
Hon Hai Precision
Texas Instruments
24.14%
21.16%
17.17%
12.22%
11.53%
6.03%
1.65%
1.37%
0.79%
0.75%
0.71%
0.7%
0.65%
0.57%
0.56%
Proportion of Vendors
Cumulative Proportion
0.0%
5.0%
10.0%
15.0%
20.0%
25.0%
Proportion of Each Vendor in Detected Global MAC


(b)
Market Share of Cellphones in China
2018
2019
Cumulative Percentage
Marker Share of Vendors
2019 Cumulative Market Share
40.0%
35.0%
30.0%
25.0%
20.0%
15.0%
10.0%
5.0%
0.0%
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
38.5%
26.5%
20.1%
17.8%
19.6%
17.0%
12.4%
10.5%
12.7%
8.7%
8.7%
7.5%
Huawei Technologies
OPPO
vivo Mobile
Xiaomi
Others
Apple, Inc

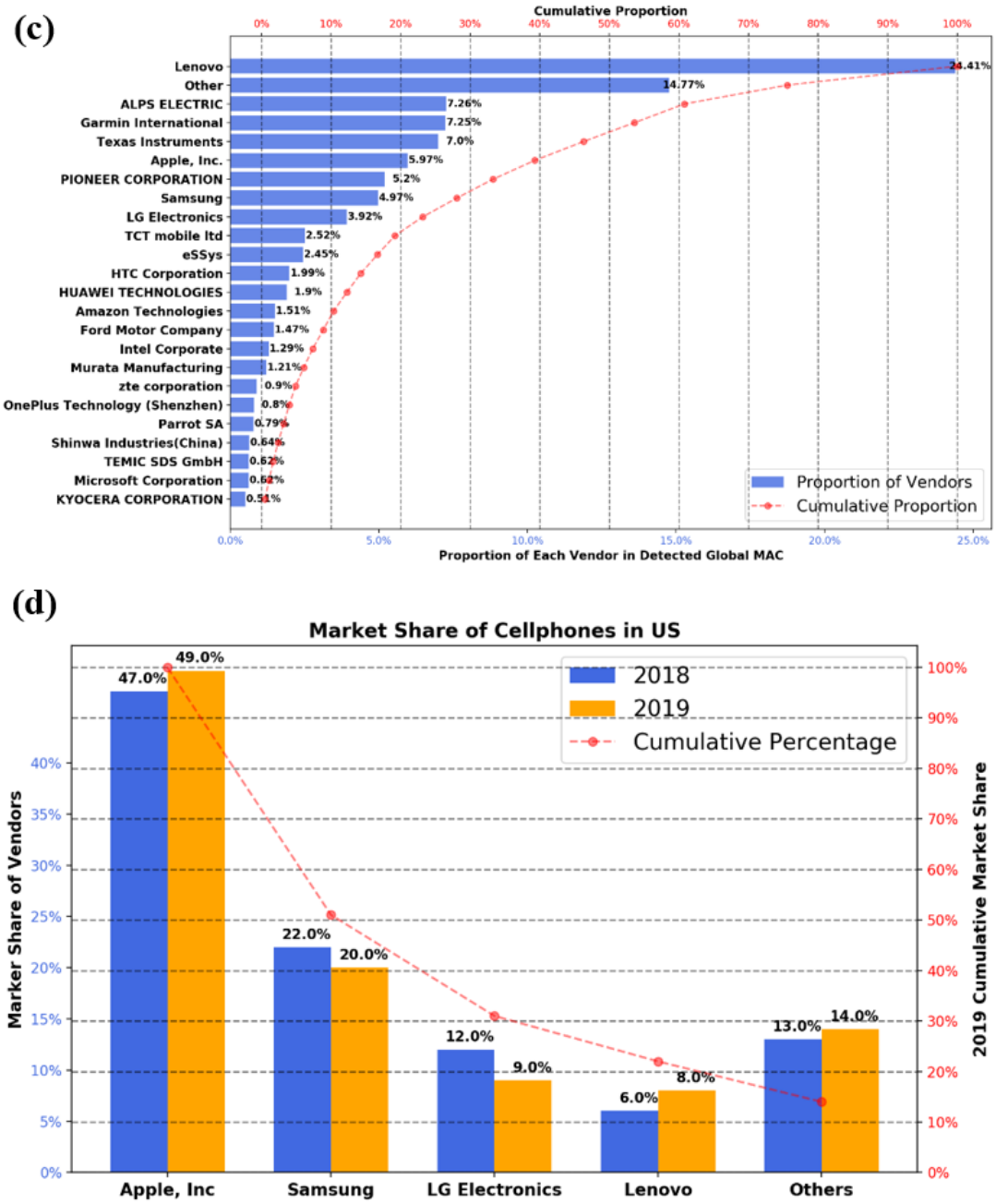


Figure 6-3 Vendor Distribution of Global MAC Address and Market Sharing of Vendors. (a) Vendor Distribution of Global MAC Address Collecting in Tongji University, (b) Market Share of Cellphone Vendors in China, (c) Vendor Distribution of Global MAC Address Collecting in Seattle, and (d) Market Share of Cellphone Vendors in US

Chapter 7. CONCLUSIONS AND FUTURE WORK

7.1 SUMMARY OF CONTRIBUTIONS

As data-driven traffic modeling approaches enhance traffic management and control systems with smarter intelligence, lower latency, better reliability than traditional traffic analysis methods, the advancement of traffic modeling for mobility pattern exploration, traffic control, and network optimization relies on real-time traffic data with more weights. The existing traffic sensing methods work reasonably well on sensing dynamic movements of motorized vehicles. However, collecting traffic data of transit rider and non-motorized still faces challenges. As the prevalent usage of smartphones, previous studies developed sensing methods for transit riders and non-motorized traffic by identifying WBM devices instead of detecting travelers directly. The findings of previous research presented the feasibility and effectiveness of such sensing technology which is called wireless sensing. Nevertheless, it is still limited in sensing accuracy and sample representativeness due to three types of uncertainties, including traffic mode uncertainty, localized spatial uncertainty, and population uncertainty. This dissertation proposes methodologies to reduce the errors caused by the introduced three types of uncertainties for transit rider and non-motorized traffic data acquisition. The proposed algorithms are validated by comparing the modeling results and the ground truth data. The major contributions of this dissertation can be summarized as follows.

Transit Ridership Flow Monitoring

1) An algorithm is designed to separate the wireless signals of passengers and non-passengers based on Fuzzy C-Means clustering.

2) The population number of passengers is estimated based on the proposed algorithm based on Random Forest regression.
3) The architecture of the system for monitoring real-time public transit ridership flow is designed based on the customized Wi-Fi and BT MAC address detector.
4) The proposed system is implemented on three transit routes in Seattle for collecting wireless sensing data. The ground truth data is collected manually for validating the performance of the proposed algorithms by comparing with other selected baseline models. The evaluation results indicate that the proposed algorithms are accurate.
5) The performance of the proposed algorithm is also compared with the existing filtering methods. The experimental results indicate the proposed algorithm can highly improve the estimation accuracy.

Multi-Modal Traffic Speed Monitoring

1) An algorithm is developed to correct the estimated traffic speed based on Received Signal Strength Indicator of Wi-Fi and BT signals. Ground truth speed measurements and the estimated traffic speed of 408 trips are compared to validate the performance. The accuracy of the proposed algorithm can be implied based on the comparison results.
2) Traffic mode identification algorithm is proposed based on a designed semi-supervised Possibilistic Fuzzy C-Means clustering algorithm. Multiple baseline algorithms are selected for the evaluation purpose. The evaluation results demonstrate the advantage of the algorithm in terms of accuracy.
3) A real-time multi-model traffic speed monitoring system is designed to explicitly describe the required components and the data streaming among the components. The proposed architecture can be used as a guideline for implementations.

4) A multi-modal traffic speed estimation algorithm is established for estimating traffic speed of the road networks covered by customized MAC address detector in a real-time way. The accuracy of the proposed algorithm is evaluated based on the comparison of estimated results and ground truth data. The evaluation results indicate the proposed algorithm is accurate for all three traffic modes, including walk, bike, and car modes.

Pedestrian Dynamic Movements Detection

1) An algorithm framework is developed for Wi-Fi CSI signal denoising method based on Hampel Identifier, Linear Signal Interpolation, Kalman Filter and Wavelet Transform.
2) Pedestrian existence detection method is designed based on the level of fluctuations of the normalized average CSI amplitude in the time domain.
3) An algorithm for identifying pedestrian moving direction is built based on Fresnel Zone theory.
4) Pedestrian moving speed estimation algorithm is established based on extracted features of Wi-Fi CSI signals.
5) The proposed algorithms are evaluated by the data collecting in indoor and outdoor environments with different Wi-Fi CSI sampling ratios. The evaluation results demonstrate the accuracy of the proposed algorithms.

7.2 FUTURE WORKS

This dissertation provides solutions to address three major uncertainties that cause considerable errors to the traffic parameters extracted from wireless sensing data. However, there are still some issues that need to be solved when it is applied in large-scale implementations or under more complex traffic conditions. Thus, future research in this direction includes the following specific topics.

Addressing Route Uncertainty of Wireless Sensing

This dissertation provides the solution to identify traffic modes and estimate traffic speed for a roadway network which has wireless sensors installing at every intersection. In this case, only one route between two adjacent sensing points. However, when it is implemented in a large-scale road network for traffic monitoring, it is hard to deploy sensors at every intersection. There are might have several potential routes for travelling between the adjacent sensing points. Thus, such route uncertainty can potentially hurt the sensing results in terms of accuracy. Future research should target a solution dedicating to reduce the impacts caused by route uncertainty, therefore enhancing the whole system.

Solving Population Uncertainty Caused by MAC Address Randomization

Since 2014, cellphone manufacturers initialized MAC address randomization mechanism due to privacy concerns. In Chapter 4, the status of MAC address randomization is analyzed based on long-term wireless sensing data collecting in two places. Based on the analysis results, there are no cellphone manufacturers newly initialize randomized MAC address during the recent years. Besides, it is found that the ratio of randomized MAC is about 80% in China and 40% in the US. In China, the major cellphone vendors have not deployed MAC address randomization, but it has been deployed by US major cellphone vendors. Although it demonstrates that multi-modal traffic speed monitoring is not highly impacted by MAC address randomization, the population uncertainty of wireless sensing data is certainly influenced by it. Thus, establishing solution to solve population uncertainty caused by randomized MAC address is a significant research direction.

Enhancement of Device-Free Wireless Sensing for Non-Motorized Traffic Monitoring

In Chapter 5 of this dissertation, the feasibility of device-free wireless sensing for pedestrian monitoring is thoroughly demonstrated. It certainly reduces the population uncertainty of device-based wireless sensing to some extent. However, there is way more to go for real implementations of such technology. The future research includes distinguishing bicyclist, pedestrian, and wheelchairs, identifying overlapped travelers, and crowd event identification based on device-free wireless sensing data as well.

BIBLIOGRAPHY

[1] X. Ma, Z. Tao, Y. Wang, H. Yu, and Y. Wang, "Long short-term memory neural network for traffic speed prediction using remote microwave sensor data," *Transp. Res. Part C Emerg. Technol.*, vol. 54, pp. 187–197, 2015.

[2] Y. Lv, Y. Duan, W. Kang, Z. Li, and F. Y. Wang, "Traffic Flow Prediction With Big Data: A Deep Learning Approach," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 2, pp. 865–873, 2014.

[3] C. Peng, X. Jin, K.-C. Wong, M. Shi, and P. Liò, "Collective human mobility pattern from taxi trips in urban area," *PLoS One*, vol. 7, no. 4, 2012.

[4] Y. Zheng, L. Capra, O. Wolfson, and H. Yang, "Urban computing: concepts, methodologies, and applications," *ACM Trans. Intell. Syst. Technol.*, vol. 5, no. 3, pp. 1–55, 2014.

[5] Y. Zheng, Q. Li, Y. Chen, X. Xie, and W.-Y. Ma, "Understanding mobility based on GPS data," in *Proceedings of the 10th international conference on Ubiquitous computing*, 2008, pp. 312–321.

[6] W. Genders and S. Razavi, "Using a deep reinforcement learning agent for traffic signal control," *arXiv Prepr. arXiv1611.01142*, 2016.

[7] Y. Wang and Z. Zeng, "Overview of Data-Driven Solutions," in *Data-Driven Solutions to Transportation Problems*, Elsevier, 2019, pp. 1–10.

[8] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[9] 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, p. 117550, 2019.

[10] 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.

[11] V. R. Vuchic, *Urban transit: operations, planning, and economics*. John Wiley & Sons, 2017.

[12] V. R. Vuchic, *Urban transit systems and technology*. John Wiley & Sons, 2007.

[13] M. Ben-Akiva, P. P. Macke, and P. S. Hsu, *Alternative methods to estimate route-level trip tables and expand on-board surveys*, no. 1037. 1985.

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

[15] T. Kusakabe and Y. Asakura, "Behavioural data mining of transit smart card data: A data

fusion approach," *Transp. Res. Part C Emerg. Technol.*, vol. 46, pp. 179–191, 2014.

[16] 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.

[17] 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. Part C Emerg. Technol.*, vol. 58, pp. 193–207, 2015.

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

[19] 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.

[20] Y. Malinovskiy, N. Saunier, and Y. Wang, "Analysis of pedestrian travel with static bluetooth sensors," *Transp. Res. Rec.*, vol. 2299, no. 1, pp. 137–149, 2012.

[21] P. Reisman, O. Mano, S. Avidan, and A. Shashua, "Crowd detection in video sequences," in *IEEE Intelligent Vehicles Symposium, 2004*, 2004, pp. 66–71.

[22] H. Cho, P. E. Rybski, and W. Zhang, "Vision-based bicycle detection and tracking using a deformable part model and an EKF algorithm," in *13th International IEEE Conference on Intelligent Transportation Systems*, 2010, pp. 1875–1880.

[23] I. P. Alonso *et al.*, "Combination of feature extraction methods for SVM pedestrian detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 8, no. 2, pp. 292–307, 2007.

[24] M. S. Shehata *et al.*, "Video-based automatic incident detection for smart roads: The outdoor environmental challenges regarding false alarms," *IEEE Trans. Intell. Transp. Syst.*, vol. 9, no. 2, pp. 349–360, 2008.

[25] S. Hankey *et al.*, "Estimating use of non-motorized infrastructure: Models of bicycle and pedestrian traffic in Minneapolis, MN," *Landsc. Urban Plan.*, vol. 107, no. 3, pp. 307–316, 2012.

[26] W. Schwartz, *Bicycle and pedestrian data: Sources, needs, and gaps*. 2000.

[27] W. D. Cottrell and D. Pal, "Evaluation of pedestrian data needs and collection efforts," *Transp. Res. Rec.*, vol. 1828, no. 1, pp. 12–19, 2003.

[28] Y. Wang and N. L. Nihan, "Can single-loop detectors do the work of dual-loop detectors?," *J. Transp. Eng.*, vol. 129, no. 2, pp. 169–176, 2003.

[29] Y. Malinovskiy, Y.-J. Wu, and Y. Wang, "Video-based monitoring of pedestrian movements at signalized intersections," *Transp. Res. Rec.*, vol. 2073, no. 1, pp. 11–17, 2008.

[30] B. De Mersseman and S. Decker, “Sensor system with radar sensor and vision sensor.” Google Patents, 2006.

[31] H. Cho, Y.-W. Seo, B. V. K. V. Kumar, and R. R. Rajkumar, “A multi-sensor fusion system for moving object detection and tracking in urban driving environments,” in *2014 IEEE International Conference on Robotics and Automation (ICRA)*, 2014, pp. 1836–1843.

[32] D. Z. Wang, I. Posner, and P. Newman, “What could move? finding cars, pedestrians and bicyclists in 3d laser data,” in *2012 IEEE International Conference on Robotics and Automation*, 2012, pp. 4038–4044.

[33] M. Tons, R. Doerfler, M.-M. Meinecke, and M. A. Obojski, “Radar sensors and sensor platform used for pedestrian protection in the EC-funded project SAVE-U,” in *IEEE Intelligent Vehicles Symposium, 2004*, 2004, pp. 813–818.

[34] D. A. Noyce, A. Gajendran, and R. Dharmaraju, “Development of bicycle and pedestrian detection and classification algorithm for active-infrared overhead vehicle imaging sensors,” *Transp. Res. Rec.*, vol. 1982, no. 1, pp. 202–209, 2006.

[35] Z. Zheng, L. Zheng, and Y. Yang, “Pedestrian alignment network for large-scale person re-identification,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 10, pp. 3037–3045, 2018.

[36] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, “Person re-identification by symmetry-driven accumulation of local features,” in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010, pp. 2360–2367.

[37] K. Liu, B. Ma, W. Zhang, and R. Huang, “A spatio-temporal appearance representation for viceo-based pedestrian re-identification,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3810–3818.

[38] B. C. Chee, M. Lazarescu, and T. Tan, “Detection and monitoring of passengers on a bus by video surveillance,” in *14th International Conference on Image Analysis and Processing (ICIAP 2007)*, 2007, pp. 143–148.

[39] F. Wang, J. Wang, J. Cao, C. Chen, and X. J. Ban, “Extracting trips from multi-sourced data for mobility pattern analysis: An app-based data example,” *Transp. Res. Part C Emerg. Technol.*, vol. 105, pp. 183–202, 2019.

[40] N. D. Lane, E. Miluzzo, H. Lu, D. Peebles, T. Choudhury, and A. T. Campbell, “A survey of mobile phone sensing,” *IEEE Commun. Mag.*, vol. 48, no. 9, pp. 140–150, 2010.

[41] W. Wen, “An intelligent traffic management expert system with RFID technology,” *Expert Syst. Appl.*, vol. 37, no. 4, pp. 3024–3035, 2010.

[42] M. Trépanier, N. Tranchant, and R. Chapleau, “Individual trip destination estimation in a

transit smart card automated fare collection system," *J. Intell. Transp. Syst.*, vol. 11, no. 1, pp. 1–14, 2007.

[43] C. Slobogin, "Public privacy: camera surveillance of public places and the right to anonymity," *Miss. lJ*, vol. 72, p. 213, 2002.

[44] R. Weinstein, "RFID: a technical overview and its application to the enterprise," *IT Prof.*, vol. 7, no. 3, pp. 27–33, 2005.

[45] J. B. Kenney, "Dedicated short-range communications (DSRC) standards in the United States," *Proc. IEEE*, vol. 99, no. 7, pp. 1162–1182, 2011.

[46] J. S. Wasson, J. R. Sturdevant, and D. M. Bullock, "Real-time travel time estimates using media access control address matching," *ITE J.*, 2008.

[47] "• Global mobile phone internet user penetration 2019 | Statistic." [Online]. Available: https://www.statista.com/statistics/284202/mobile-phone-internet-user-penetration-worldwide/. [Accessed: 13-Mar-2019].

[48] "• Smartphone penetration in the US (share of population) 2010-2021 | Statistic." [Online]. Available: https://www.statista.com/statistics/201183/forecast-of-smartphone-penetration-in-the-us/. [Accessed: 13-Mar-2019].

[49] D. Johnston and J. Walker, "Overview of IEEE 802.16 security," *IEEE Secur. Priv.*, vol. 2, no. 3, pp. 40–48, 2004.

[50] 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*, no. 2595, pp. 11–17, 2016.

[51] 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*, no. 2544, pp. 123–130, 2016.

[52] 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.*, p. 0361198118776153, 2018.

[53] 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, 2014.

[54] 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, 2017.

[55] 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,"

2018.

[56] V. Kostakos, T. Camacho, and C. Mantero, “Wireless detection of end-to-end passenger trips on public transport buses,” in *13th International IEEE Conference on Intelligent Transportation Systems*, 2010, pp. 1795–1800.

[57] A. Lesani and L. F. Miranda-Moreno, “Development and Testing of a Real-Time WiFi-Bluetooth System for Pedestrian Network Monitoring and Data Extrapolation,” 2016.

[58] 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.

[59] Y. Wang, Y. Malinovskiy, Y.-J. Wu, U. K. Lee, and M. Neeley, “Error modeling and analysis for travel time data obtained from Bluetooth MAC address matching,” *Dep. Civ. Environ. Eng. Univ. Washingt.*, 2011.

[60] Y. Malinovskiy, U.-K. Lee, Y.-J. Wu, and Y. Wang, “Investigation of bluetooth-based travel time estimation error on a short corridor,” 2011.

[61] J. D. Porter, D. S. Kim, M. E. Magaña, P. Poocharoen, and C. A. G. Arriaga, “Antenna characterization for Bluetooth-based travel time data collection,” *J. Intell. Transp. Syst.*, vol. 17, no. 2, pp. 142–151, 2013.

[62] A. M. Hainen, J. S. Wasson, S. M. L. Hubbard, S. M. Remias, G. D. Farnsworth, and D. M. Bullock, “Estimating route choice and travel time reliability with field observations of Bluetooth probe vehicles,” *Transp. Res. Rec.*, vol. 2256, no. 1, pp. 43–50, 2011.

[63] B. N. Araghi, J. Hammershøj Olesen, R. Krishnan, L. Tørholm Christensen, and H. Lahrmann, “Reliability of bluetooth technology for travel time estimation,” *J. Intell. Transp. Syst.*, vol. 19, no. 3, pp. 240–255, 2015.

[64] Y. Aliari and A. Haghani, “Bluetooth sensor data and ground truth testing of reported travel times,” *Transp. Res. Rec.*, vol. 2308, no. 1, pp. 167–172, 2012.

[65] A. Bhaskar and E. Chung, “Fundamental understanding on the use of Bluetooth scanner as a complementary transport data,” *Transp. Res. Part C Emerg. Technol.*, vol. 37, pp. 42–72, 2013.

[66] Y. Malinovskiy, Y.-J. Wu, Y. Wang, and U. K. Lee, “Field experiments on bluetooth-based travel time data collection,” 2010.

[67] M. Martchouk, F. Mannering, and D. Bullock, “Analysis of freeway travel time variability using Bluetooth detection,” *J. Transp. Eng.*, vol. 137, no. 10, pp. 697–704, 2011.

[68] N. Abedi, A. Bhaskar, E. Chung, and M. Miska, “Assessment of antenna characteristic effects on pedestrian and cyclists travel-time estimation based on Bluetooth and WiFi MAC addresses,” *Transp. Res. Part C Emerg. Technol.*, vol. 60, pp. 124–141, 2015.

[69] T. M. Brennan Jr, J. M. Ernst, C. M. Day, D. M. Bullock, J. V Krogmeier, and M. Martchouk, “Influence of vertical sensor placement on data collection efficiency from bluetooth MAC address collection devices,” *J. Transp. Eng.*, vol. 136, no. 12, pp. 1104–1109, 2010.

[70] A. Bhaskar, M. Qu, and E. Chung, “Bluetooth vehicle trajectory by fusing Bluetooth and loops: Motorway travel time statistics,” *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 1, pp. 113–122, 2014.

[71] L. Jie, H. Van Zuylen, L. Chunhua, and L. Shoufeng, “Monitoring travel times in an urban network using video, GPS and Bluetooth,” *Procedia-Social Behav. Sci.*, vol. 20, pp. 630–637, 2011.

[72] C. Bachmann, M. J. Roorda, B. Abdulhai, and B. Moshiri, “Fusing a bluetooth traffic monitoring system with loop detector data for improved freeway traffic speed estimation,” *J. Intell. Transp. Syst.*, vol. 17, no. 2, pp. 152–164, 2013.

[73] H. Park and A. Haghani, “Optimal number and location of bluetooth sensors considering stochastic travel time prediction,” *Transp. Res. Part C Emerg. Technol.*, vol. 55, pp. 203–216, 2015.

[74] S. Yang and Y.-J. Wu, “Travel mode identification using bluetooth technology,” *J. Intell. Transp. Syst.*, vol. 22, no. 5, pp. 407–421, 2018.

[75] D. Schrank, B. Eisele, T. Lomax, and J. Bak, “2015 urban mobility scorecard,” 2015.

[76] D. D. Puckett, M. J. Vickich, and others, “Bluetooth-based travel time/speed measuring systems development.,” 2010.

[77] W. Qiao, A. Haghani, and M. Hamedi, “A nonparametric model for short-term travel time prediction using bluetooth data,” *J. Intell. Transp. Syst.*, vol. 17, no. 2, pp. 165–175, 2013.

[78] J. Barcelö, L. Montero, L. Marqués, and C. Carmona, “Travel time forecasting and dynamic origin-destination estimation for freeways based on bluetooth traffic monitoring,” *Transp. Res. Rec.*, vol. 2175, no. 1, pp. 19–27, 2010.

[79] T. Tsubota, A. Bhaskar, E. Chung, and R. Billot, “Arterial traffic congestion analysis using Bluetooth duration data,” 2011.

[80] Z. Mei, D. Wang, and J. Chen, “Investigation with Bluetooth sensors of bicycle travel time estimation on a short corridor,” *Int. J. Distrib. Sens. Networks*, vol. 8, no. 1, p. 303521, 2012.

[81] R. J. Haseman, J. S. Wasson, and D. M. Bullock, “Real-time measurement of travel time delay in work zones and evaluation metrics using bluetooth probe tracking,” *Transp. Res. Rec.*, vol. 2169, no. 1, pp. 40–53, 2010.

[82] C. M. Day, T. M. Brennan, A. M. Hainen, S. M. Remias, and D. M. Bullock, “Roadway

system assessment using Bluetooth-based automatic vehicle identification travel time data," 2012.

[83] J. J. V. Díaz, A. B. R. González, and M. R. Wilby, "Bluetooth traffic monitoring systems for travel time estimation on freeways," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 1, pp. 123–132, 2015.

[84] A. Lesani and L. 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.*, 2018.

[85] Y. Zhou, B. P. L. Lau, Z. Koh, C. Yuen, and B. K. K. Ng, "Understanding Crowd Behaviors in a Social Event by Passive WiFi Sensing and Data Mining," *IEEE Internet Things J.*, vol. 7, no. 5, pp. 4442–4454, 2020.

[86] A. Saeedi, S. Park, D. S. Kim, and J. D. Porter, "Improving accuracy and precision of travel time samples collected at signalized arterial roads with bluetooth sensors," *Transp. Res. Rec.*, vol. 2380, no. 1, pp. 90–98, 2013.

[87] B. Namaki Araghi, R. Krishnan, and H. Lahrmann, "Mode-specific travel time estimation using bluetooth technology," *J. Intell. Transp. Syst.*, vol. 20, no. 3, pp. 219–228, 2016.

[88] D. Geronimo, A. M. Lopez, A. D. Sappa, and T. Graf, "Survey of pedestrian detection for advanced driver assistance systems," *IEEE Trans. Pattern Anal. Mach. Intell.*, no. 7, pp. 1239–1258, 2009.

[89] P. Tate, Z. Sapp, L. Wells, L. Yang, and J. Meier, "Advanced accessible pedestrian system for signalized traffic intersections." Google Patents, 2018.

[90] D. Halperin, W. Hu, A. Sheth, and D. Wetherall, "Predictable 802.11 packet delivery from wireless channel measurements," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 41, no. 4, pp. 159–170, 2011.

[91] W. Wang, A. X. Liu, and M. Shahzad, "Gait recognition using wifi signals," in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 2016, pp. 363–373.

[92] W. Xi *et al.*, "Electronic frog eye: Counting crowd using wifi," in *IEEE INFOCOM 2014-IEEE Conference on Computer Communications*, 2014, pp. 361–369.

[93] S. Di Domenico, M. De Sanctis, E. Cianca, and G. Bianchi, "A trained-once crowd counting method using differential wifi channel state information," in *Proceedings of the 3rd International on Workshop on Physical Analytics*, 2016, pp. 37–42.

[94] Q. Pu, S. Gupta, S. Gollakota, and S. Patel, "Whole-home gesture recognition using wireless signals," in *Proceedings of the 19th annual international conference on Mobile computing & networking*, 2013, pp. 27–38.

[95] "FLIR Acyclica RoadTrend." [Online]. Available: https://www.flir.com/support/products/roadtrend#Specifications.

[96] W. Xue, W. Qiu, X. Hua, and K. Yu, "Improved Wi-Fi RSSI measurement for indoor localization," *IEEE Sens. J.*, vol. 17, no. 7, pp. 2224–2230, 2017.

[97] M. Quan, E. Navarro, and B. Peuker, "Wi-fi localization using rssi fingerprinting," 2010.

[98] S. Yiu, M. Dashti, H. Claussen, and F. Perez-Cruz, "Wireless RSSI fingerprinting localization," *Signal Processing*, vol. 131, pp. 235–244, 2017.

[99] 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.*, 2019.

[100] 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, 2016.

[101] 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. its Appl.*, vol. 489, pp. 78–93, 2018.

[102] 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.

[103] M. Kotaru, K. Joshi, D. Bharadia, and S. Katti, "Spotfi: Decimeter level localization using wifi," in *Proceedings of the 2015 ACM Conference on Special Interest Group on Data Communication*, 2015, pp. 269–282.

[104] R. Zhou, X. Lu, P. Zhao, and J. Chen, "Device-free presence detection and localization with SVM and CSI fingerprinting," *IEEE Sens. J.*, vol. 17, no. 23, pp. 7990–7999, 2017.

[105] Z. Chen, L. Zhang, C. Jiang, Z. Cao, and W. Cui, "WiFi CSI based passive human activity recognition using attention based BLSTM," *IEEE Trans. Mob. Comput.*, vol. 18, no. 11, pp. 2714–2724, 2018.

[106] D. Zhang, H. Wang, Y. Wang, and J. Ma, "Anti-fall: A non-intrusive and real-time fall detector leveraging CSI from commodity WiFi devices," in *International Conference on Smart Homes and Health Telematics*, 2015, pp. 181–193.

[107] B. A. Forouzan, *TCP/IP protocol suite*. McGraw-Hill, Inc., 2002.

[108] J. Postel, "User datagram protocol," *Isi*, 1980.

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

[110] V. Tzivaras, *Raspberry Pi Zero W Wireless Projects*. Packt Publishing Ltd, 2017.

[111] E. ESRI, “Shapefile Technical Description: An ESRI White Paper, 1998.” ESRI http://www. esri. com/library/whitepapers/pdfs/shapefile. pdf, accessed May, 2014.

[112] M. Haklay and P. Weber, “Openstreetmap: User-generated street maps,” *IEEE Pervasive Comput.*, vol. 7, no. 4, pp. 12–18, 2008.

[113] W. Xu *et al.*, “Internet of vehicles in big data era,” *IEEE/CAA J. Autom. Sin.*, vol. 5, no. 1, pp. 19–35, 2017.

[114] J. C. Bezdek, R. Ehrlich, and W. Full, “FCM: The fuzzy c-means clustering algorithm,” *Comput. Geosci.*, vol. 10, no. 2–3, pp. 191–203, 1984.

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

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

[117] R. Xu and D. C. Wunsch, “Survey of clustering algorithms,” 2005.

[118] 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.

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

[120] 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 networks Learn. Syst.*, vol. 30, no. 2, pp. 601–614, 2018.

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

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

[123] Z. Pu, Z. Li, R. Ke, X. Hua, and Y. Wang, “Evaluating the Non-Linear Correlation between Vertical Curve Features and Crash Frequency on Highways using Random Forests,” *J. Transp. Eng. Part A Syst.*, 2020.

[124] 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, 2018.

[125] L. Breiman, *Classification and regression trees*. Routledge, 2017.

[126] D. Reynolds, “Gaussian mixture models,” *Encycl. biometrics*, pp. 827–832, 2015.

[127] 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, 1998.

[128] H. Attias, “A variational baysian framework for graphical models,” in *Advances in neural information processing systems*, 2000, pp. 209–215.

[129] D.-S. Lee, J. J. Hull, and B. Erol, “A Bayesian framework for Gaussian mixture background modeling,” in *Proceedings 2003 International Conference on Image Processing (Cat. No. 03CH37429)*, 2003, vol. 3, pp. III--973.

[130] J. Wu, H. Xiong, and J. Chen, “Adapting the right measures for k-means clustering,” in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2009, pp. 877–886.

[131] Y. Liu, Z. Li, H. Xiong, X. Gao, and J. Wu, “Understanding of internal clustering validation measures,” in *2010 IEEE International Conference on Data Mining*, 2010, pp. 911–916.

[132] Y. Zhao and G. Karypis, “Evaluation of hierarchical clustering algorithms for document datasets,” in *Proceedings of the eleventh international conference on Information and knowledge management*, 2002, pp. 515–524.

[133] P.-N. Tan, *Introduction to data mining*. Pearson Education India, 2018.

[134] P. J. Rousseeuw, “Silhouettes: a graphical aid to the interpretation and validation of cluster analysis,” *J. Comput. Appl. Math.*, vol. 20, pp. 53–65, 1987.

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

[136] D. L. Davies and D. W. Bouldin, “A cluster separation measure,” *IEEE Trans. Pattern Anal. Mach. Intell.*, no. 2, pp. 224–227, 1979.

[137] D. Analysis, “Clustering Quality Assessment,” 2011.

[138] J. Han and C. Science, “CS 412 Intro . to Data Mining,” 2017.

[139] 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.

[140] L. Mikkelsen, R. Buchakchiev, T. Madsen, and H. P. Schwefel, “Public transport occupancy estimation using WLAN probing,” in *2016 8th International Workshop on Resilient Networks Design and Modeling (RNDM)*, 2016, pp. 302–308.

[141] 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, 2019.

[142] 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 *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers*, 2019, pp. 781–787.

[143] U. Mehmood, I. Moser, P. P. Jayaraman, and A. Banerjee, "Occupancy Estimation using WiFi: A Case Study for Counting Passengers on Busses," in *2019 IEEE 5th World Forum on Internet of Things (WF-IoT)*, 2019, pp. 165–170.

[144] R. Krishnapuram and J. M. Keller, "A possibilistic approach to clustering," *IEEE Trans. fuzzy Syst.*, vol. 1, no. 2, pp. 98–110, 1993.

[145] B. Kulis, S. Basu, I. Dhillon, and R. Mooney, "Semi-supervised graph clustering: a kernel approach," *Mach. Learn.*, vol. 74, no. 1, pp. 1–22, 2009.

[146] S. Basu, M. Bilenko, and R. J. Mooney, "A probabilistic framework for semi-supervised clustering," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2004, pp. 59–68.

[147] S. Basu, A. Banerjee, and R. Mooney, "Semi-supervised clustering by seeding," in *In Proceedings of 19th International Conference on Machine Learning (ICML-2002*, 2002.

[148] A. Likas, N. Vlassis, and J. J. Verbeek, "The global k-means clustering algorithm," *Pattern Recognit.*, vol. 36, no. 2, pp. 451–461, 2003.

[149] K. Wagstaff, C. Cardie, S. Rogers, S. Schrödl, and others, "Constrained k-means clustering with background knowledge," in *Icml*, 2001, vol. 1, pp. 577–584.

[150] L. Tari, C. Baral, and S. Kim, "Fuzzy c-means clustering with prior biological knowledge," *J. Biomed. Inform.*, vol. 42, no. 1, pp. 74–81, 2009.

[151] L. Davies and U. Gather, "The identification of multiple outliers," *J. Am. Stat. Assoc.*, vol. 88, no. 423, pp. 782–792, 1993.

[152] S. Palipana, D. Rojas, P. Agrawal, and D. Pesch, "FallDeFi: Ubiquitous fall detection using commodity Wi-Fi devices," *Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technol.*, vol. 1, no. 4, p. 155, 2018.

[153] R. G. Brown, P. Y. C. Hwang, and others, *Introduction to random signals and applied Kalman filtering*, vol. 3. Wiley New York, 1992.

[154] C. Vonesch, T. Blu, and M. Unser, "Generalized Daubechies wavelet families," *IEEE Trans. Signal Process.*, vol. 55, no. 9, pp. 4415–4429, 2007.

[155] T. S. Rappaport and others, *Wireless communications: principles and practice*, vol. 2. prentice hall PTR New Jersey, 1996.

[156] D. Wu, D. Zhang, C. Xu, Y. Wang, and H. Wang, "WiDir: walking direction estimation using wireless signals," in *Proceedings of the 2016 ACM international joint conference on pervasive and ubiquitous computing*, 2016, pp. 351–362.

[157] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *European conference on computer vision*, 2016, pp. 630–645.

[158] D. Halperin, W. Hu, A. Sheth, and D. Wetherall, "Tool release: Gathering 802.11 n traces with channel state information," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 41, no. 1, p. 53, 2011.

[159] B. S. da Silva, G. T. Laureano, A. S. Abdallah, and K. V. Cardoso, "Widmove: Sensing movement direction using ieee 802.11 n interfaces," in *2018 IEEE Canadian Conference on Electrical & Computer Engineering (CCECE)*, 2018, pp. 1–4.

[160] A. Lindgren and F. Chen, "State of the art analysis: An overview of advanced driver assistance systems (adas) and possible human factors issues," *Hum. factors Econ. Asp. Saf.*, vol. 38, p. 50, 2006.

[161] Z. Wang and K. Jia, "Frustum convnet: Sliding frustums to aggregate local point-wise features for amodal 3d object detection," *arXiv Prepr. arXiv1903.01864*, 2019.

[162] E. Arnold, O. Y. Al-Jarrah, M. Dianati, S. Fallah, D. Oxtoby, and A. Mouzakitis, "A survey on 3d object detection methods for autonomous driving applications," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 10, pp. 3782–3795, 2019.

[163] W. Song, G. Xiong, and H. Chen, "Intention-aware autonomous driving decision-making in an uncontrolled intersection," *Math. Probl. Eng.*, vol. 2016, 2016.

[164] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 3354–3361.

[165] N. Jayaweera, N. Rajatheva, and M. Latva-aho, "Autonomous driving without a burden: View from outside with elevated lidar," in *2019 IEEE 89th Vehicular Technology Conference (VTC2019-Spring)*, 2019, pp. 1–7.

[166] Y. Wang, W.-L. Chao, D. Garg, B. Hariharan, M. Campbell, and K. Q. Weinberger, "Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8445–8453.

[167] R. Faster, "Towards real-time object detection with region proposal networks," *Adv. Neural Inf. Process. Syst.*, p. 9199, 2015.

[168] J. Redmon and A. Farhadi, “Yolov3: An incremental improvement,” *arXiv Prepr. arXiv1804.02767*, 2018.

[169] X. Chen, K. Kundu, Z. Zhang, H. Ma, S. Fidler, and R. Urtasun, “Monocular 3d object detection for autonomous driving,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2147–2156.

[170] M. Winters, M. Brauer, E. M. Setton, and K. Teschke, “Built environment influences on healthy transportation choices: bicycling versus driving,” *J. urban Heal.*, vol. 87, no. 6, pp. 969–993, 2010.

[171] B. E. Saelens and S. L. Handy, “Built environment correlates of walking: a review,” *Med. Sci. Sports Exerc.*, vol. 40, no. 7 Suppl, p. S550, 2008.

[172] L. Zhang, J. Hong, A. Nasri, and Q. Shen, “How built environment affects travel behavior: A comparative analysis of the connections between land use and vehicle miles traveled in US cities,” *J. Transp. Land Use*, vol. 5, no. 3, pp. 40–52, 2012.

[173] M. Solutions, “Analysis of IOS 8 MAC Randomization on Locationing,” 2014.

[174] C. Matte, “Wi-Fi tracking: Fingerprinting attacks and counter-measures,” 2017.

[175] I. T. Standards *et al.*, “Standard Group MAC Addresses : A Tutorial Guide,” vol. 10039, no. Llc, pp. 1–4.