FINAL REPORT

Contract No. BED30-977-03

Evaluating and Validating Technology Options for Estimating Transit Vehicle Occupancy in Real Time

Prepared for:
Research Center
Florida Department of Transportation
605 Suwannee Street, M.S. 30
Tallahassee, FL 32399-0450

Project Manager:
Gabe Matthews and Chris Wiglesworth
Public Transit Office
Florida Department of Transportation
605 Suwannee Street, Tallahassee, FL 32399
mailto:gabrielle.matthews@dot.state.fl.us

Prepared by:
Florida State University

Principal Investigator: Yanshuo Sun, Ph.D.

Co-Principal Investigators: Qianwen Guo, Ph.D.
Zhaomiao Guo, Ph.D.
Albert Gan, Ph.D.

Graduate Research Assistants: Esther Ndemasi Mneney
Md Rakibul Alam
Nattakarn Surangsrirout

January 2024

Disclaimer

Disclaimer: The opinions, findings, and conclusions expressed in this publication are those of the author(s) and not necessarily those of the Florida Department of Transportation.

Metric Conversion Table

SI* (MODERN METRIC) CONVERSION FACTORS				
APPROXIMATE CONVERSIONS TO SI UNITS				
SYMBOL	**WHEN YOU KNOW**	**MULTIPLY BY**	**TO FIND**	**SYMBOL**
		LENGTH		
in	inches	25.4	millimeters	mm
ft	feet	0.305	meters	m
yd	yards	0.914	meters	m
mi	miles	1.61	kilometers	km
		AREA		
in^2	square inches	645.2	square millimeters	mm^2
ft^2	square feet	0.093	square meters	m^2
yd^2	square yard	0.836	square meters	m^2
ac	acres	0.405	hectares	ha
mi^2	square miles	2.59	square kilometers	km^2
		VOLUME		
fl oz	fluid ounces	29.57	milliliters	mL
gal	gallons	3.785	liters	L
ft^3	cubic feet	0.028	cubic meters	m^3
yd^3	cubic yards	0.765	cubic meters	m^3
	NOTE: volumes greater than 1000 L shall be shown in m^3			
		MASS		
oz	ounces	28.35	grams	g
lb	pounds	0.454	kilograms	kg
T	short tons (2000 lb)	0.907	megagrams (or "metric ton")	Mg (or "t")
APPROXIMATE CONVERSIONS FROM SI UNITS				
SYMBOL	**WHEN YOU KNOW**	**MULTIPLY BY**	**TO FIND**	**SYMBOL**
		LENGTH		
mm	millimeters	0.039	inches	in
m	meters	3.28	feet	ft
m	meters	1.09	yards	yd
km	kilometers	0.621	miles	mi
		AREA		
mm^2	square millimeters	0.0016	square inches	in^2
m^2	square meters	10.764	square feet	ft^2
m^2	square meters	1.195	square yards	yd^2
ha	hectares	2.47	acres	ac
km^2	square kilometers	0.386	square miles	mi^2
		VOLUME		
mL	milliliters	0.034	fluid ounces	fl oz
L	liters	0.264	gallons	gal
m^3	cubic meters	35.314	cubic feet	ft^3
m^3	cubic meters	1.307	cubic yards	yd^3
		MASS		
g	grams	0.035	ounces	oz
kg	kilograms	2.202	pounds	lb
Mg (or "t")	megagrams (or "metric ton")	1.103	short tons (2000 lb)	T

*SI is the symbol for the International System of Units. Appropriate rounding should be made to comply with Section 4 of ASTM E380.

Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
4. Title and Subtitle Evaluating and Validating Technology Options for Estimating Transit Vehicle Occupancy in Real Time		**5. Report Date** October 2023
		6. Performing Organization Code Florida State University
7. Author(s) Yanshuo Sun, Qianwen Guo, Zhaomiao Guo, Albert Gan, Esther Mneney, Md Rakibul Alam, Nattakarn Surangsrirout.		**8. Performing Organization Report No.**
9. Performing Organization Name and Address FAMU-FSU College of Engineering, Florida State University 2525 Pottsdamer St, Tallahassee, FL 32310		**10. Work Unit No. (TRAIS)**
		11. Contract or Grant No. BED30-977-03
12. Sponsoring Agency Name and Address Florida Department of Transportation 605 Suwannee Street, Tallahassee, FL 32399		**13. Type of Report and Period Covered** Final Report May 2022 – January 2024
		14. Sponsoring Agency Code 99700-3596-119

15. Supplementary Notes
Gabe Matthews and Chris Wiglesworth have both served as the project managers.

16. Abstract
The primary goal of this project was to evaluate and validate various technologies for collecting transit vehicle occupancy information in real time. The specific objectives include the following: (1) Identify a list of potential technology alternatives for transit vehicle occupancy estimation by scanning the academic literature, news, and technical reports, as well as interviewing transit practitioners; (2) Evaluate all potential technologies from a technical perspective (e.g., measurement accuracy, latency, reliability, level of automation, ease of implementation and use, and maintenance needs) involving both hardware and software, and a nontechnical perspective (e.g., cost efficiency, privacy impact, and user acceptance); (3) Develop detailed documentation of promising technologies, covering technical capabilities, privacy, barriers to implementation, risks, cost, and possible vendors; (4) Conduct pilot studies and validate selected technologies in three representative transit systems in Florida; (5) Complete a technical report on selecting and implementing vehicle occupancy estimation technology; (6) Develop and deliver a webinar to disseminate the project findings. After pilot studies in three transit systems in Florida, we noted that MAC (Media Access Control) address randomization in Wi-Fi probing presented a significant challenge for estimating vehicle occupancy accurately. However, a data-driven learning algorithm, after training and testing on substantial data, can achieve a satisfactory predictive performance, such as achieving an R^2 value of 0.84 in the case study of Route Evergreen in Tallahassee, FL. The predictive performance may be even higher when continuous data are available, and noise in training data is better filtered.

17. Key Word	18. Distribution Statement
Bus occupancy, Technology Assessment, Wi-Fi Probing, MAC address randomization, Data-driven Prediction, Case Studies	No restrictions

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	92	

Acknowledgements

The research team would like to thank Gabrielle Matthews, David Sherman, and Chris Wiglesworth of the Florida Department of Transportation (FDOT) Public Transit Office for their assistance and feedback throughout the project. The team is grateful for the support from the transit agencies StarMetro in Tallahassee, Lynx Orlando, and Miami-Dade Transit in Miami, Florida, for every individual who attended the meetings and for their cooperation throughout the pilot studies.

Executive Summary

Despite the significance of real-time transit vehicle occupancy in evaluating the travel experience of riders and measuring the operational efficiency of transit services, it remains a challenge to accurately estimate transit vehicle occupancy, especially in a real-time and cost-effective manner. This is in stark contrast to real-time transit vehicle location information, which is already widely provided by transit agencies throughout the U.S. The primary goal of this project was to evaluate and validate various technologies for collecting transit vehicle occupancy information in real time. To achieve this goal, the following research tasks have been conducted.

First, the project team conducted a full scan of potential technologies for estimating transit vehicle occupancy (both rail and bus) by conducting detailed and in-depth reviews of the relevant literature and practice. We identified the following technology options for the real-time estimation of transit vehicle occupancy: automatic passenger counter, automatic fare collection, crowdsourcing, onboard survey, mobile ticketing, temperature sensing, Wi-Fi, Bluetooth, cellular, optical and thermal cameras, lidar, and ultrasonic sensors.

Then, an in-depth evaluation covering both technical and nontechnical factors was performed for each technology identified earlier. Technical factors included measurement accuracy, latency, reliability, level of automation, ease of implementation and use, and maintenance needs. Non-technical factors included cost efficiency, privacy impact, and user acceptance.

Next, considering the technical and non-technical evaluations of each technology and the technology availability in three Florida transit systems (StarMetro, Lynx, and Miami-Dade Transit), Wi-Fi was identified as a promising technology, among others. Notably, Wi-Fi probing is promising because it requires little hardware, data is available in real time, and no privacy issues exist.

To evaluate the selected promising technologies in the field, pilot studies were conducted at three locations. With inputs from transit agencies, the research team selected three transit routes or lines for pilot studies at three different locations. Specifically, route Evergreen was selected for StarMetro in Tallahassee; route 104 was selected for Lynx in Orlando; and two Metro Mover loops, namely Omni Loop and Brickell Loop, were selected for Miami-Dade Transit.

One technical challenge arising from MAC (Media Access Control) address randomization in Wi-Fi probing was noted. When MAC addresses are randomized, it is impossible to uniquely identify MAC addresses for estimation purposes. Therefore, the conventional approach of counting the number of unique MAC addresses as a proxy for the number of passengers will no longer work. We therefore proposed that other types of Wi-Fi frames or methods must be considered to explore the best potential of Wi-Fi data in estimating bus occupancy. We then identified many possible features and developed a data-driven approach for estimating vehicle occupancy. After feature engineering and hyperparameter tuning, satisfactory predictive results were obtained for the collected Wi-Fi frames in Tallahassee, Florida. For instance, the highest R^2 value that has been achieved is 0.84, which is considerably high. Comparable results were also achieved for data collected from Orlando and Miami. Therefore, this project demonstrated Wi-Fi frames can be used to estimate vehicle occupancy even though MAC addresses are randomized.

One shortcoming in the current analysis is that the data collection lasted for two weeks. When data are collected over a longer period, preferably continuously over time, the predictive performance can be further improved. Therefore, a data-driven learning algorithm that has been trained and tested on large-scale Wi-Fi frames can be used for transit vehicle occupancy estimation in practice.

Table of Contents

List of Figures

List of Tables

CHAPTER 1. INTRODUCTION

1.1 Background

Transit system planners and operators need transit vehicle occupancy data to enhance service delivery, evaluate the comparative effectiveness of transit operations, and evaluate the in-vehicle travel experience of transit passengers. The provision of real-time transit vehicle occupancy information is also crucial in assisting public transit passengers in making well-informed travel choices, including deciding whether to utilize the incoming bus or wait for the next available bus. Unlike the real-time transit vehicle arrival information that is already widely provided by transit agencies across the United States, real-time transit vehicle occupancy is rarely available.

Traditional methods for determining the occupancy of transit vehicles commonly involve the utilization of manual surveys, data from automatic passenger counters (APCs), or data from automated fare collecting (AFC) systems. Each of these strategies has its disadvantages. The process of conducting manual surveys requires a significant amount of labor and is consequently associated with high costs. In a considerable number of cases, manual surveys are limited in their applicability considering the small sample size, and the data collected are also prone to human error, especially in areas with high levels of boarding and alighting. Although APCs have the potential to offer relatively precise passenger counts, their cost remains high for many transit agencies. Consequently, many agencies can only afford to implement APCs on partial routes. Furthermore, it should be noted that APCs are not designed to be utilized for real time data applications. Furthermore, the usage of AFC systems to gather occupancy data is limited to rail transport systems that are fully controlled and require ticket validation at both the points of entry and exit. Nevertheless, it is worth noting that only a limited number of transit systems in the United States possess AFC systems that effectively record both entry and exit transactions. As a result, the task of accurately determining the number of occupants in transit vehicles, especially in real-time and with little expense, remains a major challenge.

The utilization of emerging information and communication technologies, including mobile ticketing, Wi-Fi, temperature monitoring, radio-frequency identification, and crowdsourcing, provides potential in enhancing the estimation of real time transit vehicle occupancy. Upon conducting a comprehensive analysis of the literature, it is evident that there is a lack of research that has assessed and substantiated the effectiveness of these technologies in accurately estimating transit vehicle occupancy. Thus, this project assessed and validated prospective information and communication technologies that have the potential to accurately determine the transit vehicles occupancy in real time. This study assessed a range of Florida transit agencies, taking into consideration their diverse characteristics such as size, service regions, types of services (bus and rail), and information technology infrastructure support. It is important to acknowledge that the compatibility of a particular technology may differ among transit agencies, as what works effectively for one may not be the preferred choice for another.

This research has the potential to advance the current practice of collecting, validating, and disseminating real-time transit vehicle occupancy data. The obtained vehicle occupancy data can assist system operators in comprehending how transit capacity is utilized by location, route, travel orientation, and time. Once real-time vehicle occupancy data are available, transit agencies will be able to integrate these data into their mobile apps and station signage so that passengers can readily access it. The additional travel information, in addition to the real-time vehicle

arrival, provides a more comprehensive view of the transit service, thereby possibly attracting more travelers to the transit system. This report can also assist transit agencies that are not a part of this project's pilot study in selecting their preferred vehicle occupancy data collection technology.

1.2 Research Overview and Structure

In this section we will provide a general overview of the research project and the structure of this report. The goal of this project is to test and verify several technologies for gathering and estimating real-time transit vehicle occupancy. The objectives include but are not limited to the following:

- Identify a list of prospective technology options for transit vehicle occupancy estimation by examining academic literature, news, and technical reports.
- Evaluate all potential technologies from both a technical (e.g., measurement accuracy, latency, reliability, level of automation, ease of implementation and use, and maintenance requirements) and non-technical (e.g., cost efficiency, privacy impact, and user acceptance) standpoint.
- Create extensive documentation of viable technologies, including their technical capabilities, privacy issues, implementation difficulties, risks costs, and potential vendors.
- Conduct the pilot studies and validate selected technologies in Tallahassee, Orlando and Miami, Florida.

The report is divided into six chapters. This chapter gives an overview of the project and the outline of the report.

In Chapter 2, the literature review of relevant studies is presented and the list of the identified potential technologies for the real-time estimation of transit vehicle occupancy is also given. In addition, a summary of the evaluation of the potential technologies based on their technical and technical factors is presented. Furthermore, based on the evaluation, applications of Wi-Fi technologies in public transportation are presented, followed by the identified research gaps from the review.

Wi-Fi frame collection is described in Chapter 3. The fundamentals of Wi-Fi association process are described, so are the hardware needed and the setup of the hardware. Outdoor experiments and pilot studies were conducted in Tallahassee, Miami, and Orlando; and the description of the collected data is presented.

In Chapter 4, the data driven methodology used to estimate the number of passengers is presented. Also, initial analysis of the data is presented.

The results and discussion are presented in Chapter 5. Outdoor experiments result together with the results of the data driven approach for the pilot studies conducted are presented.

Chapter 6 provides conclusions of the research project highlighting the major findings and recommended future work.

CHAPTER 2. LITERATURE REVIEW

2.1 Relevant Studies Review by Technology

Many individuals utilize public transportation on a daily basis. The quality of service is an essential consideration for passengers when deciding whether to use the bus system. Enhancements in the domains of scheduling and passenger information are heavily dependent on the availability of real-time data pertaining to bus location and in-vehicle congestion. The availability of bus location information through GPS is easily accessible and cost-effective. Obtaining accurate and up-to-date occupancy information in cost-effective and easily scalable ways represents a greater challenge. The availability of crowding information is of great significance, particularly when combined with short-term congestion predictions. This combination empowers operators to engage in dynamic scheduling practices, enabling them to manage a fluctuating load effectively. Additionally, it enhances passengers' understanding of the present and anticipated conditions of the transportation systems.

A detailed literature and practice review were conducted to identify all potential technologies for occupancy estimations. In addition to manual surveys, APCs, and AFC systems, the additional technologies that were identified include Crowdsourcing, Wi-Fi, Bluetooth, Cellular Networks, Optical, and Thermal Cameras, LiDAR, and Ultrasonic Sensors. Each of the identified technologies is evaluated based on technical and non-technical capabilities. The criteria used for the technical evaluation are as follows:

a) Measurement accuracy: how close the technology's measurements are to the ground truth value.
b) Latency: how long it takes to provide occupancy estimation findings after processing and computing.
c) Reliability: the capacity to work successfully despite real-world disturbances (e.g., weather changes, vehicle brightness.
d) Level of automation: the extent of intervention from vehicle operators and other personnel needed to keep the system working.
e) Implementation and use: the projected effort and complexity of using a proposed technology.
f) Maintenance: support staff attempt to restore system functionality. The above criteria apply to both the software and hardware sides. Scholarly literature, technical standards, and pertinent reports were used to evaluate identified technologies.

After the technical evaluations, a non-technical analysis was conducted. Technology may have all the desirable technical features, but it may not be selected for deployment due to non-technical factors. The non-technical evaluation involved re-evaluating those technologies with good technical capabilities through the following aspects:

a) Cost: the estimated amount of capital and operating investments required.
b) Privacy: protecting personal information (both vehicle operators and riders) from being observed by unauthorized parties,
c) Public acceptance: attitudes from riders towards the occupancy estimation technologies
d) Transit agency acceptance: perception of usefulness and ease of maintaining these technologies,

e) Vendor acceptance: vendors' incentives for developing the required systems for the transit agencies.

2.1.1 Automatic Passenger Counter

Automatic Passenger Counters (APC) are special devices installed on transit vehicles (i.e., buses and trains) to keep track of the boarding and alighting movements of passengers. There are multiple types of APCs depending on the underlying technology. Infrared sensors can form a light barrier to detect the passage of people; pressure-sensitive switches (e.g., mechanical treadle mats) can be activated as passenger weight is applied on them; load sensors installed on suspensions of trains can also be used to estimate the number of passengers (Darsena et al., 2022; Nielsen et al., 2014).

An example of a study that involves APC was conducted by Khomchuk et al. (2018). Since most train cars have installed electronic weighing sensors (E&T, 2020), the researchers used such weight data from the Copenhagen rail network to infer the train load. They indicated that the weight sensors could provide comparable passenger counts as infrared sensors. They also developed an algorithm to predict the crowding level for each train car by combining the weight data with some historical passenger data. Table 2-1 provides a list of some relevant studies involving APC technology.

Table 2-1 APC Technology-Related Studies

Study	Research Topic	Mode	Data Classification	Region
Rakebrandt (2007)	The concept and applications of APC systems.	Bus/Rail	N/A	New York City, US
Nielsen et al. (2014)	A passenger counting technique that used the weighing systems installed in the majority of current trains to control brakes.	Rail	Real Time	Denmark
Khomchuk et al. (2018)	Crowding level information to passengers waiting at arriving station through APC data.	Rail	Real Time	Massachusetts, US
Traunmueller et al. (2018)	A low-cost approach to APC by using accelerometer measurement for estimating passenger loading.	Rail	Real Time	Massachusetts, US
Jenelius (2020)	Crowding prediction problem based on real-time load data and examined the performance of numerous data-driven prediction algorithms.	Rail	Real Time	Stockholm, Sweden.

Study	Research Topic	Mode	Data Classification	Region
Murdan et al. (2020)	Design, development, and testing of an Internet of Things-based system that uses an array of sensors to detect events in the vehicle and display real-time vehicle occupancy.	Bus	Real Time	Mauritius
Grgurević et al. (2022)	An overview and study of the most widely employed methods, technology, and designs for on board passenger counting in public urban transport vehicles.	Bus/Rail	Non-Real Time	N/A

2.1.2 Automatic Fare Collection

The Automatic Fare Collection (AFC) system is an automated ticketing system being adopted by an increasing number of transit agencies. AFC typically consists of ticket vending machines, ticket checking machines, and automatic gate machines. It is noted that the primary purpose of AFC is for fare collection although AFC-generated data can be used to estimate ridership or vehicle occupancy (Asim et al., 2022). It is also common to jointly analyze APC and AFC data to comprehensively understand passenger activities. An example application of the AFC data is provided by Jang (2010), who conducted travel time and transfer pattern analyses using AFC data from Seoul, South Korea. As both the trip origin and destinations were recorded by the AFC system in the study, the number of boarding and alighting passengers for a vehicle at each stop could be collected. Sun & Schonfeld (2016) used train schedules and AFC data from the Beijing Subway to assign passengers to train trajectories, which further yielded the number of passengers taking each train. In this assignment, the entry time, entry station, exit time, and exit stations were all available. Table 2-2 lists some related studies involving AFC.

Table 2-2 AFC Technology-Related Studies

Study	Research Topic	Mode	Data Classification	Region
Cui (2006)	Development of an algorithm to estimate bus passenger trip origin-destination matrix (OD matrix) using data from the Automatic Data Collection systems such AFC.	Bus	Non-Real Time	Massachusetts, US

Study	Research Topic	Mode	Data Classification	Region
Jang (2010)	The potential applications of automatically gathering smart card data and explored the usage of AFC data for transportation planning applications.	Rail	Non-Real Time	Massachusetts, US
Pelletier et al. (2011)	The examination of smart card data usage in the context of public transportation. The study addressed data utilization at management's strategic, tactical, and operational levels (ridership statistics and performance indicators).	Bus	Non-Real Time	Montreal, Canada
Sun & Schonfeld (2016)	AFC data were utilized to construct a schedule-based passenger path-choice prediction model for a multi-operator rail transport network.	Rail	Non-Real Time	Beijing, China
Yang et al. (2021)	Short-term passenger volume forecasting for urban train networks using smart-card data and deep learning	Rail	Non-Real Time	Beijing, China
Özgün et al. (2022)	Using reverse direction boarding to estimate alighting counts for public transportation vehicle occupancy levels.	Bus	Non-Real Time	Antalya, Turkey

2.1.3 Crowdsourcing

The Texas A&M Transportation Institute (2022) website defines crowdsourcing as an innovative approach to engaging the public in decision-making. This strategy enables many individuals to act as information collectors and broadcasters to others in the crowd. Several research studies have explored the concept of crowdsourcing to collect and provide real-time transit information, as summarized in Table 2-3.

Table 2-3 Crowdsourcing-Related Studies.

Study	Research Topic	Mode	Data Classification	Region
Stirling (2012)	Tiramisu Transit App that uses crowdsourced information from riders to provide the predicted number of bus arrival times and occupancies.	Bus	N/A	Syracuse, New York State, US
Tejas et al. (2015)	Crowd Sourcing to provide real-time information about buses.	Bus	Real Time	India

Study	Research Topic	Mode	Data Classification	Region
Vemula et al. (2015)	Improving the usability of public transportation through crowdsourcing.	Bus	Real Time	India
Chaudhary et al. (2016)	Occupancy prediction using crowdsourced data from smartphones.	Bus	Real Time	India
Haywood et al. (2017)	Investigation of the cost of public transportation crowding in terms of passenger welfare.	Bus	Non-Real Time	Paris, France
Mukheja et al. (2017)	Real-time positioning method using crowdsourced information from a smartphone.	Bus	Real Time	India
Lazo (2017)	Commuters take charge of their commute through crowdsourced information from app-based services.	Bus	N/A	Washington, US
Wanek-Libman (2020)	Sharing crowding information in transit vehicles through crowdsourcing.	Bus	N/A	Illinois, US
Texas A&M Transportation Institute (2022)	Description of what crowdsourcing is and its impact on the community.	Rail/ Bus	Non-Real Time	Texas, US

As an example, Tejas et al. (2015) described a system that gathered bus occupancy information through crowdsourcing. The system used consisted of an Android module, a cloud module, and a QR code module. In this system, passengers scanned the QR codes painted on a bus to upload occupancy information to the cloud. Chaudhary et al. (2016) proposed a similar system and tested it with help from four students traveling in the same bus in Chandigarh, India. The authors provided the crowding information in five-level scale ranging from empty (less than 10 occupied seats) to standing (more than 50 passengers on-board). Chaudhary et al. (2016) stated that certain users had reported erroneous data, as the ground truth data were collected manually and used to validate the information.

A transportation reporter (Lazo, 2017) from the Washington Post described how mobile apps were used to collect travel preferences (such as route, drop-off point, and departure time) from commuters to customize the commuter bus services. The reporter presented that Chariot, a commuter shuttle service provider, allowed users to pitch their preferred new route. However, Lazo (2017) did not mention any vehicle occupancy information collected or shared through crowdsourcing. In addition, Kim et al. (2019) performed an online survey to reduce overcrowding by investigating the effects of occupancy information designed to promote passenger behavior modification. The majority of participants, roughly 93.3%, indicated that the information was useful and that they were willing to shift to a less crowded carriage. Nevertheless, they indicated that further information beyond occupancy data such the number of empty seats would be required.
The unique advantage of crowdsourcing is that no major capital investments in hardware are

needed, as passengers can report vehicle occupancy information voluntarily. The shortcoming of crowdsourcing lies in the quality of data submitted by passengers. Data validation and error correction is necessary.

2.1.4 Cellular

In a cellular network (or called wireless network), users with portable transceivers (e.g., mobile phones) can communicate with other users at a different location through fixed-location transceivers (also called cellular radio towers or cell sites). The radio signals exchanged between those transceivers can be used to track mobile devices and their users (Liu et al. 2014). Darsena et al. (2022) reported that cellular signals could be used for crowd-counting. Table 2-4 displays representative studies involving cellular data.

Darsena et al. (2022) provided an example of a study that used radio wave signals of cellular communication and DL (Deep Learning) techniques for crowd estimation. The results of the study were 78% accurate. However, the authors indicated several setbacks associated with using cellular data for passenger estimation; these complications include the need to collect information from separate mobile phone operators, privacy issues, and the inability to distinguish passengers from the public. The authors added that the incoming 5G technology that utilizes very small cells would permit more precise monitoring, such as bus stops.

There are a few challenges in estimating vehicle occupancy using cellular data. First, cellular data are owned by mobile carriers, such as Verizon or AT&T, rather than public transit agencies. Such cellular location data have to be acquired from multiple carriers. Second, cellular data need to be processed (e.g., removing data for those mobile users who are classified as non-transit passengers), which can take time. Third, the latency is high, because such cellular data are not available to transit operators in real-time.

Table 2-4 Cellular Network Technology-Related Studies.

Study	Research Topic	Mode	Data Classification	Region
Aguilera et al. (2013)	Cellular data for passenger flow measurement in an underground transit system.	Rail		France
Ramachandran (2013)	Methods for determining and quantifying the relationship between the number of cell phone users and the overall number of persons in a given location.	Transit/ Non-Transit	Non-Real Time	New Jersey, US
Shibata & Yamamoto (2019)	Recommendation of a new crowd density observation technique for monitoring people flow. The suggested method measures the signal	Non-Transit	N/A	Japan

Study	Research Topic	Mode	Data Classification	Region
	intensity of cellular communication radio waves.			
Barbour et al. (2019)	A unique approach for estimating building occupancy at an unprecedented scale using enormous amounts of passively gathered mobile phone data.	Non-Transit	Non-Real Time	Boston, Massachusetts, US

2.1.5 Wi-Fi

Wi-Fi is a network technology used for local area networking and internet access. Wi-Fi allows neighboring digital devices to exchange data via radio waves. Wi-Fi services are provided by public transportation or other passenger transport companies in order to improve the travel experience of their customers or increase ridership. Because many commuters carry Wi-Fi devices (e.g., 80%, according to Pu et al. (2021)) and a rising number of transit operators install Wi-Fi access points on their vehicles, the idea of predicting vehicle occupancy with Wi-Fi data has been investigated by numerous researchers in the literature.

Estimating transit vehicle occupancy using Wi-Fi probe data is a promising alternative because it requires little technology and can potentially be accessed in real-time. It, like other device-based passenger counting approaches, suffers from overestimation (when a single user is carrying numerous devices) and underestimating (when a user is not carrying any devices). Those overestimation and underestimating issues, however, can be overcome with proper data filtration and re-scaling. Table 2-5 displays the representative studies involving Wi-Fi.

Table 2-5 Wi-Fi Technology-Related Studies.

Study	Research Topic	Mode	Data Classification	Region
Kang et al. (2016)	In-vehicle Wi-Fi-based tracking system that provides various analytics for transit operators.	Bus		Madison, Wisconsin, US
Pattanusorn et al. (2016)	A real-time system for determining the position and occupancy of public transportation vehicles.	Bus	Real Time	Thailand
Jain et al. (2018)	The impact of Wi-Fi on Trains to rail passengers.	Rail	Non-Real Time	UK
Traunmueller et al. (2018)	Analyzing human mobility patterns in cities using Wi-Fi probe and locational data	Non-Transit		Manhattan, New York City, US.

Study	Research Topic	Mode	Data Classification	Region
Oransirikul et al. (2019)	Estimation of the number of passengers by evaluating signals from their Wi-Fi devices and identifying them using a real-time filtering method as originating from passenger or non-passenger devices.	Bus	Real Time	Japan
Mehmood et al. (2019)	Wi-Fi-based system for occupancy estimation in buses.	Bus	Real Time	Australia
Paradeda et al. (2019)	Evaluating the feasibility of bus passenger estimates based on the detection of Wi-Fi MAC addresses of portable devices.	Bus	Non-Real Time	Brazil
Vieira et al. (2020)	A system that estimated the number of passengers in buses or metro through the user's smartphone and mathematical modeling.	Bus	Real Time	Belo Horizonte, Brazil
Nitti et al. (2020)	A Wi-Fi-based Automatic Bus Passenger Counting System (iABACUS). The iABACUS aimed to observe and analyze urban mobility by tracking passengers on public transit vehicles during their whole route without requiring the passengers to take any action.	Bus	Real Time	Italy
Ryu et al. (2020)	Practical application of Wi-Fi detecting system for predicting passengers' origin-to-destination (O/D) trip and bus stop waiting times.	Bus	Non-Real Time	Charlottesville, Virginia, US
Tang et al. (2020)	A passive Wi-Fi radar system for occupancy detection and people counting.	Non-Transit	Non-Real Time	UK
Asim et al. (2022)	Wi-Fi technology application to collect	Bus	Real Time	Canada

Study	Research Topic	Mode	Data Classification	Region
	information on passenger activity for transit service planning and management.			

2.1.6 Bluetooth

Bluetooth is a short-range wireless communication technology used to link electronic devices through radio waves. Weppner et al. (2014) presented an application of Bluetooth scanning for urban crowd monitoring. The dataset collected during a three-day event in Zurich, Switzerland involved 1,000 Bluetooth scanners, around 20,000 unique Bluetooth devices in discoverable mode, and nearly 200,000 discoveries. In the area of public transportation, Kostakos et al. (2010) also employed Bluetooth technology to estimate the origin-destination matrix in a case study in Portugal. They installed a Bluetooth scanner on the ceiling of a bus near the exit to detect passengers with their Bluetooth devices in discoverable mode. Kostakos et al. (2010) indicated that the proportion of users who set their Bluetooth devices in discoverable mode was around 7.5%. They validated their estimation results by conducting a correlation analysis between the number of Bluetooth passengers and the number of validated tickets (considered ground truth). Kostakos et al. (2010) reached the conclusion that in the future, they would implement the GPRS connection to have remote access to their data in real time. Privacy concerns were raised by the authors because their system records precise passenger location data. RFID tickets, according to Kostakos et al. (2010), represent the same threat to passenger privacy. In addition, the low penetration rate is a significant shortcoming of Bluetooth-based passenger detection. Table 2-6 lists some related studies involving Bluetooth technology.

Table 2-6 Bluetooth Technology-Related Studies.

Study	Research Topic	Mode	Data Region	Classification
Kostakos et al. (2010)	A unique, low-cost wireless system that employed commercially available Bluetooth technology and identified and recorded end-to-end passenger travels.	Bus	N/A	Maderia Island, Portugal
Weppner et al. (2014)	The possibility of employing Bluetooth scanning to monitor crowds in metropolitan environments.	Bus	Real Time	Zurich, Switzerland
Basalamah (2016)	A description of using Bluetooth low energy (BLE) tagging as an alternate way for crowd counting at bus stops; a large population carried BLE proximity tags that function as beacons and whose presence is detected by a few volunteers' smartphones.	Bus	Non-Real Time	Saudi Arabia

2.1.7 Optical and Thermal Cameras

It has been widely known that optical and thermal cameras can be used to detect and estimate the number of people in various settings (e.g., indoor or outdoor). Passengers of public transportation may be left behind on the train platform due to extreme congestion. Sipetas et al. (2020) provides an example of application in the domain of public transportation. Surveillance videos of a train station in Boston, Massachusetts were analyzed with image processing and people detection software to estimate the number of passengers who had departed the station without boarding. Using manual estimates, the estimation results were verified. It was determined that the estimation error did not exceed 10%.

Darsena et al. (2022) indicated that the majority of optical camera-based crowd counting and detection research relies on computer vision technologies where crowds are identified in videos using particular features, such as motion tracking or facial recognition. The Office of Vehicle Technologies of the US Department of Energy also funded a project led by the Chattanooga Area Regional Transportation Authority, where computer vision models were used to collect ridership data from onboard camera videos through anonymous tracking of passengers (United States Department of Energy, 2020).

In addition to optical cameras, thermal cameras can be used to detect people in low-light settings, complete darkness, or other difficult conditions (Darsena et al. 2022). Thermal cameras can be used to monitor crowding situations in both indoor and outdoor settings. Table 2-7 lists some related studies involving optical and thermal cameras.

Table 2-7 List of Relevant Studies Involving Optical and Thermal Cameras.

Study	Research Topic	Mode	Data Classification	Region
Yu et al. (2007)	A method for estimating the real-time passenger crowd flow in a bus with a complex background using image processing.	Bus	Real Time	China
Chen et al. (2008)	A system using video processing to automatically count passengers entering and exiting buses	Bus	Real Time	China
Junior et al. (2010)	An overview of crowd analysis with computer vision techniques, covering a variety of topics including people tracking, crowd density estimation, event detection, validation, and simulation.	Non-Transit	Non-Real Time	Brazil
Lengvenis et al. (2013)	An automated computer vision system for counting passengers. Four algorithms were developed to estimate the number of people using public transportation, and the benefits and challenges were examined.	Bus	Non-Real Time	Kaunas, Lithuania Europe
Hashimoto et al. (2015)	Using a two-layer laser-based scanner set to track people in a group.	Non-Transit	N/A	Japan

Study	Research Topic	Mode	Data Classification	Region
Liu et al. (2017)	A passenger counting system that combined the CNN detection model and the Spatio-temporal context model.	Bus	Non-Real Time	China
Sipetas et al. (2020)	Surveillance camera feeds analysis using image processing and object detection software to determine the number of passengers left behind at station platforms.	Rail	Non-Real Time	Boston, Massachusetts, US
Hsu et al. (2020)	An approach for estimating the number of bus passengers using deep learning in a variety of scenarios.	Bus	Real Time	Taiwan
Wei et al. (2021)	A CNN (convolutional neural network)-based network named the MP-CNN (metro platform-CNN) was designed to properly count people on metro platforms.	Rail	N/A	Zhengzhou, China

2.1.8 Mobile Ticketing

Mobile ticketing is a technology that enables users to purchase fares through their smartphones. Apanasevic & Rudmark (2021) demonstrated that ticketing is one of the core sources of information used for transit planning by most agencies. An interview conducted by the authors with a company in Denmark called Movia revealed that more passengers were switching to mobile ticketing solutions and that only 40 to 50 percent of their Movia passengers still used smartcards while the remaining use other means such as mobile ticketing. Moovit, a mobile ticketing app, provides user-reported occupancy information to help other riders plan their travels (Moovit, 2021).

Rahman et al. (2016) used mobile ticketing data to estimate the origin-destination matrix for the East River Ferry (ERF), a privately-operated ferry service in New York. With the back-end data provided by the mobile ticketing app developer, Rahman et al. (2016) estimated O-D matrices and conducted a comparison with the traditional onboard surveys. Rahman et al. (2016) concluded that mobile ticketing was a good source of data for understanding where and when passengers use ferry services.

Sørensen et al. (2019) reported that ticketing databases (including mobile ticketing) would be used for analytical purposes such as ridership, travel patterns, boarding at particular stops, and other information that will enable the providers to plan their services better. As mobile ticketing is part of the AFC system, mobile ticketing data are usually included in farebox data.

2.1.9 Hybrid

Multiple technologies can be combined to estimate transit vehicle occupancy or other metrics. For instance, Pu et al. (2019) detected both Wi-Fi and Bluetooth devices when estimating passenger flows. In an experiment conducted in Sydney, Australia, Moser et al. (2019) installed four different passenger counting technologies on a bus, two cameras, one of which was in the

front, two infrared sensors to cover both doors on the bus, five Wi-Fi sensors, and a sensor mat at the rear door. The experiment results were as follows: 80% from Wi-Fi sensing, 84% from video sensing, and 90% from sensor mat. The authors reported that the team successfully collected infrared data and gained insight into the appropriate positioning of the infrared sensor. Moreover, the authors noted that sensor mats could not be used for real-time estimation because data collected cannot be accessed instantly; however, infrared sensors and video sensing will be further assessed in the future for real-time estimation. Table 2-8 lists some related studies involving multiple technologies.

Table 2-8 Relevant Studies Involving Multiple Technologies.

Study	Research Topic	Mode	Data Classification	Region
Kouyoumdjieva et al. (2019)	A detailed investigation of non-image-based people counting methods.	Transit/ Non-Transit		Stockholm, Sweden
Sørensen et al. (2019)	A review of how the number of passengers on trains is measured, including technologies and practices of measuring actual ridership.	Rail	Non-Real Time	Norway
Moser et al. (2019)	A methodology for empirically evaluating APC systems for deployment in public transport bus services.	Bus	Non-Real Time	Australia
Pu et al. (2019)	Application of Wi-Fi and Bluetooth sensing devices to monitor real-time public transit ridership flow.	Bus	Real Time	Seattle, US
El-Tawab et al. (2020)	Smart city data sensing using an IoT framework.	Bus	Real Time	Virginia, US
Jenelius (2020)	A system used to deliver customized, anticipated in-vehicle crowding information to commuters on public transportation via mobile applications or at-stop screens.	Bus	Real Time	Stockholm, Sweden.
Noursalehi et al. (2021)	A real-time predictive decision support platform that addresses both operations control and customer information requirements.	Rail	Real Time	Massachusetts, US.
Jiang et al. (2021)	Advanced sensing and networking technologies that collect and analyze multimodal, multi-perspective, and real-time crowding data pertinent to crowd management.	Non-Transit	Real Time	Saudi Arabia
Drabicki et al. (2022)	Investigation of whether providing real-time crowding information (RTCI) at the stop on the two upcoming vehicle	Bus	Non-Real Time	Warsaw, Poland

Study	Research Topic	Mode	Data Classification	Region
	departures can encourage customers to wait for a less packed departure, hence reducing the bunching impact.			
Zhao et al. (2022)	Using AFC and Wi-Fi data to investigate an effective method for detecting passengers in a metro system.	Rail	Non-Real Time	Shenzhen, China

2.1.10 Miscellaneous

2.1.10.1 Manual Surveys

Manual passenger counting was the most used passenger counting method in transit systems as of 1998 (Boyle 2008). Later, more transit agencies combined automated and manual methods in collecting ridership data. Based on a survey completed by New York City Transit, Masters et al. (2003) also indicated that manual surveys were largely used for passenger counting despite all the technological advancements in automated counting. Since manual counting is a versatile method, it can be used to count passengers in any mode of transportation, whether passengers are boarding or alighting from vehicles or changing routes. Masters et al. (2003) demonstrated that manual counts were mainly used to count people entering and exiting vehicles. Due to the high accuracy, manual surveys remain as the only way to collect the ground truth data in order to validate the estimation results using other counting technologies (Asim et al. 2022). However, it should be noted that manual surveys are time consuming, susceptible to human error and expensive. Manual surveys are not ideal for crowded areas, and data could be affected in scenarios such as unpredicted change in weather conditions. Hence most agencies end up combining manual surveys with other methods such as AFC and APC.

2.1.10.2 Lidar (Light Detection and Ranging)

LiDAR is a technology for locating far-off objects and figuring out their position, speed, or other details by examining the pulsed laser light reflected from their surfaces. LiDAR can maintain the number of people passing through an entry, thus enabling efficient access control and crowd management (Ganz Security, 2021). Lesani et al. (2020) employed LiDAR to count and identify the travel direction of each individual in environments with high pedestrian flows. The pedestrian counting was based high-frequency (everyone 20 ms) distance measurements with a 16-channel LiDAR sensor. After comparing with the manual counting results (ground truth), Lesani et al. (2020) found overall 97% of pedestrians were accurately detected and counted. The authors proposed comparing 2D LiDAR and computer vision technology in terms of performance computing and cost as their future work. Kouyoumdjieva et al. (2020) reported that although LiDAR technology could achieve higher accuracy, the associated costs for equipment are high and may prevent large-scale deployments. It should be noted that most buses do not LiDAR sensors at present.

2.1.10.3 Acoustic and Ultrasound Sensors

According to Darsena et al. (2022), acoustic sensor-based methods count people using audio signals produced by speaking individuals or provided by cellphones. Kannan et al. (2012) developed a crowd counting technique based on audio tones using the microphones and

speakerphones that are often present on mobile phones. The solution was applied on 25 Android phones that were used for several studies at bus stops, on buses, in cafeterias, and classrooms. Another study on the acoustic-based solution using ultrasonic sounds was conducted by Kouyoumdjieva et al. (2020). The authors stated that the number of occupants could be deduced from the features of the received reverberation of transmitted waves, such as the receive time or the signal decay. The fundamental premise behind this method is that as the number of people in the room increases, the signal decays more quickly. As a result, the reverberation time can be utilized as a feature to evaluate occupancy. Another approach that can be used is to calculate the population by tracking the energy of the acoustic signal over time.

2.1.10.4 Carbon Dioxide Concentration

Environmental parameters are becoming popular for occupancy estimation. For instance, every person in a room emits CO_2, the concentration of CO_2 in the air can be used to estimate the number of individuals. Most of the research involving environmental parameters have been done within indoor/building settings. For example, Candanedo & Feldheim (2016) estimated workplace occupancy using data from light, temperature, humidity, and carbon dioxide sensors. In addition, Jiang et al. (2016) developed an indoor occupancy estimator which can estimate the number of real-time indoor occupants using a CO_2 sensor which is part of a standard HVAC (Heat, Ventilation and Air-conditioning) system. Kouyoumdjieva et al. (2020) noted that numerous other parameters, such as venting mechanisms that continually lower the CO_2 content, may affect the estimating accuracy. Thus, these aspects need to be taken into consideration.

2.2 Technical and Non-Technical Evaluation Summary

In this section, we will provide the summary of the evaluation of potential technologies; we first considered the technological factors such as accuracy, latency, dependability, automation degree, ease of installation, and maintenance. We then assessed each technology using non-technical criteria such as cost, privacy, public acceptance, and transit agency acceptance. Table 2-9 provides a high-level summary of the findings.

Table 2-9 Evaluation of Potential Technologies

Technology	Accuracy	Reliability	Latency	Level of Automation	Level of Implementation	Maintenance	Cost	Privacy	Public Acceptance	Transit Agency Acceptance
Automatic Passenger Counter	High	High	Medium	Low	Low	Medium	Medium	Low	Yes	Yes
Automatic Fare Collection	High	High	Medium	Low	Low	Medium	Medium	Medium	Yes	Yes
Crowdsourcing	Medium	High	Low	Medium	Low	Medium	Low	Low	Yes	Yes
Wi-Fi	Medium	Medium	Low	Medium	Low	Low	Low	High	Yes	Yes
Bluetooth	Medium	Medium	Low	Low	Low	Low	Low	High	Yes	Yes
Cellular Network	Medium	Low	Medium	Medium	Medium	Medium	Medium	Medium	Yes	Yes
Optical and thermal Cameras	High	Medium	Medium	Medium	Low	Low	High	High	-	-
Mobile Ticketing	Medium	Medium	Medium	Low	Medium	Low	Medium	High	Yes	Yes
Hybrid	High	High	Low	Medium	Medium	Medium	Medium	Medium	-	-
Manual Surveys	High	High	Medium	Low	Medium	Low	Medium	Low	Yes	Yes
LiDAR	High	Medium	Medium	Medium	Medium	Medium	Medium	Medium	-	-
Acoustic and Ultrasonic Sensors	Low	Low	Medium	Medium	Medium	Low	Low	Medium	-	-
Carbon dioxide Concentration	Low	Low	Medium	Medium	Low	Low	Medium	Low	Yes	-

Not all the technologies evaluated above are available in Florida's transit systems. Table 2-10 show the technology availability survey results. We can find those available technologies include APC, AFC, Cellular Data, and Wi-Fi. In addition, manual surveys are used for obtaining ground truth data. Note that those technologies can be combined to achieve the best possible estimation accuracy and reliability.

Table 2-10 Technology Availability Survey Results

Technology	StarMetro	LYNX	Miami-Dade County Transit
Automatic Passenger Counter (APC)	Yes	Yes	Yes
Automatic Fare Collection (AFC)	Yes	Yes	Yes
Crowdsourcing	No	No	No
Wi-Fi	Yes	Yes	Yes
Bluetooth	No	No	No
Cellular Data	-	Yes*	-
Optical and Thermal Cameras	No	No	No
Mobile Ticketing	No	Yes	Yes
Hybrid	No	No	No
Manual Surveys	Yes	Yes	Yes
LiDAR	No	No	No
Acoustic/Ultrasound Sensors	No	No	No
Carbon dioxide Concentration	No	No	No

* Data could be available through a different agency upon request

2.3 Wi-Fi Applications Review

Through the evaluation of the prospective technologies, Wi-Fi probing technology is found to have the most potential for real-time vehicle occupancy estimation. Public transit operators and other passenger transportation companies provide free Wi-Fi to improve the travel experience of their customers or get more people to use their services. One study of people who ride the Capitol Corridor trains in California found that free Wi-Fi led to a 2.7% rise in train ridership Dong et al. (2015). Wi-Fi probing is a promising option because it requires little hardware, and data can be made available in real time.

Smartphones with Wi-Fi and Bluetooth are widely used, accounting for more than 80% of the market in the U.S. (Asim et al., 2022); approximately 307 million people are smartphone users, and the numbers are projected to continue to increase over time. The authors reported that most smartphone users leave their Wi-Fi mode on. It is also observed that Wi-Fi has a greater connection speed, range, and level of internet access than Bluetooth. Its detection rate is also substantially higher than that of Bluetooth. The discovery time for Bluetooth is roughly 10.21

seconds, while the discovery time for Wi-Fi is approximately 1 second (Videa & Wang, 2021). Wi-Fi is also more likely to be enabled on a mobile device than Bluetooth. Thus, the use of Wi-Fi technology to estimate vehicle occupancy has been explored in the literature, with some studies involving combining Wi-Fi with other technologies, such as Bluetooth.

The number of smartphones on board is, therefore, a potentially useful proxy for the number of passengers aboard the bus. In estimating the on-board crowding, one can monitor Wi-Fi signals, classify them as originating from non-passenger or passenger devices, and use the result as a proxy to estimate the number of passengers on board the bus. Wi-Fi-capable devices are assigned a unique identifier or ID. A smartphone can be identified by its international mobile equipment identification (IMEI) number or its MAC address (Hidayat et al., 2020). A MAC address is a unique code that does not contain any user-specific information. Using the probe request, or other Wi-Fi frames MAC addresses can be detected. Wi-Fi-enabled consumer devices routinely perform a wireless probe process by transmitting a "probe request" frame (Oliveira et al., 2019). The purpose of this method is for surrounding Wi-Fi Access Points (APs) to transmit information regarding accessible wireless networks. The scanning process occurs regardless of whether the user is connected to a wireless network or not, as long as the Wi-Fi interface is enabled. Even after connecting to a network, the device continues to scan for networks with stronger signal strength.

The interval between sending two probe requests varies, depending on the type of operating system, such as Android, iOS, and Windows, and if the screen mode is on or off. Mobile devices send probe requests up to 55 times per hour on average (Freudiger, 2015). Overall, Android has the smallest interval, followed by Windows and iOS. When an iOS or Windows phone connects to a Wi-Fi network, it has a large interval of around 1,200 seconds. However, the Android phone maintains a short interval of 2.11 and 2.15 seconds (Li et al., 2016). The different probe request interval is caused by a differentiated energy-saving design of smartphones in Wi-Fi registered mode.

Thus, a Wi-Fi sniffer can be used to monitor the transmission of different frames, such as probe requests. This method is easy to implement, non-participatory like passengers do not need to install an app or visit a particular website, and it is non-invasive; it does not affect passengers' normal use of their devices. Several studies have explored the application of Wi-Fi technology in public transportation systems. Asim et al. (2022) detailed research done in Calgary, Canada using Wi-Fi sensors to learn about the habits of light rail riders. They noted that Wi-Fi sensing showed promise in comprehending passenger journeys, but that validation of Wi-Fi sensing calculations required manually obtained data or surveillance footage.

Relatedly, Gu et al. (2021) tried to reconstruct the spatio-temporal trajectory of a rail transit passenger from partially captured Wi-Fi probe data in Shanghai, China. The potential of using Wi-Fi and/or Bluetooth to extract Origin-Destination (O-D) movements is presented by (Dunlap et al., 2016; Mishalani et al., 2016; Pu et al., 2019.). Mishalani et al. (2016) presented that using Wi-Fi data to identify transit passenger O-D flows holds potential when aggregated across numerous bus runs for a period of the day. This possibility is increased when Wi-Fi observations are linked with other data sources, particularly boarding and alighting counts obtained using APC technologies.

However, a few setbacks have been presented in the studies, which include capturing devices of non-passengers, thus creating an overestimation of occupancy. Also, not all passengers carry mobile devices or have Wi-Fi disabled on their mobile devices, thus leading to underestimation. Those challenges can be overcome through filtering and adjusting data. Vieira et al. (2020) provided one related case study of Wi-Fi probing in Belo Horizonte, Brazil. Kalman filters were used by Vieira et al. (2020) to adjust the passenger count.

Paradeda et al. (2023) summarized that applications of Wi-Fi technology in public transportation have focused on the estimation of various variables such as bus stop, station, or terminal data (number of waiting passengers and/or their waiting times and/or in and outflows); OD matrixes; bus load; boarding and/or alighting; frequency of bus route use; travel times and transfers, as shown in Table 2-11.

Table 2-11 Summary Applications of Wi-Fi in Public Transportation (Paradeda et al., 2023).

Study	Bus stop, station, or terminal data	OD matrixes	Bus load	Boarding and/or alighting
Algomaiah and Li (2022)	—	Yes	Yes	Yes
Asim et al. (2022)	Yes	—	Yes	Yes
EL-Tawab et al. (2019)	Yes	Yes	—	—
Fukuda et al. (2017)	—	Yes	Yes	Yes
Hakegard et al. (2018)	—	Yes	Yes	—
Hidayat et al. (2020b)	—	—	Yes	—
Ji et al. (2017)	—	Yes	Yes	—
Junior and Medrano (2018)	—	—	Yes	Yes
Junior et al. (2022)	—	—	Yes	Yes
Myrvoll et al. (2017)	—	—	Yes	—
Nitti et al. (2020)	—	—	Yes	—
Oransirikul et al. (2016)	Yes	—	—	—
Oransirikul et al. (2019)	—	—	Yes	—
Pu et al. (2021)	—	Yes	—	—
Ryu et al. (2020)	Yes	Yes	Yes	—
Vieira et al. (2020)	—	—	Yes	—

2.4 Research Gaps

Many researchers have used Wi-Fi to estimate the number of people utilizing public transportation. For example, Mishalani et al. (2016), Mikkelsen et al. (2016), Myrvoll et al. (2017), and Mehmood et al. (2019) are just a few examples. The previously mentioned review of the literature reveals that practically most of the previous research mainly considered probe requests only. This may be due in large part to the fact that in the original research, the source MAC address of a probe request was not randomly generated, suggesting that a specific mobile device may be uniquely identified. However, most smartphone manufacturers have implemented

MAC address randomization to safeguard user privacy. Vanhoef and Piessens (2016) argue that randomizing mobile devices' MAC addresses is a crucial step toward protecting users' privacy in an increasingly connected world. However, this does present a challenge in uniquely identifying devices for estimation purposes.

MAC address randomization varies between devices and operating systems. Devices' MAC addresses can be changed after each network connection, on a regular basis in response to other events such as the mode of the screen of the device. The mobile device industry is extremely diverse, with a wide range of manufacturers and operating systems in use. Because each of these systems executes MAC address randomization differently, estimating device numbers becomes even more difficult. In addition, the frequency with which devices connect to Wi-Fi networks or other wireless networks varies, which also influences the accuracy of device counting. This makes establishing a steady baseline for device counting challenging.

Therefore, because of the unpredictability and diversity of randomization algorithms, this privacy enhancement complicates calculating the number of devices in a network. Despite the randomization challenge, Vanhoef et al. (2016) suggested that it is possible to fingerprint the information elements of the probe requests and therefore identify a device and that authenticity of a MAC address can be determined by examining the OUI (Organization Unique Identifier) component, which can be cross-referenced with the IEEE table of registered vendors. If the OUI is not present in the table, the MAC address is classified as randomized; otherwise, it is categorized as non-random as suggested by Koç (2022) in the Master's thesis. However, Rusca et al. (2023) reported that the randomization process is predominantly conducted on a per-burst basis, with a minor increase in burst duration observed while transitioning from iOS to Android. The authors added that the consistency of MAC randomization implementation varies among models from the same manufacturer, as evidenced by the differences observed between the iPhone 6 and the iPhone 11.

Oliveira et al. (2019) reported that iPhone devices running iOS 10.1.1 exhibit a novel MAC randomization process in various scenarios. These scenarios include the device being locked or unlocked, the activation or deactivation of a Wi-Fi interface, and the establishment or attempted establishment of a connection to a Wi-Fi access point. Therefore, it may be argued that the determination of the time duration during which a random MAC address is utilized in certain mobile devices is not feasible, as this timeframe is contingent upon the manner in which individuals engage with their smartphones. When all MAC addresses are randomized, it is impossible to uniquely identify MAC addresses for estimation purposes; We argue that other types of Wi-Fi frames or methods must be considered to explore the best potential of Wi-Fi data in estimating transit vehicle occupancy.

CHAPTER 3. WI-FI FRAME DATA COLLECTION

3.1 Fundamentals of Wi-Fi Association Process

The IEEE standard for Wireless Local Area Networks (WLAN) MAC (Media Access Control) and Physical Layer (PHY) is 802.11. The IEEE.802.11 standard documentation specifies the requirements for the exchange of information. The 802.11 protocol is used by Wi-Fi devices to transmit wireless network packets. The 2.4 GHz channel is utilized by mobile devices. 802.11b and 802.11g use the 2.4 GHz band while operational, while 802.11n and 802.11ac use either the 2.4 GHz or 5 GHz band (Kalikova & Krcal, 2017). Wi-Fi frames are gathered in accordance with IEEE 802.11 specifications. IEEE 802.11 defines management, control, and data frames as the three distinct types of frames.

Management frames are utilized for joining and exiting wireless networks. These are referred to as type 0 frames. The eight subtypes of management frames are beacon (0x8), probe request (0x4), probe response (0x5), association request (0x0), association response (0x1), reassociation request (0x2), reassociation response (0x3), authentication (0xb), de-authentication (0xc), disassociation (0xa), and action (0xd). To join the BSS (Basic Service Set), stations (such as mobile phones) submit association requests to access points. Control frames manage both frame acknowledgment and medium access. The control frames are identified as type 1. ACK-Acknowledgement, Block ACK request, RTS-Request to Send, CTS-Clear to Send, Block ACK, PS-Poll, CF-End, and CF-End/CF-Ack are subtypes of Type 1. Data frames are referred to as type 2 and are employed for transmitting data.

The Wi-Fi association process is a set of actions that a client device (for example, a smartphone or laptop) takes to connect to a Wi-Fi access point (AP) or router. This procedure is critical for establishing a safe and dependable wireless connection. The following are the important steps in the Wi-Fi association process:

- Scanning: When a client device is turned on or comes into contact with a Wi-Fi network, it begins searching for available networks. The device listens for beacon frames broadcasted by neighboring APs during this phase. These beacons include critical network information such as the network's name (SSID), security protocols supported, and signal strength.
- Network Selection: After scanning, the client device displays a list of available networks to the user. The user chooses a network to join based on its SSID and signal strength.
- Authentication: After selecting a network, the client device begins the authentication procedure. For open networks (those without security), this step is skipped, and the device connects directly to the selected AP. For secured networks, however, to establish the authenticity of it, the client must supply the right pre-shared key (PSK) or credentials. Only authorized users can connect to the network as a result of this process.
- After successfully authenticating, the client device sends an association request to the selected AP. The AP validates the credentials and accepts the association if they are correct. The client is now regarded to have connected to the network, and data transmission can commence.

Wi-Fi Association Security Considerations

In Wi-Fi networks, security is of the utmost importance. The association procedure includes various security features to ensure data confidentiality and integrity:

- Encryption: To protect data transfer, Wi-Fi networks often use encryption protocols such as WPA2 (Wi-Fi Protected Access 2) or its successor WPA3. These protocols encrypt data as it passes between the client and the AP, making it difficult for eavesdroppers to intercept and understand the information.
- Authentication: As previously stated, clients must give valid credentials in order to connect to the network. This prevents unauthorized individuals from connecting to the Wi-Fi network.
- Some networks employ the "hidden SSID" functionality, which means that the SSID is not broadcast in beacon frames. To connect, clients must manually enter the SSID, giving another degree of mystery to the network.

In summary, the Wi-Fi association process can be explained as follows: The access point routinely emits a beacon frame to indicate its presence and provide stations (e.g., mobile phones) with the information needed to connect to the wireless network. Any devices with Wi-Fi enabled actively send probe request frames on a regular basis (Oliveira et al., 2019). Probe requests highlight the station's transmission speeds and 802.11 features. When receiving a probe request, an access point verifies if the station can support at least one common data rate. If their data rates are the same, the access point's SSID, supported data rates, encryption types (if needed), and other 802.11 features are advertised in a probing response.

3.2 Hardware Setup for Wi-Fi Frame Capture

The hardware needed for Wi-Fi frame capturing is displayed in Figure 3-1. The initial setup includes the use of a mouse and keyboard; however, there is an option of a remote connection using a laptop instead, as displayed in Figure 3-2. In the initial step of the setup, the required three
files were downloaded to boot the micro-computer, the Raspian OS was installed in the micro-SD card and then placed in the micro-computer. Then, the Wi-Fi driver was installed, followed by installing remote connection software such as Putty or Nomachine software. The final step was enabling Wi-Fi sniffing software Wireshark. Once installation is complete packet sniffing can be conducted.

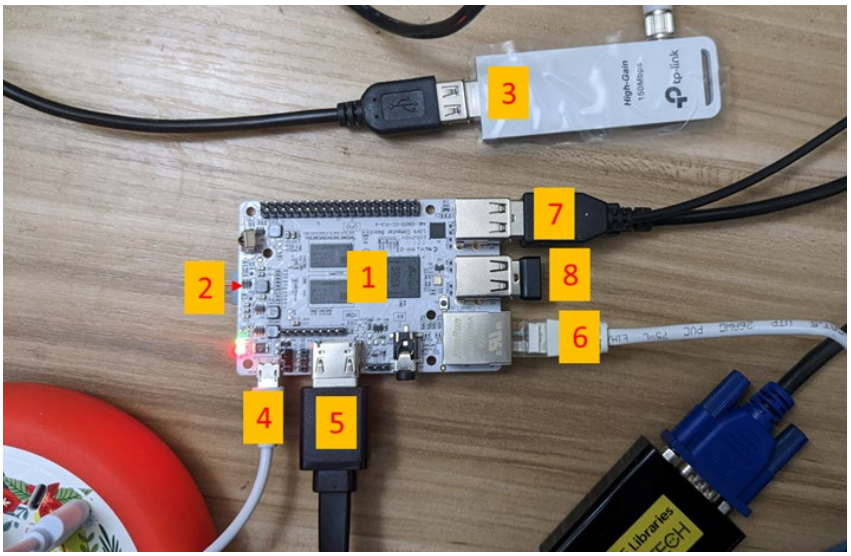


Figure 3-1 Initial hardware setup

1. Micro-computer board: Le Potato (AML-S950X-CC)
2. Micro SD card: 32 GB SanDisk Extreme (Inside Micro-computer)
3. Wi-Fi adapter: TP Link W722N v2/v3
4. Power supply: 5V-3A
5. Standard HDMI to VGA cable
6. Ethernet cable
7. Keyboard
8. Mouse

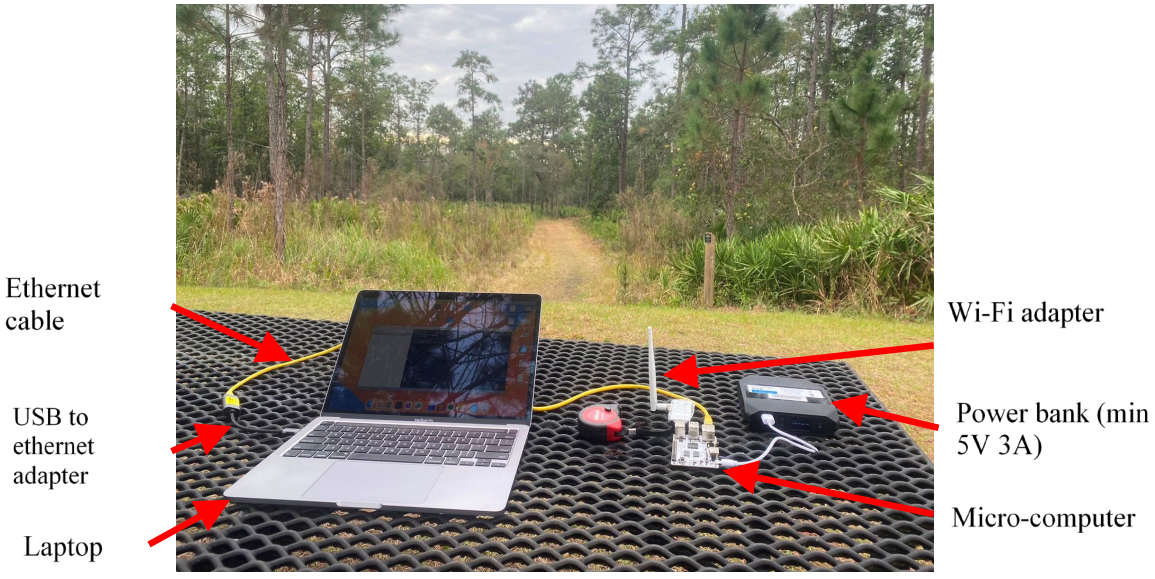


Figure 3-2 Hardware after remote connection.

3.3 Outdoor Experiments

The field experiments were conducted to understand the Wi-Fi probing mechanism and randomization of MAC addresses; explore the relationship between signal strength in probe requests and distance of the mobile devices from sniffers, and the time and frequency of probe requests from mobile devices. The experiments were done by the Florida State University (FSU), the University of Central Florida (UCF) and Florida International University (FIU) researchers in open space at J.R Alford Greenway in Tallahassee, UCF Arboretum and FIU football field. The test fields are shown in Figure 3-3, Figure 3-4, and Figure 3-5.

The experiments were conducted in two phases. In phase one, two types of experiments were conducted. In the first experiment, a series of tests were conducted using a Wi-Fi sniffer to detect the presence of a single iPhone at varying distances of 0, 10, 20, and 30 feet and randomization was on, phone settings left as default. Data on probe requests was collected for a duration of 10 minutes at various distances. It is important to note that during the data collection process, the iPhone's screen remained off but the device itself was switched on. Similarly, in the second experiment at varying distances the iPhone was repositioned at regular intervals of 3 minutes, with the screen mode on.

In phase two, the research team conducted additional experiments turning on and off MAC address randomization feature. Three types of experiments were conducted:

- First experiment: Probing for a Dell laptop at distances 0, 10, 20, and 30 ft from the Wi-Fi sniffer. Collecting data for 10 minutes at each distance Laptop screen set to on and then off.
- Second experiment: Probing for an iPhone 11 at distances 0, 10, 20, and 30 ft from the Wi-Fi sniffer. Collecting data for 10 minutes at each distance Laptop screen set to on and then off.
- Third experiment: Probing for Samsung Z flip (Android), at distances 0, 10, 20, and 30 ft from the Wi-Fi sniffer. Collecting data for 10 minutes at each distance Laptop screen set to on and then off.

These experiments were conducted to better understand Wi-Fi probing mechanisms and MAC randomization to help oh how to better analyze the data in the next steps.

Figure 3-3 UCF test site

Figure 3-4 FSU test site

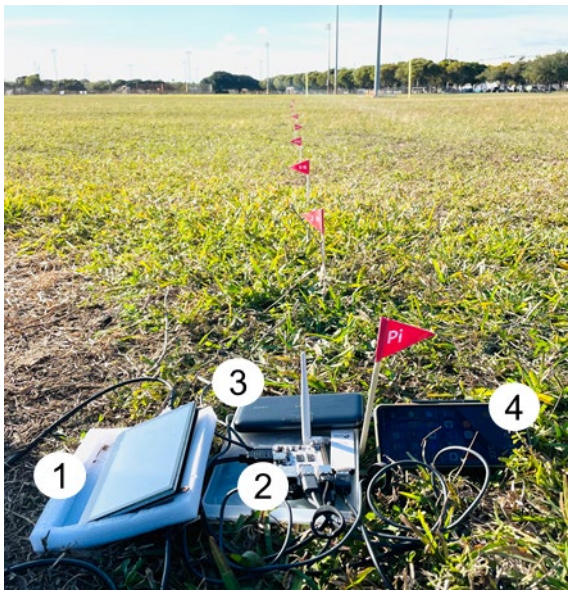


Figure 3-5 FIU test site

3.4 Pilot Studies

In this section we will provide a brief description of the conducted pilot studies in Tallahassee, Orlando and Miami Florida.

3.4.1 StarMetro

3.4.1.1 Automated Data Collection

The bus route selected for pilot studies in Tallahassee, FL is Evergreen, which connects Tallahassee Community College (TCC) to Apalachee Parkway Walmart Shopping Center. The roundtrip travel time is approximately two hours. Two types of automated data, namely Wi-Fi frames and vehicle location data, were collected by the research team. Figure 3-6 shows the hardware used for collecting the Wi-Fi frames:

- Raspberry Pi 3 Model B v1.2 with operating system installed on a micro-SD card.
- Wi-Fi USB adapter: TP-Link (TL-WN722N)
- Portable Power Banks (10,000mAH / 5V)
- Portable monitor, keyboard, and mouse

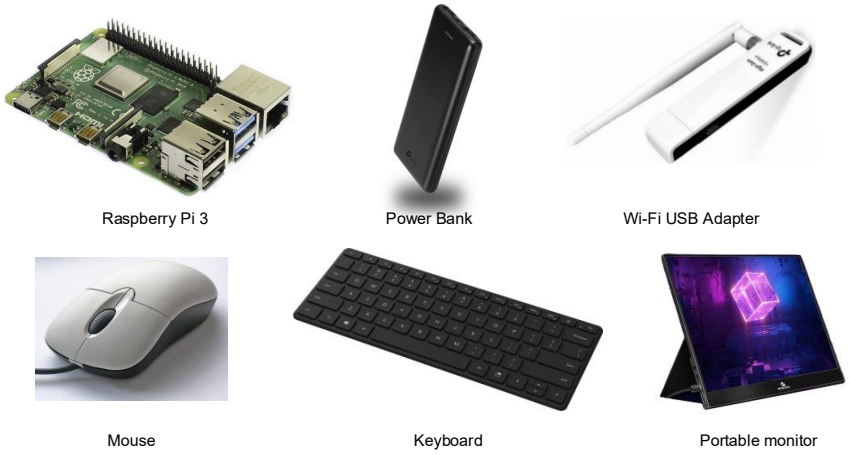


Figure 3-6 Hardware used for Wi-Fi frame collection.

An open-source packet analyzer, called Wireshark, was installed on the Raspberry Pi to capture the Wi-Fi frames and later analyze such collected data. In addition, a smartphone with the GPS Tracks app installed was used to collect bus location data over time. An example bus trajectory from March 29^{th,} 2023, is displayed in Figure 3-7.

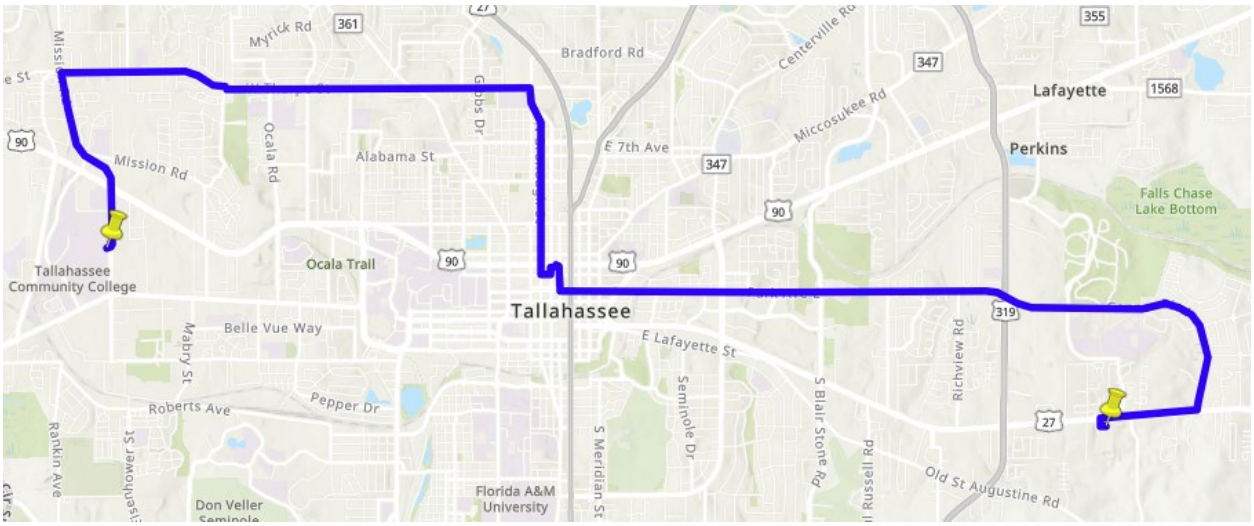


Figure 3-7 GPS trajectory of an Evergreen bus (TCC to Apalachee Parkway Walmart)

3.4.1.2 Manual Data Collection

While collecting the automated data, a checker manually counted the number of people getting on and off at every stop. The manually collected bus occupancy data were used as ground truth to validate the occupancy estimations from Wi-Fi frames. Some sample occupancy data are shown in Table 3-1. The actual bus departure time from each stop was also recorded.

Table 3-1 A sample of a Customized Manual Survey Form for March 29th, 2023

ON-BOARD PASSENGER CHECK	CHECKER	EM		
DATE 03/29/2023				
		PASSENGERS		
STOPS	TIME	ON	OFF	LOAD
Tallahassee Community College	11:05 AM			30
Appleyard Drive and Tennessee Street	11:06 AM	1	2	29
N Mission Road and Appleyard Drive				29
W Mission Road and Greenon Lane				29
N Mission Road and Rexwood Drive	11:08 AM	1		30
W Tharpe Street and N Mission Road				30
N Mission Road and Tharpe Street				30
W Tharpe Street and Burns Street				30
W Tharpe Street and Trimble Road				30
W Tharpe Street and Falcon Crest				30
W Tharpe Street and Devra Drive	11:10 AM	1		31

Two graduate students jointly completed the collection of Wi-Fi frame data and manual survey for a period of two consecutive weeks, starting from March 27th to April 7th, 2023. On each day, the students were expected to collect the data for two round trips. The data collection schedule was designed considering the availability of the involved graduate students.

3.4.1.3 Descriptions of Collected Data

The Wi-Fi frames were collected under the IEEE 802.11 specifications using Wireshark. Management, control, and data frames are the three different types of frames defined in IEEE 802.11. A general list of subtypes of management frames is provided in Table 3-2. In addition, Figure 3-8 shows an example probe request from a Wireshark capture file.

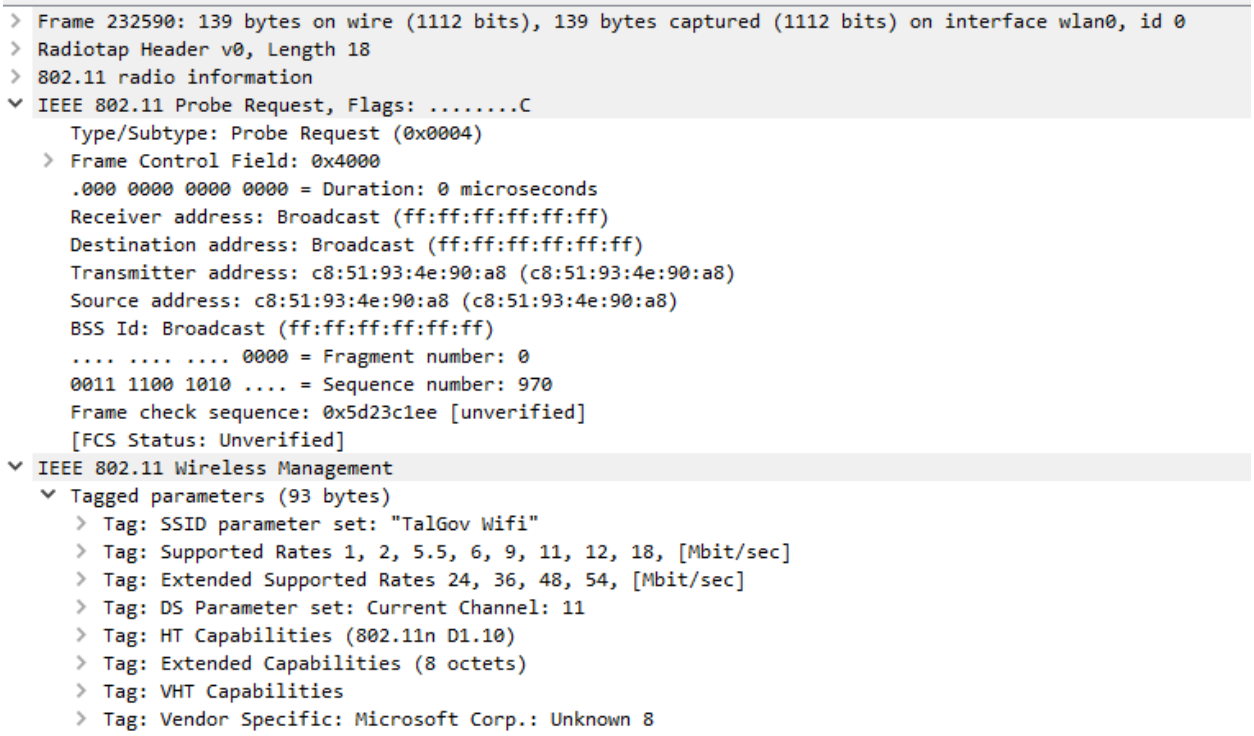


Figure 3-8 Example of management frame (type 0) with subtype known as probe request.

Table 3-2 List of Management Frame Subtypes

Subtype Field	Description
0000	Association request
0010	Reassociation request
0100	Probe request
1000	Beacon
1010	Disassociation
1100	De-authentication
1011	Authentication
1110	Action
0001	Association response
0011	Reassociation response
0101	Probe response

Note: Adapted from (howiwifi.com)

Control frames are used for frame acknowledgment and to manage access to the medium. Control frames are described as type 1. Type 1 subtypes are listed in Table 3-3.
One of the control frame subtype examples is Request-To-Send (RTS). Stations transmit RTS frames to reserve the medium for the time specified in microseconds in the duration field, as shown in Figure 3-9.

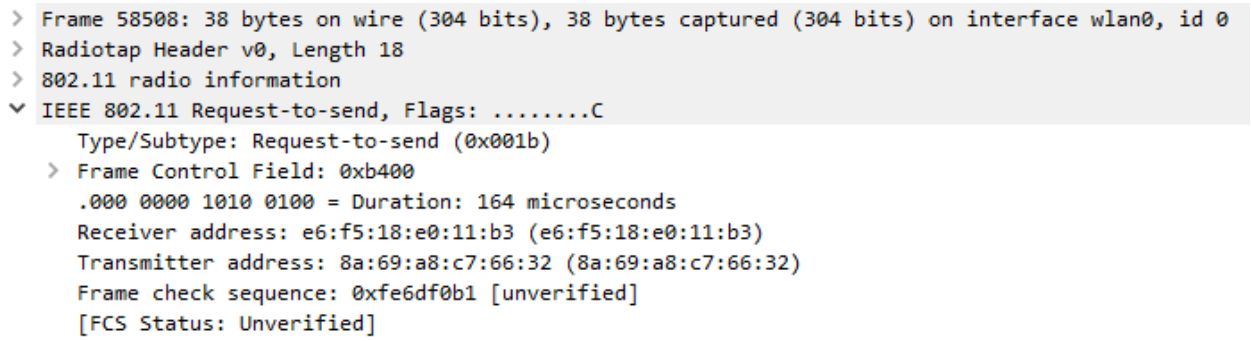


Figure 3-9 Example of control frame with subtype Request-to-Send (RTS).

Table 3-3 List of Control Frame Subtypes

Subtype Field	Description
0100	Beamforming Report Poll
0101	VHT/HE NDP Announcement
0110	Control Frame Extension
0111	Control wrapper
1000	Block ACK request
1001	Block ACK
1010	PS-Poll
1011	RTS
1100	CTS
1101	ACK
1110	CF-End
1111	CF-End+CF-Ack

Note: Adapted from (howiwifi.com)

Data frames are used to transmit information and are described as type 2. The subtypes for data frames are shown in Table 3-4. Figure 3-10 depicts an example of a QoS data frame. QoS data is utilized whenever a QoS station is transmitted to another QoS station.

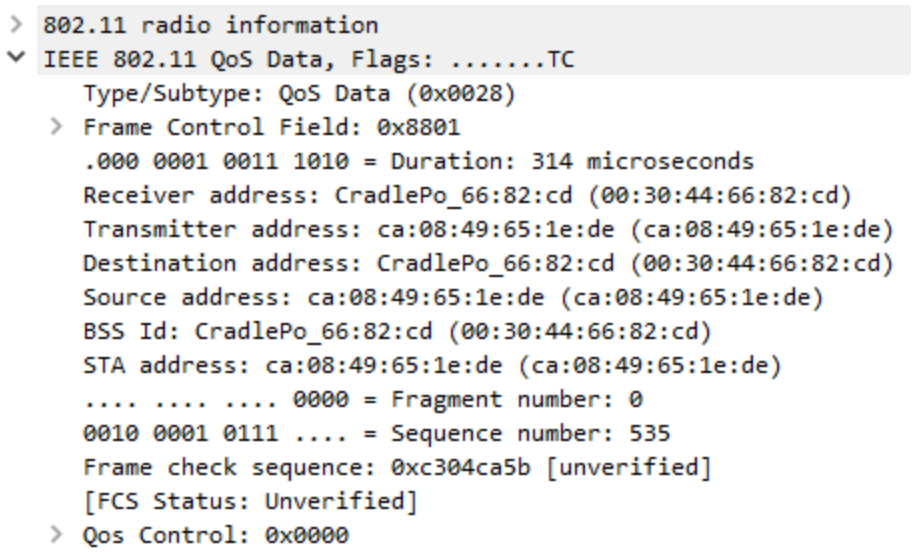


Figure 3-10 Example of data frame (type 2) with subtype QoS Data.

Table 3-4 List of Data Frames Subtypes

Subtype Field	Description
0000	Data
0001	Data+CF-Ack
0010	Data+CF-Poll
0011	Data+CF-Ack+CF-Poll
0100	Null (no data)
0101	CF-Ack (no data)
0110	CF-Poll (no data)
0111	CF-Ack+CF-Poll (no data)
1000	QoS Data
1001	QoS Data+CF-Ack
1010	QoS Data+CF-Poll
1011	QoS Data+CF-Ack+CF-Poll
1100	QoS Null (no data)
1101	Reserved
1110	QoS CF-Poll (no data)
1111	QoS CF-Ack+CF-Poll (no data)

Note: Adapted from (howiwifi.com)

The various Wi-Fi frames collected were stored in PCAP files which stands for packet capture. One packet capture file was associated with each trip. The average size of the packet capture files was 40 MB. The file sizes ranged from 5.6 MB to 110 MB. Variation is dependent on variables such as the number of commuters and the packet capture duration. Due to the capture of more packets, trips with a greater number of passengers typically have larger file sizes. In a similar manner, longer travel times result in longer capture durations, resulting in larger file sizes. PyShark was utilized to read the files and export the files to Excel format. The variables that were selected for the subsequent analysis include average captured length, average data rate,

average signal strength, average duration, the unique number of MAC addresses of (source, transmitter, transmitter, receiver, destination) of specific subtypes, number of packets and unique number of MAC addresses by selected type and subtype.

The extracted information came from different layers. First, the radio tap layer contained the channel frequency, length, and data rates. Second, the WLAN radio layer contained the signal strength and duration information. Third, the WLAN layer included: type, subtypes, source address, transmitter address, receiver address, destination address and BSSID. Lastly, the WLAN management layer contained the SSID, wps_vendor_id, WLAN-supported rates, and WLAN HT capabilities. This layer could be used to identify the manufacturers of the devices/ stations, hence providing useful information when analyzing the data for occupancy estimations.

The heat map was to visualize and explore the collected data. Figure 3-11 displays the total number of packets for different types and subtypes. Some types and subtypes had more packets compared to others. For example, it is observed that type 1 with request to send, clear to send, and type 0 for probe request and probe responses, as highlighted in Figure 3-11, cover the most information and were explored further to discover any potential relations with the vehicle occupancy.

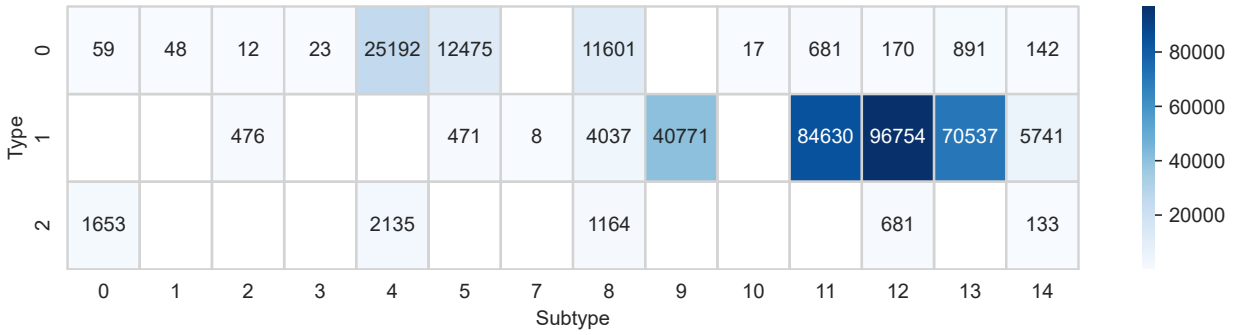


Figure 3-11. Total number of packets for different types and subtypes.

Furthermore, the number of unique source addresses for different types and subtypes was also visualized, as shown in Figure 3-12. Probe requests were observed to contain more data in type 0. Similarly, Figure 3-13 illustrates the unique number of MAC addresses of the receiver address where Acknowledgment, clear-to-send, and probe responses are observed to contain most of the data in type 1.

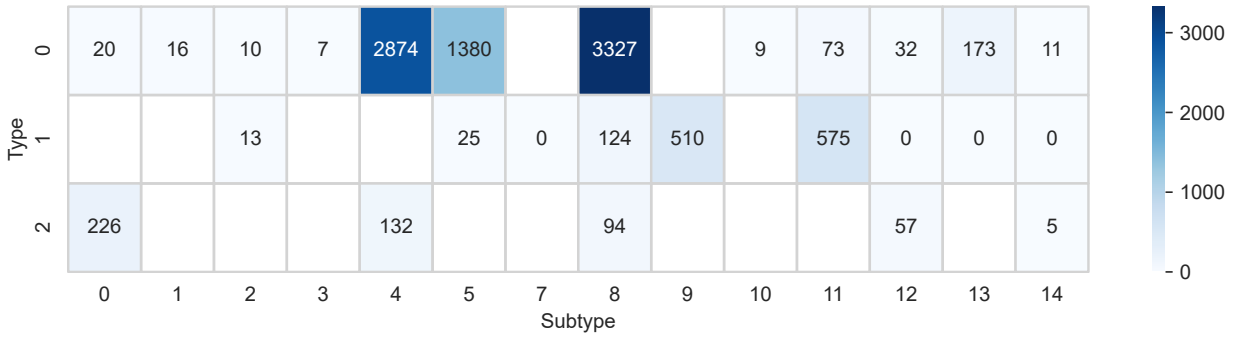


Figure 3-12 Unique number of source MAC addresses for types and subtypes.

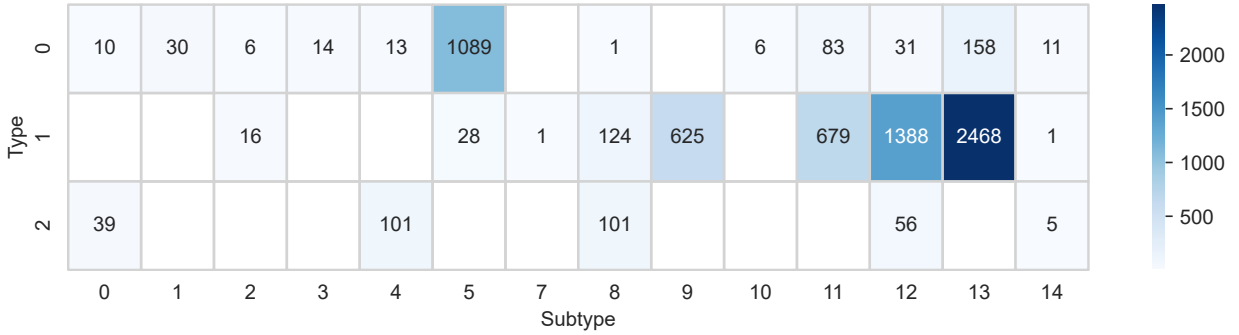


Figure 3-13 Unique number of receiver MAC addresses for types and subtypes.

The number of unique MAC addresses by frame type, subtype, and address type is used as the features. For instance, in a probe request, the source address or transmitter address represents the MAC address of a mobile device while the destination address or receiver address is the address of an access point. Nonetheless, in a probe response, the source address becomes the address of an access point. After extracting the key attributes from the Wireshark capture files and merging data from all days, we can compute the values of features (independent variables) over time assuming different time intervals, e.g., 1 minute, 2 minutes, and 5 minutes.

3.4.2 Lynx

3.4.2.1 Automated Data Collection

In the pilot study in Orlando, the automated data collection refers to the collection of Wi-Fi frame or packet data using hardware shown in Figure 3-14. Wi-Fi frame data include the wireless signal information coming from mobile devices such as cellphones, laptops, and tablets, from the vicinity of the Wi-Fi sniffer. In the pilot study, the sniffer was located on the bus.

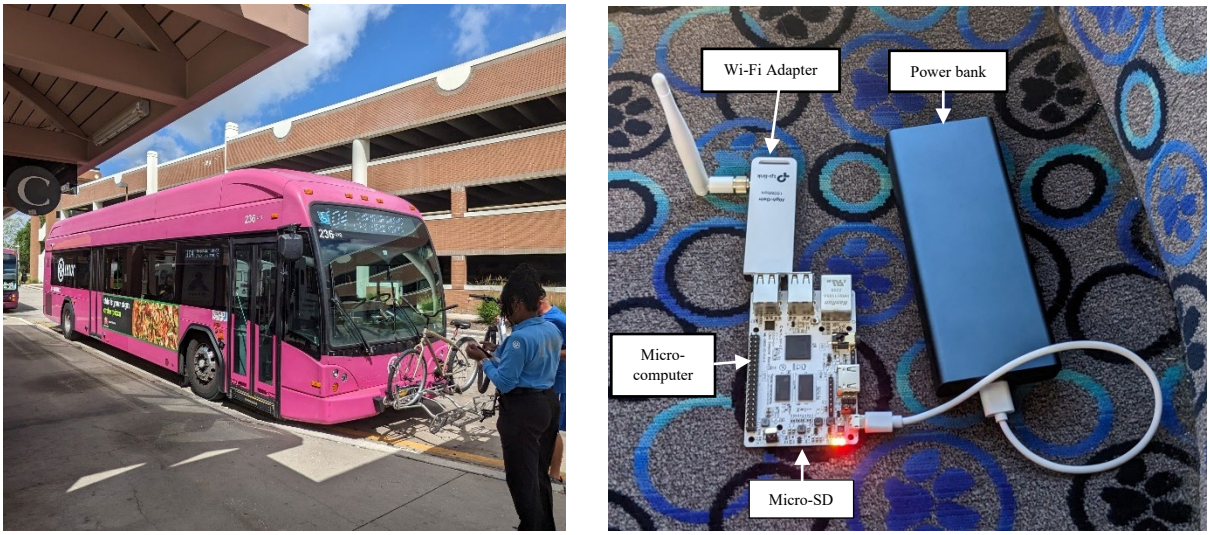


Figure 3-14 Hardware (right) for Wi-Fi probe data collection in Lynx bus (left)

Two types of frame collection software Wireshark (GUI application) and tshark (without GUI application) were installed in the Raspberry OS of Le Potato board. TP Link W722N Wi-Fi adapter was plugged into the board. The monitor mode feature of this adapter aims to capture Wi-Fi frame data from all Wi-Fi traffic received on a wireless channel. Once the bash scripting code for collecting the Wi-Fi probe was executed from the terminal of the Le Potato board, the probe data collection began. As the Le Potato board did not have its own monitor and keyboard, we connected a laptop using ethernet and sent codes in the terminal of the Le Potato board. The Wi-Fi frame data were automatically stored on a micro-SD card. The power was supplied from a power bank to smoothly run the data collection.

The hardware was located at the seat just behind the back door (almost at the center of the bus) so that Wi-Fi probe data from all the passengers can be collected without any bias of distance.

3.4.2.2 Manual Data Collection

The manual data collection refers to the passenger count data at different bus stop locations along the subject route. In the pilot study, route 104 was the selected route for the pilot study as shown in Figure 3-15. A logbook was used to write down the number of passengers who got on/off the bus at bus stops, and the stop locations were recorded using the “GPS Coordinates” Android app. The location information was later merged with logbook data based on stoppage time.

The data were collected from the seat (slightly elevated) just behind the back door which helped to observe the passengers getting off/on.

3.4.2.3 Data Collection Schedule

Both manual and automated data were collected simultaneously in the Lynx bus. A total of ten business days from March 27, 2023, to April 7, 2023, were selected for this data collection along Route 104 in the city of Orlando.

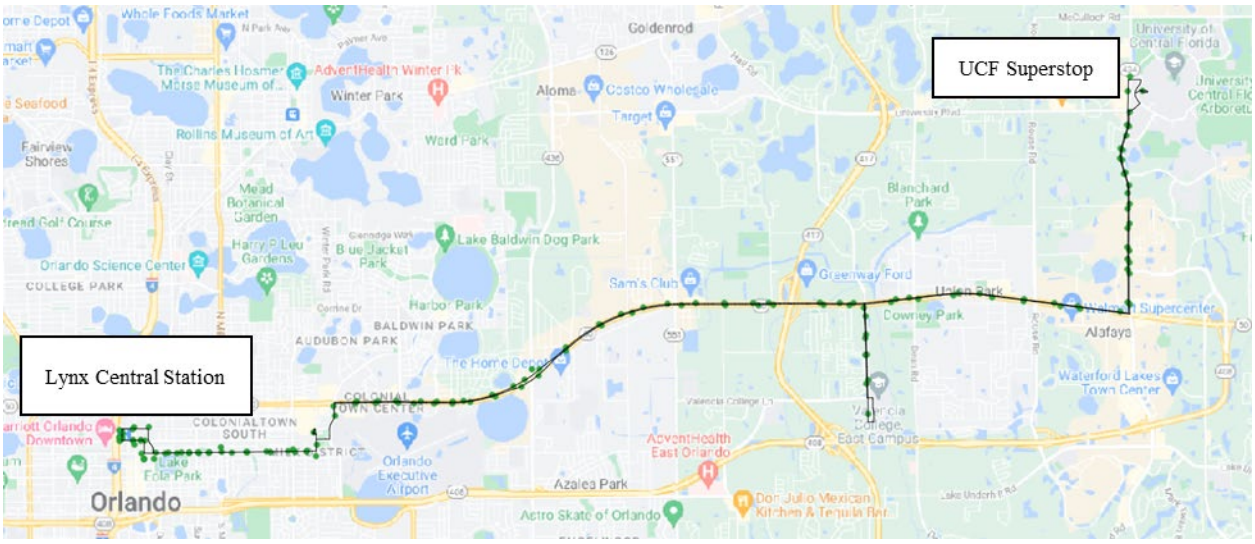


Figure 3-15 Lynx route 104 connecting the university and downtown Orlando

Every day two trips were chosen for data collection. The first trip (from UCF Campus Superstop to Lynx Central Station in downtown Orlando) was scheduled to depart at 8:55 am. The second trip was a return trip with around 10-15 minutes of break time that was scheduled to depart from Lynx Central Station at 10:15 am. The total duration of data collection from both trips was around 1 hr. and 15 minutes.

3.4.2.4 Descriptions of Collected Data

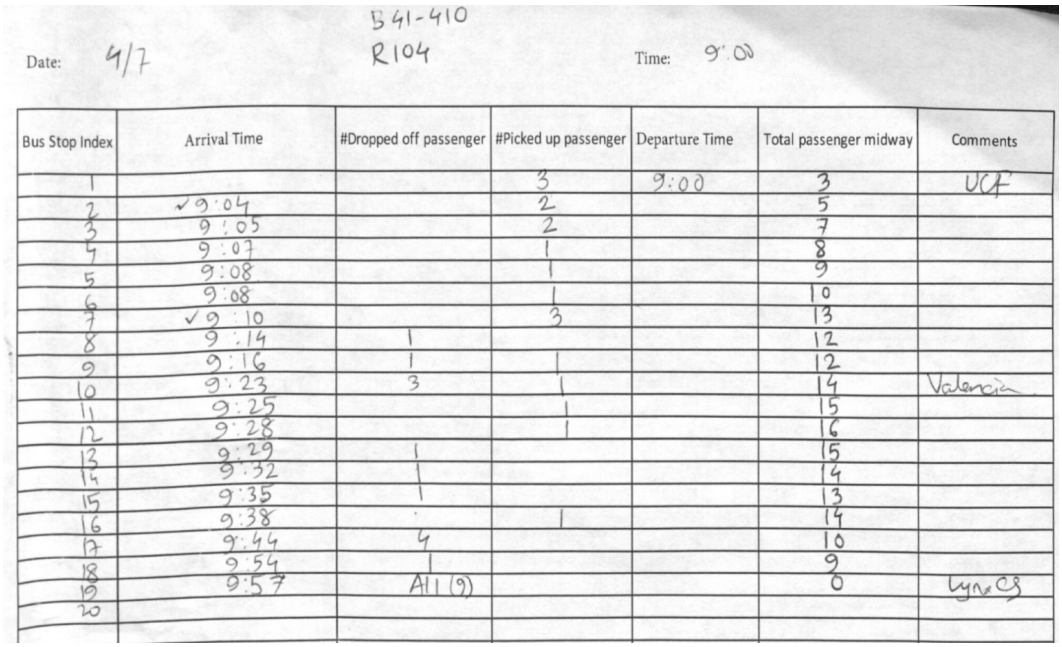
Figure 3-16 represents a sample of the automated collected data after converting to a .csv file from the original pcap (Wireshark) file. The conversion was done using a bash script code in terminal. This dataset has six columns such as frame.number, frame.time, wlan.sa, wlan.da, wlan_radio.signal_dbm, and _ws.col.Info.

The first column, “frame.number,” describes the order of the Wi-Fi probe signal (i.e., frame or packet) which was filtered based on the trip schedule. “Frame.time” describes the local date and time when the probe signal was captured. The MAC addresses of a sender and a receiver device for that particular probe signal were represented by “wlan.sa” and “wlan.da,” respectively. The signal strength in dBm units was collected in “wlan_radio.signal_dbm” column. The final column, “ws.col.Info” has information such as signal/frame type, subtype, SSID, etc.

frame.number	frame.time	wlan.sa	wlan.da	wlan_radio.signal_dbm	_ws.col.Info
146386	Apr 7, 2023 09:00:00.018160110 EDT	88:da:1a:06:f3:82	ff:ff:ff:ff:ff:ff	-55	Probe Request, SN=6, FN=0, Flags=.....C, SSID=LYNXGFI
146387	Apr 7, 2023 09:00:00.030660443 EDT	00:25:ca:3a:d0:63	ff:ff:ff:ff:ff:ff	-55	Beacon frame, SN=879, FN=0, Flags=.....C, BI=100, SSID
146388	Apr 7, 2023 09:00:00.039407693 EDT	dc:fe:23:ad:9d:c2	ff:ff:ff:ff:ff:ff	-95	Beacon frame, SN=2219, FN=0, Flags=.....C, BI=100, SSI
146389	Apr 7, 2023 09:00:00.062876485 EDT	00:30:44:6c:08:65	ff:ff:ff:ff:ff:ff	-49	Beacon frame, SN=3034, FN=0, Flags=.....C, BI=100, SSI
146390	Apr 7, 2023 09:00:00.065022527 EDT	06:30:44:6c:08:65	ff:ff:ff:ff:ff:ff	-49	Beacon frame, SN=591, FN=0, Flags=.....C, BI=100, SSID
146391	Apr 7, 2023 09:00:00.087578693 EDT	06:30:44:6c:08:65	14:eb:b6:45:44:69	-47	Probe Response, SN=3509, FN=0, Flags=.....C, BI=100,
146392	Apr 7, 2023 09:00:00.092317443 EDT	00:25:ca:3a:d0:63	14:eb:b6:45:44:69	-55	Probe Response, SN=880, FN=0, Flags=....R...C, BI=100, S
146393	Apr 7, 2023 09:00:00.096189943 EDT	00:25:ca:3a:d0:63	14:eb:b6:45:44:69	-55	Probe Response, SN=881, FN=0, Flags=.....C, BI=100, SS
146394	Apr 7, 2023 09:00:00.098936235 EDT	06:30:44:6c:08:65	14:eb:b6:45:44:69	-47	Probe Response, SN=3510, FN=0, Flags=.....C, BI=100,
146395	Apr 7, 2023 09:00:00.101181485 EDT	00:25:ca:3a:d0:4d	14:eb:b6:45:44:69	-91	Probe Response, SN=597, FN=0, Flags=....R...C, BI=100, S
146396	Apr 7, 2023 09:00:00.107812235 EDT	00:25:ca:3a:d2:d3	14:eb:b6:45:44:69	-75	Probe Response, SN=3568, FN=0, Flags=.....C, BI=100,
146397	Apr 7, 2023 09:00:00.121562610 EDT	88:da:1a:06:f3:82	ff:ff:ff:ff:ff:ff	-95	Probe Request, SN=7, FN=0, Flags=.....C, SSID=LYNXGFI
146398	Apr 7, 2023 09:00:00.125312360 EDT	00:25:ca:3a:d0:4d	ff:ff:ff:ff:ff:ff	-85	Beacon frame, SN=598, FN=0, Flags=.....C, BI=100, SSID
146399	Apr 7, 2023 09:00:00.133062027 EDT	00:25:ca:3a:d0:63	ff:ff:ff:ff:ff:ff	-55	Beacon frame, SN=882, FN=0, Flags=.....C, BI=100, SSID
146400	Apr 7, 2023 09:00:00.135184277 EDT	00:25:ca:3a:d2:d3	ff:ff:ff:ff:ff:ff	-75	Beacon frame, SN=3569, FN=0, Flags=.....C, BI=100, SSI
146401	Apr 7, 2023 09:00:00.135198318 EDT	26:76:db:67:69:ab	06:30:44:6c:08:65	-45	Null function (No data), SN=883, FN=0, Flags=.....TC

Figure 3-16 Automated Wi-Fi probe data sample for April 7, 2023

Manual data include two segments: the first segment is the passenger count data as shown in Figure 3-17 and the second segment is collected GPS data as shown in Figure 3-18. The passenger count data have a total of seven columns such as bus stop index, arrival time, #dropped off passengers, #picked up passengers, departure time, total passenger midway, and comments. The numerical order of locations where the bus stopped along the route in a trip was described by the “Bus Stop Index” column. “Arrival Time” and “Departure Time” columns represent the time when the bus arrives at a bus stop and departs from that stop, respectively. Note that this interval time duration between arrival and departure is also called the dwell time of that bus stop. The typical dwell time of bus stops except starting stop and ending stop of the trip was found to be less than 1 minute. The number of passengers getting on and off the bus at each bus stop is described by “#Picked up passenger” and “#Dropped off passenger” columns respectively. Based on these two columns, the number of total passengers between two consecutive bus stops is derived in “Total passenger midway” column. The column “Comments” described any additional information.



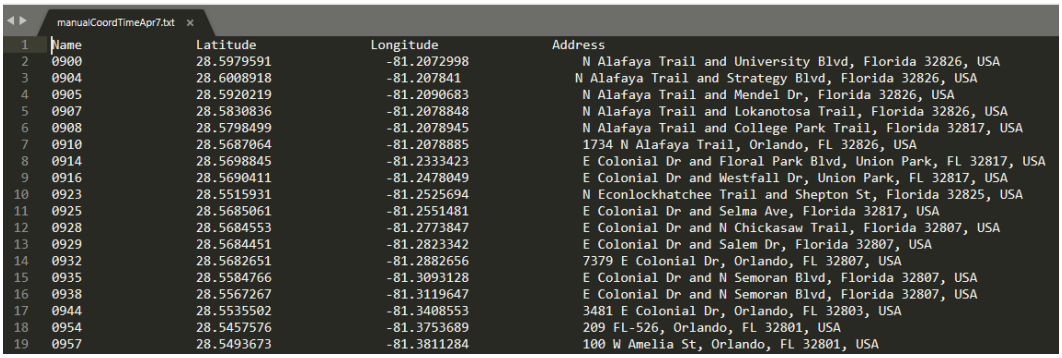
B41-410
R104

Date: 4/7 Time: 9:00

Bus Stop Index	Arrival Time	#Dropped off passenger	#Picked up passenger	Departure Time	Total passenger midway	Comments
1			3	9:00	3	UCF
2	✓9:04		2		5	
3	9:05		2		7	
4	9:07		1		8	
5	9:08		1		9	
6	9:08		1		10	
7	✓9:10		3		13	
8	9:14	1			12	
9	9:16	1	1		12	
10	9:23	3	1		14	Valencia
11	9:25		1		15	
12	9:28		1		16	
13	9:29	1			15	
14	9:32	1			14	
15	9:35	1			13	
16	9:38		1		14	
17	9:44	4			10	
18	9:54	1			9	
19	9:57	All (9)			0	Lynx C3
20						

Figure 3-17 Manual passenger count data sample for April 7, 2023 (inbound trip)

The GPS coordinate data has four columns: name, latitude, longitude, and address. The time when the bus stopped along the route in a trip was described by column “name”, whereas the location is described by columns “latitude” and “longitude”. Descriptive information of the location is described by column “address”.



manualCoordTimeApr7.txt

	Name	Latitude	Longitude	Address
2	0900	28.5979591	-81.2072998	N Alafaya Trail and University Blvd, Florida 32826, USA
3	0904	28.6008918	-81.207841	N Alafaya Trail and Strategy Blvd, Florida 32826, USA
4	0905	28.5920219	-81.2090683	N Alafaya Trail and Mendel Dr, Florida 32826, USA
5	0907	28.5830836	-81.2078848	N Alafaya Trail and Lokanotosa Trail, Florida 32826, USA
6	0908	28.5798499	-81.2078945	N Alafaya Trail and College Park Trail, Florida 32817, USA
7	0910	28.5687064	-81.2078885	1734 N Alafaya Trail, Orlando, FL 32826, USA
8	0914	28.5698845	-81.2333423	E Colonial Dr and Floral Park Blvd, Union Park, FL 32817, USA
9	0916	28.5690411	-81.2478049	E Colonial Dr and Westfall Dr, Union Park, FL 32817, USA
10	0923	28.5515931	-81.2525694	N Econlockhatchee Trail and Shepton St, Florida 32825, USA
11	0925	28.5685061	-81.2551481	E Colonial Dr and Selma Ave, Florida 32817, USA
12	0928	28.5684553	-81.2773847	E Colonial Dr and N Chickasaw Trail, Florida 32807, USA
13	0929	28.5684451	-81.2823342	E Colonial Dr and Salem Dr, Florida 32807, USA
14	0932	28.5682651	-81.2882656	7379 E Colonial Dr, Orlando, FL 32807, USA
15	0935	28.5584766	-81.3093128	E Colonial Dr and N Semoran Blvd, Florida 32807, USA
16	0938	28.5567267	-81.3119647	E Colonial Dr and N Semoran Blvd, Florida 32807, USA
17	0944	28.5535502	-81.3408553	3481 E Colonial Dr, Orlando, FL 32803, USA
18	0954	28.5457576	-81.3753689	209 FL-526, Orlando, FL 32801, USA
19	0957	28.5493673	-81.3811284	100 W Amelia St, Orlando, FL 32801, USA

Figure 3-18 Manual GPS data sample for April 7, 2023 (inbound trip)

3.4.3 Miami-Dade

The data collection for the pilot study in the Miami-Dade Transit took place in the Metromover train system, which provides frequent service. The capacity per vehicle is around 90 passengers. On average, the travel time between stations is relatively short, typically between 1 and 3 minutes. The Metromover system consists of three lines: the Omni Loop, the Brickell Loop, and the Inner Loop. Data was collected for every station on the Omni Loop and the Brickell Loop, except Freedom Tower station, as it was temporarily closed for renovations during the data collection period.

3.4.3.1 Automated Data Collection

In the Miami-Dade Transit site, automated data collection is from three different sources: sniffer devices for Wi-Fi frame data, smartphones for on-site GPS data, and Swiftly API for querying Mover location data (GPS traces).

First, a sniffer was equipped in a vehicle for capturing network traffic. The sniffer consisted of a Le Potato board, a Wi-Fi adapter, a portable monitor, a remote keyboard with a touchpad, and two power banks. Figure 3-19 illustrates the assembly diagram of the device. However, to prioritize security and passenger concerns issues during data collection, the sniffer device was concealed in a bag, except for the Wi-Fi adapter for signal reception.

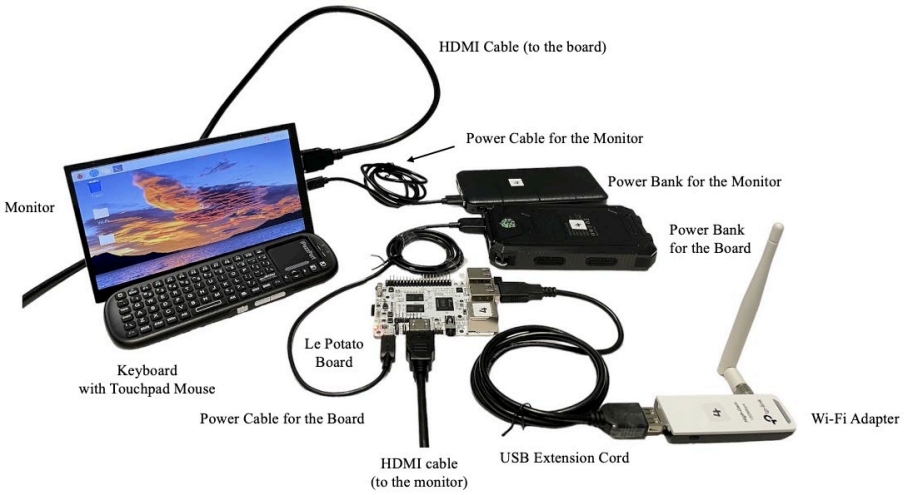


Figure 3-19 Sniffer device assembly diagram

Next, on-site GPS data were collected from smartphones installed with the necessary tracking applications. This data provided information on the geographical location and speed of the vehicles throughout their routes. For the pilot study, Android phones utilized the GPS Logger application, while iOS devices used GPS Tracks. Both applications are available for free download.

Lastly, to complement the GPS data from smartphones, we utilized the Swiftly API, a third-party service offering real-time transit data. This API provides detailed information about vehicle locations, speed, and other relevant transit data which can be utilized as a cross-referencing and a backup source of the smartphone-collected GPS data in case of any operational issues with the smartphone applications.

(a) for Android Devices (b) for iOS Devices (c) Swiftly API

Figure 3-20 GPS data source

3.4.3.2 Manual Data Collection

For manual data collection, onboard passenger count is conducted to obtain the ground truth data of vehicle occupancy. A customized passenger counting form adapted from StarMetro onboard is utilized. Figure 3-21 illustrates an example of the Omni loop On-board Passenger Check Form.

During the data collection, checkers recorded the initial number of passengers onboard and each metro mover's arrival and departure times. They also recorded the number of passengers boarding and alighting at each stop. Finally, the passenger load is calculated after the trip.

ON-BOARDPASSENGER CHECK

Checker: **Page:**

Route: OMNI LOOP **Direction:** Downtown to Omni **Start Station:** Government Center

Month: **Day:** **Year:** 2023 **Start:** AM / PM

Seats: 4 **Capacity:** 90 **On Board:**

STOP			TIME		PASSENGERS			NOTE
STOP	STATION	TRANSFER	ARR	DEP	ON	OFF	LOAD	
1	GOVERNMENT CENTER	METRORAIL						
2	THIRD STREET	BRICKELL LOOP						
3	KNIGHT CENTER							
4	BAYFRONT PARK							
5	FIRST STREET							
6	COLLEGE BAYSIDE							
7	FREEDOM TOWER							
8	PARK WEST							
9	ELEVENTH STREET							
10	MUSEUM PARK							
11	ADRIENNE ARSHT CENTER							
12	SCHOOL BOARD							
13	ADRIENNE ARSHT CENTER							
14	MUSEUM PARK							
15	ELEVENTH STREET							
16	PARK WEST							
17	FREEDOM TOWER							
18	COLLEGE NORTH	INNER LOOP						
19	WILKIE D. FERGUSON, JR	BRIGHTLINE						
20	GOVERNMENT CENTER	METRORAIL						

Figure 3-21 Omni loop passenger check form

3.4.3.3 Data Collection Schedule

The data collection schedule for the pilot study on the Omni loop and Brickell loop spanned two full weeks, from March 27 to April 9, 2023, including both weekdays and weekends. Data was collected every day within a four-hour period, specifically from 12:00 PM to 4:00 PM. During this time, two checkers were assigned per day, with one checker assigned to each loop to collect data simultaneously on each loop.

On Mondays, the schedule was adjusted due to a conflict, with the Omni loop being surveyed in the morning from 8:00 AM to 12:00 PM and the Brickell loop from 2:00 PM to 6:00 PM. It is important to note that minor adjustments were made to the schedule due to incidents or issues during the data collection. For instance, the sniffer device ran out of batteries. Moreover, some Metro mover operations disrupted the data collection process and required the checkers to re-collect the data again to ensure a complete trip, such as power problems, resulting in temporary halts. Additionally, the train was redirected from its initial loop to travel on a different loop. The travel time of one completed loop is approximately 30-40 minutes. Therefore, we aimed to complete around 3 to 4 trips per loop per day to gather sufficient data to ensure accurate and comprehensive results.

3.4.3.4 Descriptions of Collected Data

The collected data from each trip for this pilot study involved packet capture files (PCAP files) for network traffic data, manual passenger counting data, GPS data from smartphones, and GPS data from Swiftly API. During the 14-day data collection period, our primary objective was to collect around 3 to 4 trips per loop per day. We successfully achieved this goal, resulting in a total of 107 completed trips which are 52 trips from the Omni loop and 55 trips from the Brickell loop.

Each trip was associated with a single packet capture file. The size of the packet capture files varied between 5.3 MB and 40.7 MB, and the average was 27.9 MB. The variation depends on factors such as the number of passengers on board and the packet capture duration. Generally, trips with higher passenger counts tend to have larger file sizes due to the capture of more packets. Similarly, trips with longer travel times result in longer capture durations, leading to increased file sizes.

In summary, we have collected a comprehensive range of data sources. These include packet capture files, manual passenger counting data, smartphone GPS data, and GPS data obtained through the Swiftly API. We aim to enhance the accuracy and effectiveness of the occupancy estimation by combining these diverse sources of information.

CHAPTER 4. DATA-DRIVEN APPROACH

4.1 Regression

In this section, we present the conceptual design of a data-driven approach for estimating transit occupancy from collected Wi-Fi frames. For a new set of collected data such as types/subtypes, we first compute the values of those relevant features and then predict Y (occupancy). This is a data-driven approach for understanding how Y can be estimated from probe data. The true number of mobile devices on a bus is never known. We will not try to estimate or predict it. Instead, we examine a proxy, denoted Y, which is the number of observed passengers on board. For a given set of collected data (types/subtypes), we try to derive a few features to characterize or approximate X, namely X_1, X_2 etc. The target variable vehicle occupancy can be modelled as a continuous variable implying a regression task. The relationship between the features and the occupancy is described in Eq. (1).

$$Y_i = f(X_i, \beta) + e_i \qquad (1)$$

Exploratory analysis was conducted to identify the relationship between the independent variables such as average signal strength, packets count and average RSSI (Received Signal Strength Indicator) for different types and subtypes to the dependent variable (occupancy). Figure 4-1 shows the data aggregated over 5 minutes interval. The figure illustrates that overall, as the average packet count increases the average passenger head count also increases despite a few outliers that were observed.

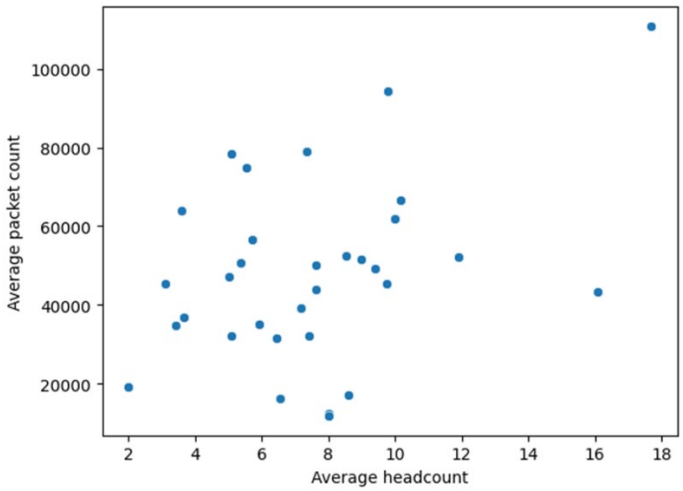


Figure 4-1 Average packet count vs. average headcount for each trip in 5-minute interval

We conducted some preliminary analysis to explore various features and their relationship to the ground truth manual passenger counts. Some examples include, the average received signal strength and the total number for packets received was compared the number of passengers. It was observed that there is a positive correlation between the number of passengers and the average signal strength as shown in Figure 4-2. As we explored our data further, we observed that as the bus got to the C.K Steele Plaza terminal, a lot of noise was captured, and several bus router packets from other buses were captured, as presented in Figure 4-3.

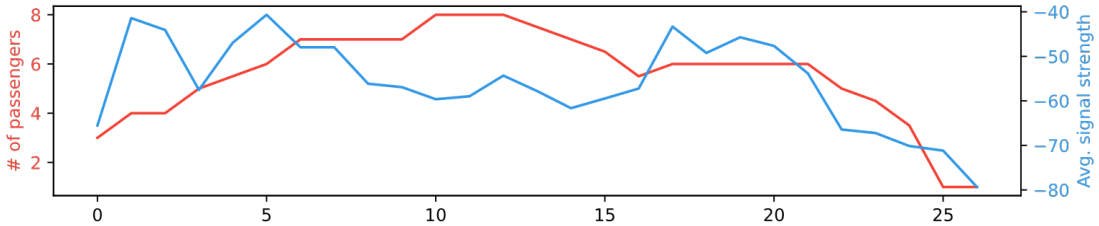


Figure 4-2 Number of passengers vs. average signal strength on April 7, 2023 (2-minute interval)

Some MAC addresses appeared in a cluster at certain times when the bus arrived at the C.K Steele Plaza terminal, and some appeared throughout the entire trip indicating onboard Wi-Fi router as shown in Figure 4-4. In our analysis, we used the OUI to remove the probe data requests that we know are related to the bus routers. The flowchart of an overview of the methodology of how data is being processed is shown in Figure 4-5. In summary, the process can be explained as follows: we first convert the Wi-Fi probe data .pcap files to .csv files and filter the needed relevant data; we then use Python to organize the data in fixed time intervals; the data was then combined with the manually collected data to generate feature variables which will be input for our machine learning model.

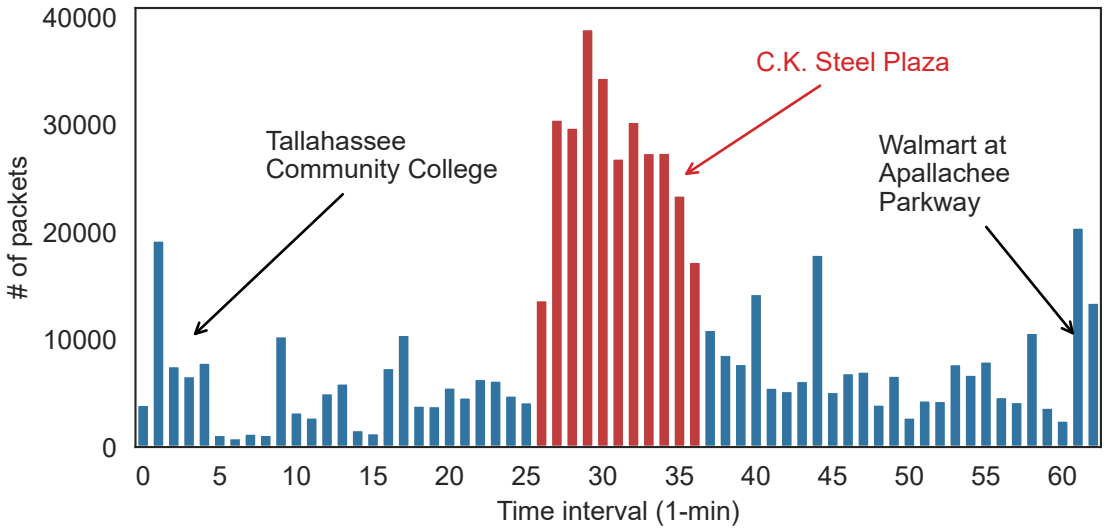


Figure 4-3 A sample of total number of packets received in relation to the bus stop

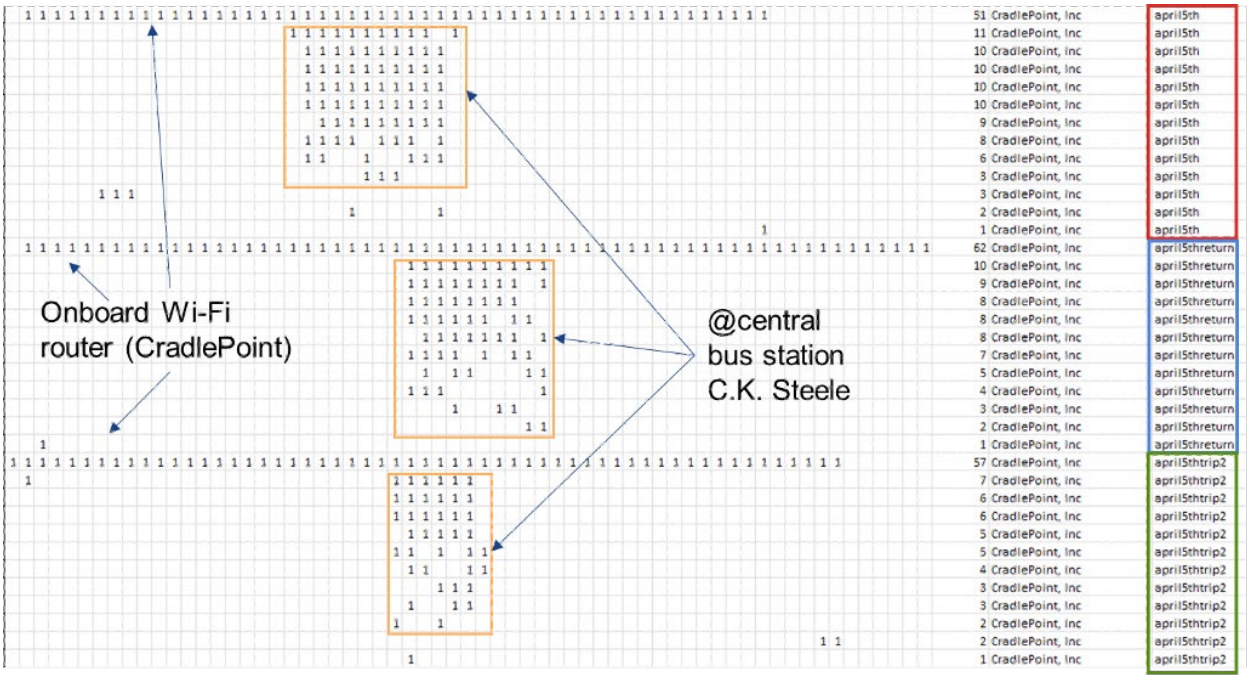


Figure 4-4 Wi-Fi routers observed through the trip on April 5, 2023

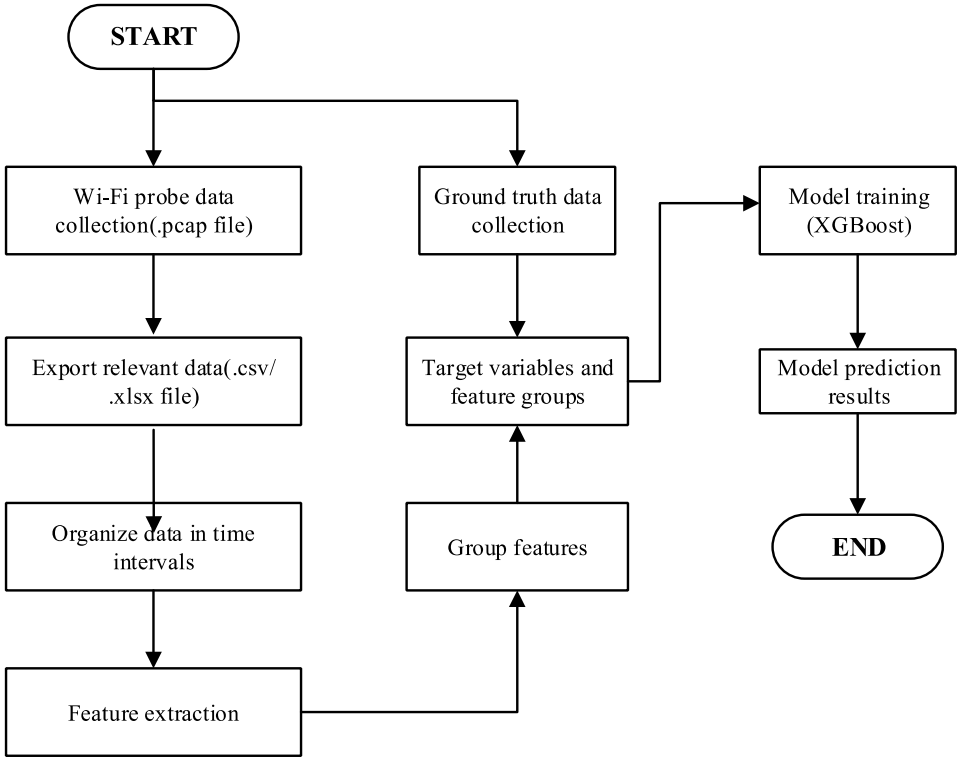


Figure 4-5 Data processing methods overview

4.2 Classification

Classification algorithms allow for the classification of data into specified classes or categories, enhancing efficient pattern identification and prediction. Machine learning classification methods are widely divided into two types: supervised learning and unsupervised learning. The most prevalent approach to classification is supervised learning, in which the algorithm is trained on labeled data, which means that each input data point is paired with a matching target label or class. The goal of this sort of learning is to create a model that can accurately predict the class labels of unknown data.

Among the most prominent supervised classification methods are:

- Logistic regression is a basic yet powerful classification method that models the link between input data and the likelihood of a binary outcome.
- Decision Trees: To classify data points, decision trees employ a hierarchical framework of decisions.
- Random Forest: A random forest is an ensemble method for improving classification accuracy and reducing overfitting by combining numerous decision trees.
- SVM: A powerful method that finds a hyperplane to split data into multiple classes while maximizing the margin between classes.
- Deep learning neural networks, specifically convolutional neural networks (CNNs) for image classification and recurrent neural networks (RNNs) for sequence data, have transformed the classification discipline.

On the other hand, unsupervised learning works with unlabeled data and tries to find underlying structures or patterns in the data. They are not conventional classification techniques, although they can be applied to tasks like clustering, which obliquely categorizes data points. Unsupervised learning methods that are widespread include:

- K-Means Clustering: K-means divides data points into clusters based on how similar they are, making it a useful tool for class discovery and segmentation.
- Hierarchical Clustering: By arranging data into a tree-like form, hierarchical clustering reveals hierarchical linkages within the data.

Classification model was also utilized in occupancy prediction which could provide the crowding level of the vehicle. For instance, the occupancy can be categorized into groups such as if the headcount is less than 5 then it's very low, if less than 10 then it's low, if less than 15 it's medium and if greater than 15 then it's high occupancy level. A new column is added as a categorical variable to indicate occupancy level.

4.3 Data Preprocessing

Data pre-processing was conducted to eliminate unnecessary information. The steps to the data preprocessing methods are as follows:

- Step 1: We removed the requests that were observed for multiple days, e.g., three days in a row that were not associated with the passengers on-board.
- Step 2: Remove requests of Cradle point which is associated with the bus routers.

- Step 3: Fill in intervals: if columns 'j1' and 'j2' are both 1 for a particular MAC, then all columns between them should be set to 1. Removing any observed short lasting and long-lasting addresses that were present throughout the trip such as the raspberry pi, drivers' device and data collector devices.

- Remove features counts that have more than 100 missing values and remove samples that have missing values.
- Burst filtering: The Wi-Fi frame requests were observed to be sent in burst. In field experiments we observed a single MAC address by turning randomization off, however we have several MAC addresses in the pilot study data. Thus, we observed different MAC addresses as shown in Figure 4-6; and how we identified them as different burst. As shown, we have two devices that send number of requests at different times. The Figure 4-7 illustrates on how the burst calculation is done; clusters of bursts are identified based on the time difference between the two consecutive requests. Then, count the number of clusters for a specific time interval for various types/subtype, signal strength filters. Figure 4-8 displays all probe requests observed in 2 minutes. Identifying bursts from all sources can be challenging; hence filtering with signal strength can help remove noises and identify bursts from nearby devices. Figure 4-9 displays bursts after applying signal strength filter of greater than -60 dBm.

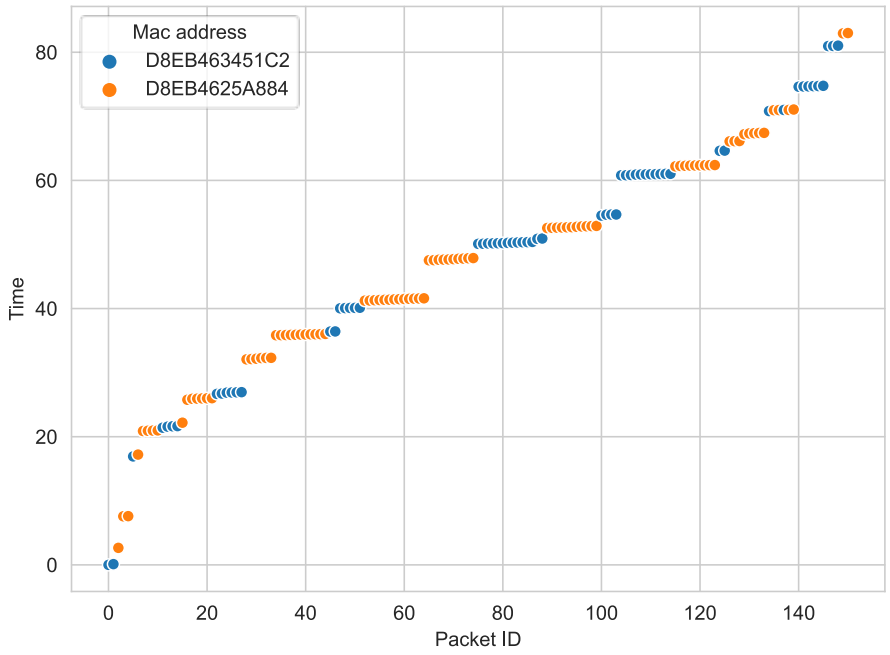


Figure 4-6 Android probe request from multiple MAC addresses

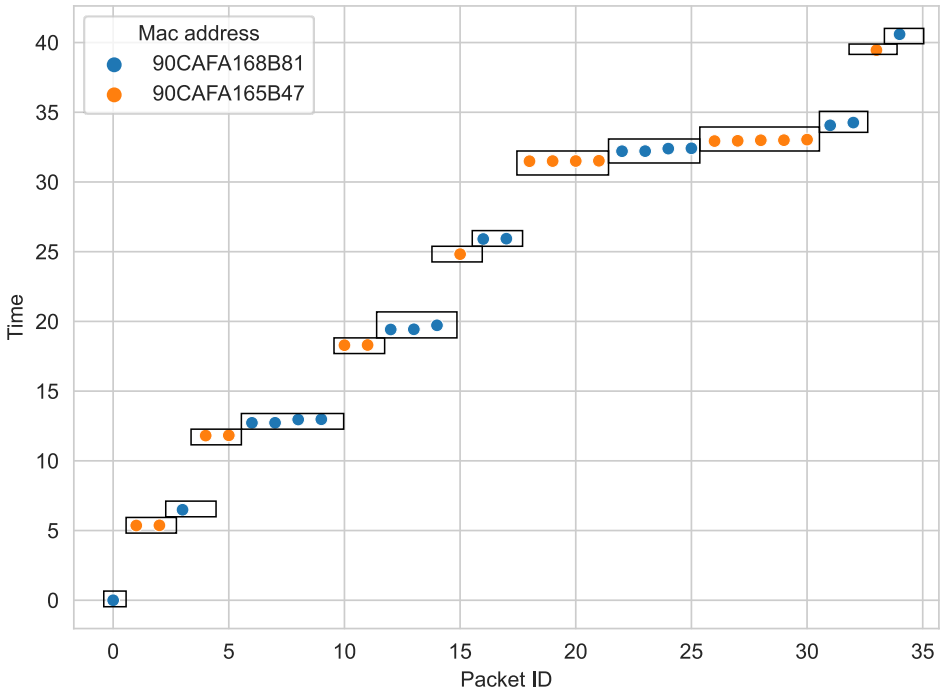


Figure 4-7 Burst count calculation by observing two different MAC addresses

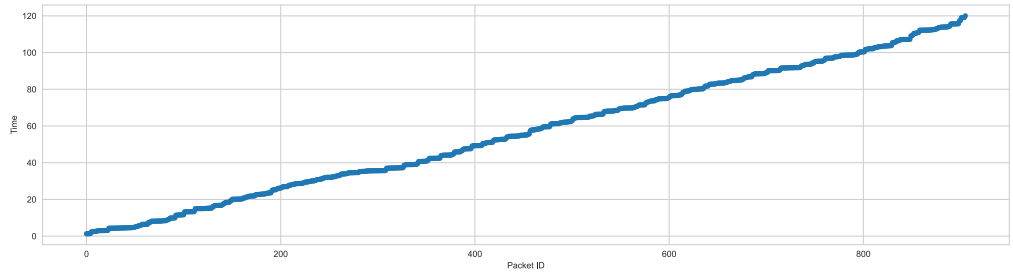


Figure 4-8 All probe requests observed in 2 minutes

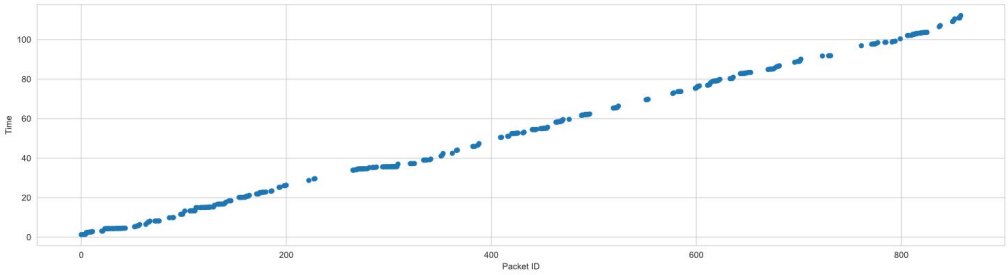


Figure 4-9 Probe requests observed in 2 minutes after applying RSSI filter greater than -60 dBm

CHAPTER 5. RESULTS AND DISCUSSION

5.1 Outdoor Experiments

5.1.1 Phase 1: Signal Strength and Distance Analysis

The results of the outdoor experiments are as follows. In the first experiment when randomization was left as default, randomized MAC addresses were observed. A notable finding was the presence of a significant number of probe requests originating from a single MAC address. Afterwards, the duplicate rows with the same MAC address were eliminated. The interval between probe requests was greater at the maximum distance compared to the shortest distance as displayed in Figure 5-1. The duration of the probe request gap was rather brief, with an average of approximately three seconds, as observed in Figure 5-2. A noticeable correlation between signal strength and distance is observed up to a distance of 20 feet. A range can be suggested for each distance based on the following average values: -45 dBm for a distance of 0 ft, -55 dBm for a distance of 10 ft, and -75 dBm for a distance of 20 ft for the first experiment as displayed in Figure 5-3. For the second experiment, as shown in Figure 5-4, there appeared to be a slight inconsistency in the pattern in signal strength range when comparing the situation with the screen turned off. In the second experiment the average range was -35 dBm for a distance of 0 ft, -55 dBm for a distance of 10 ft, and -60 dBm for a distance of 20 ft. The potential cause of this phenomenon could be attributed to a change in distance at consistent intervals of three minutes.

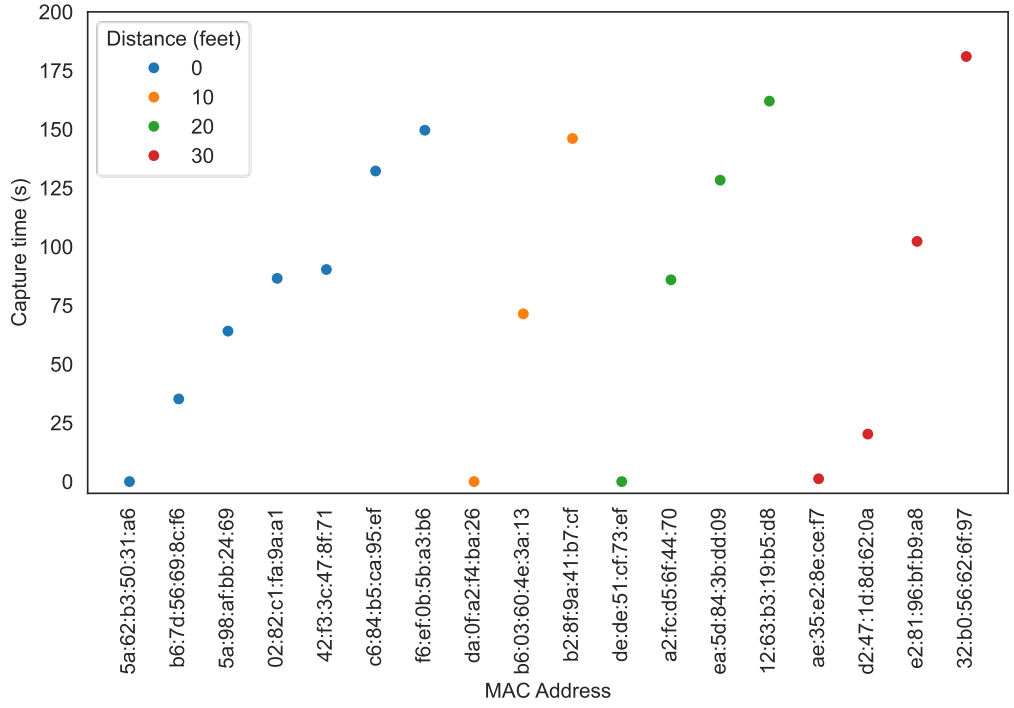


Figure 5-1 Outdoor Experiment 1 when iPhone screen is off

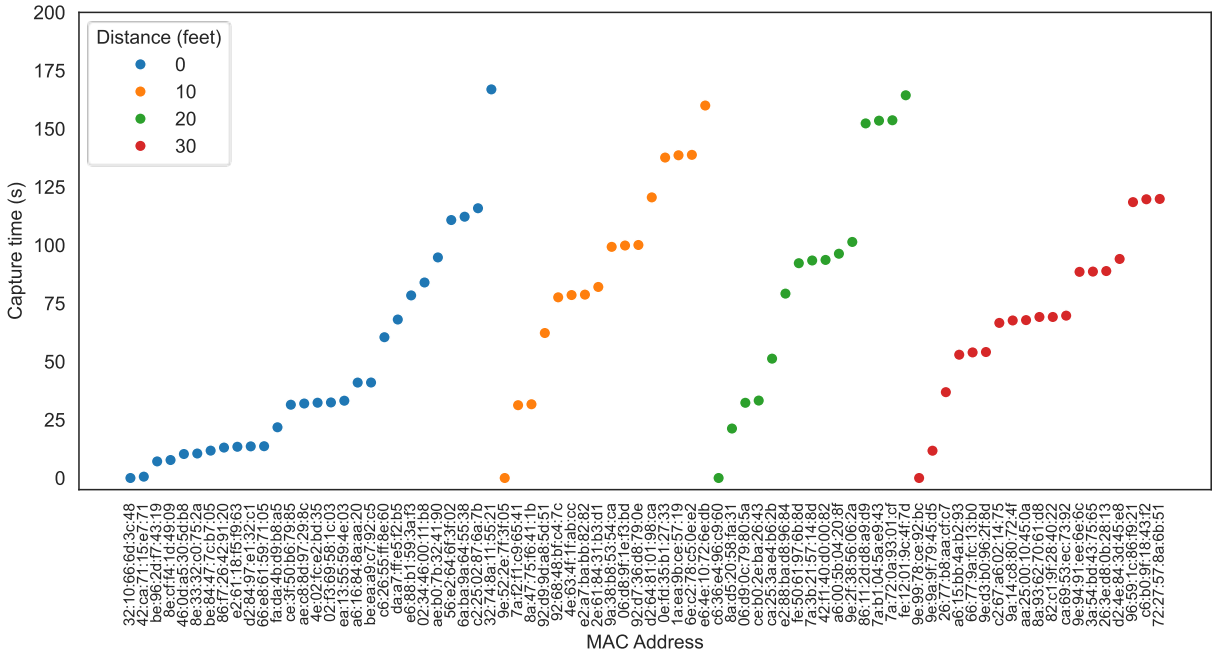


Figure 5-2 Outdoor Experiment 2 when iPhone screen is on

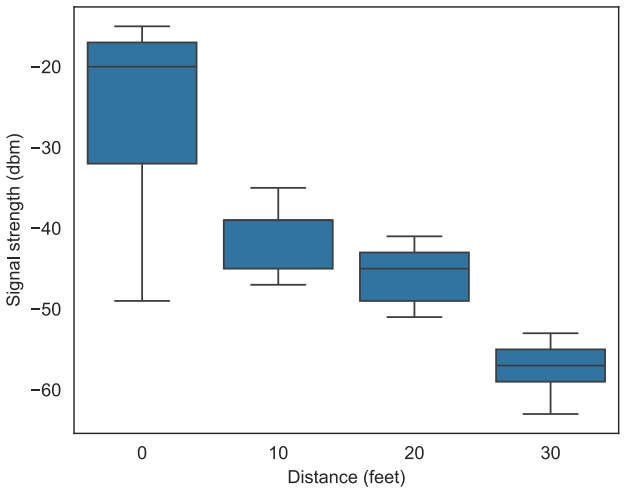


Figure 5-3 Signal strength (dBm) vs. distance (ft) when screen is off

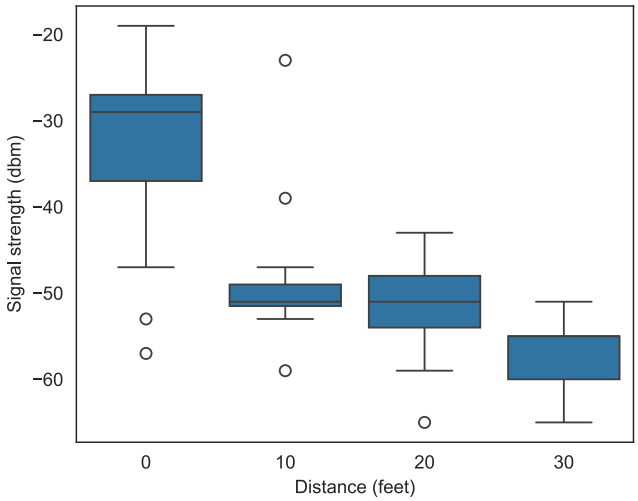


Figure 5-4 Signal strength (dBm) vs. distance (ft) when screen is on

5.1.2 Phase 2: Non-Randomized and Randomized MAC Addresses Analysis

In this section, an extensive analysis of burst filtering was carried out subsequent to the observation of the tendency for requests to be transmitted in bursts. Experiment 1 revealed that when the randomization option for the laptop's MAC address was disabled, the laptop consistently emitted probe requests at one-minute intervals while the screen was on. In contrast, no more probe requests were observed after about one minute while the screen was inactive, as shown in Figure 5-5 and Figure 5-6. Additionally, it was noteworthy that each time probe requests were detected, they manifested as a distinct burst of requests. Importantly, throughout these instances, the MAC address remained constant when randomization was disabled, signifying a stable, unchanging identifier.

When the MAC address randomization option on the laptop was enabled, the Cloud Network technology manufacturer linked with the laptop no longer appeared in the OUI (Organizationally Unique Identifier) name list. Instead, the laptop's altered MAC address had to be determined by a careful analysis of other characteristics inherent in the intercepted packets. The "TAG OUI," for example, consistently held the value 5271450 in both testing settings, regardless of MAC address randomization. The randomized MAC addresses' temporal pattern of behavior closely resembled that of the non-randomized MAC address. Even though the true MAC address was hidden while randomization was turned on, the randomized MAC address did not change over time.

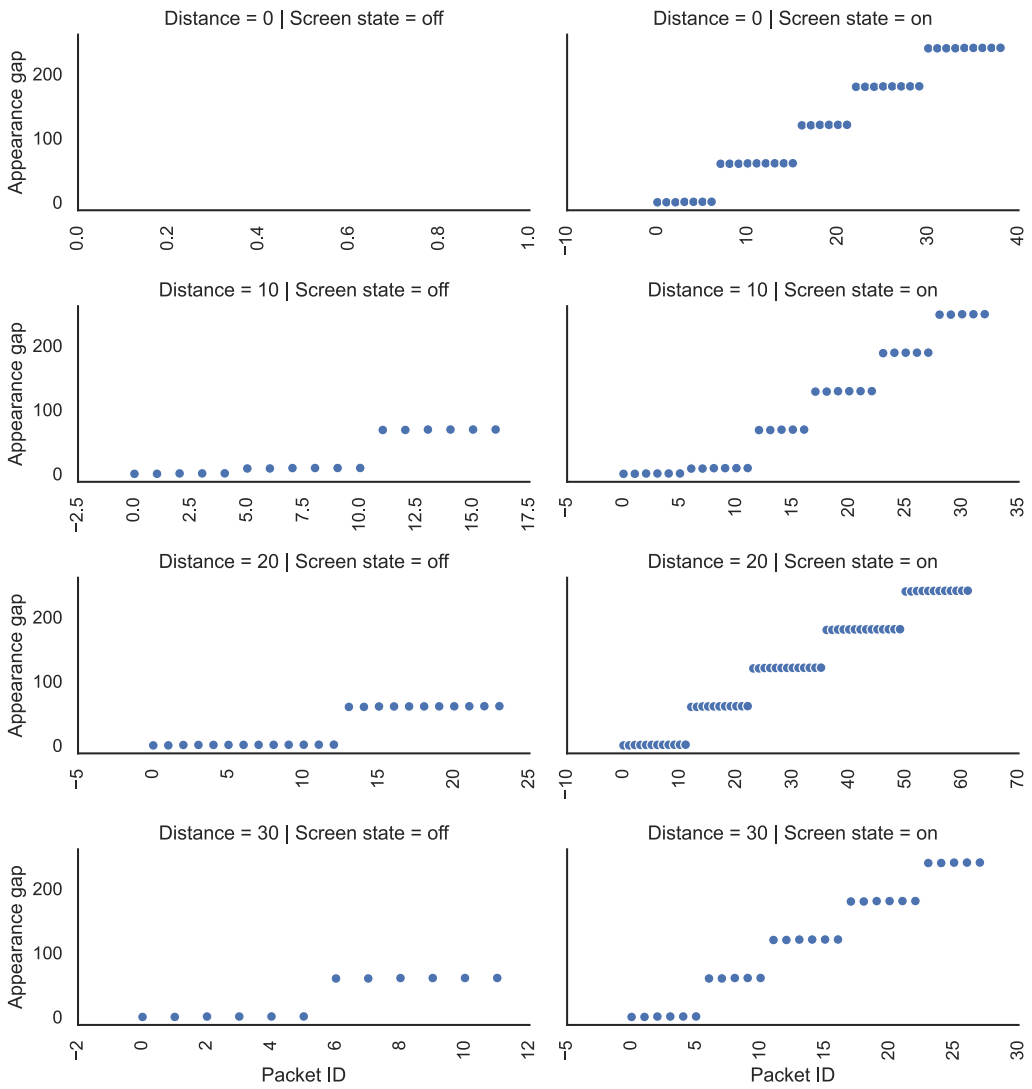


Figure 5-5 Experiment 1 laptop MAC randomization turned off

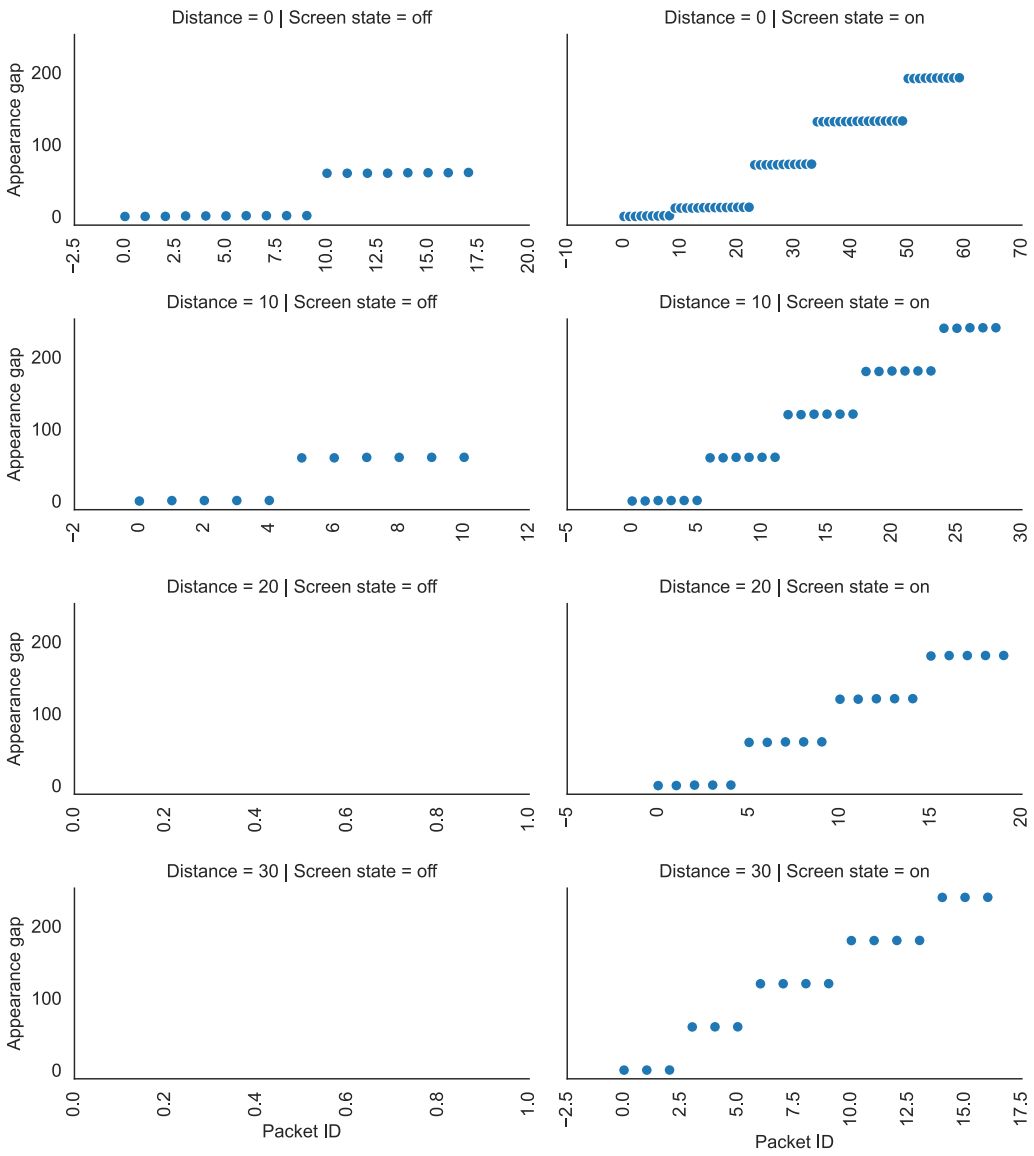


Figure 5-6 Experiment 1 laptop MAC randomization turned on

Moreover, in experiment 2, probe requests were found for only distance 0 and screen on. Thus, all types of Wi-Fi frames were considered as displayed in Figure 5-7, the requests were sent in bursts. The true MAC address was observed and did not vary over time. When the MAC randomization is on, the changed MAC address is not identifiable among noise, different MAC addresses were used even in the same burst. In experiment 3, when the Android was tested, very few probe requests are captured when the screen is off. There is no parameter that can identify the randomized MAC addresses of the iPhone and android from the database as previously it was possible to use “OUI Tags” to identify the MAC address of the laptop.

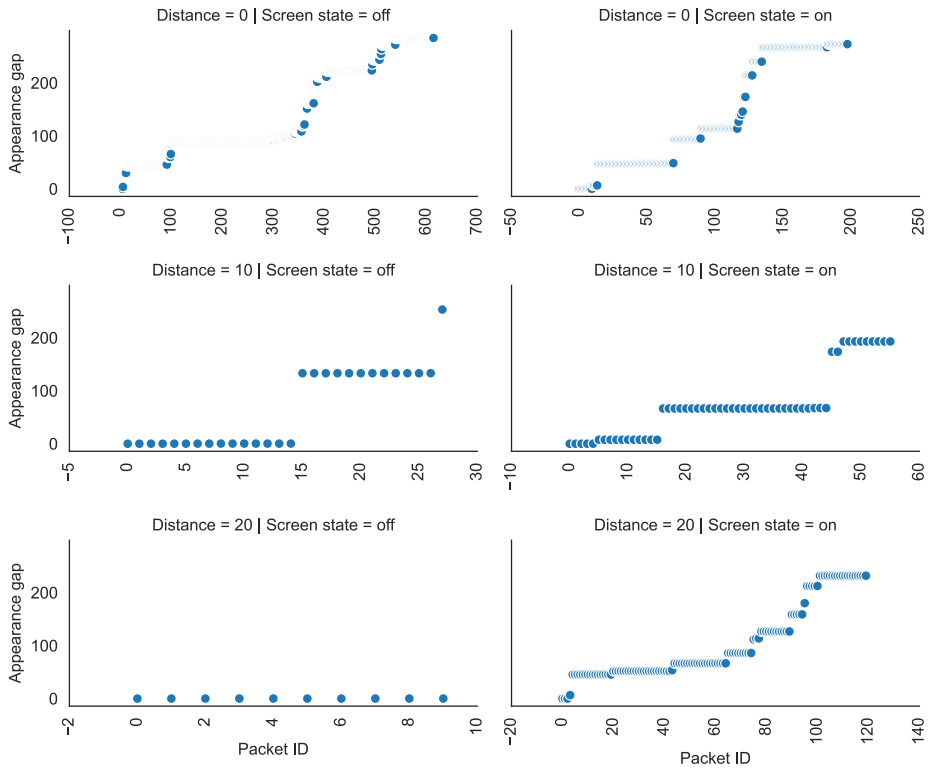


Figure 5-7 Experiment 2 iPhone MAC randomization turned off

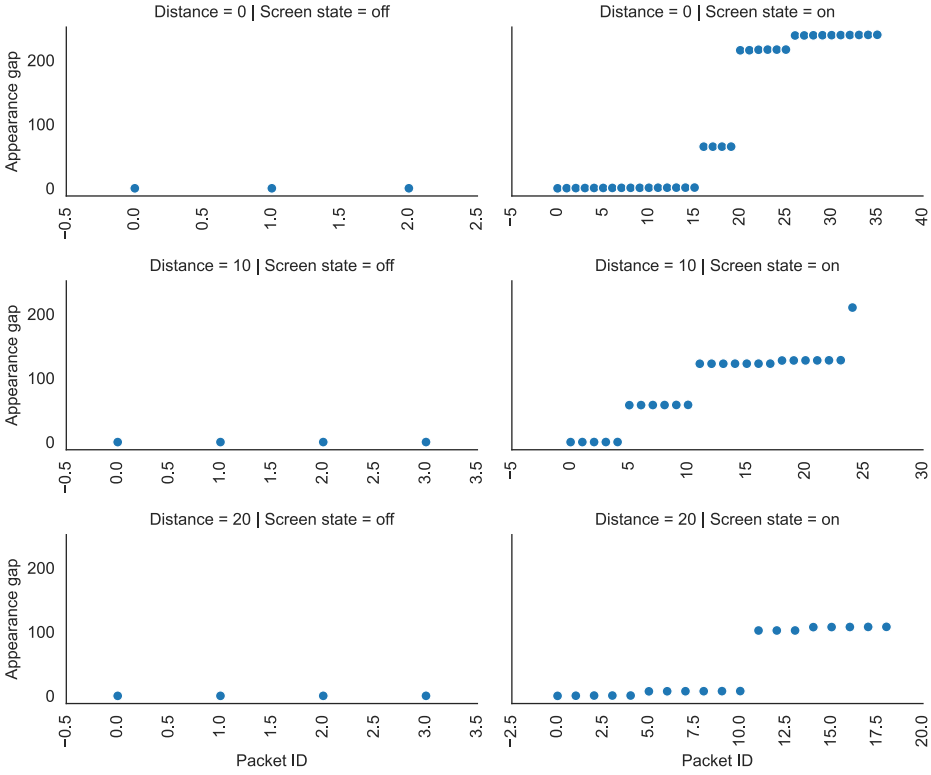


Figure 5-8 Experiment 3 Android MAC randomization turned off

5.2 StarMetro

The results are presented in two distinct subsections. In the first subsection, labeled as "MAC Addresses Presence Analysis," the focus is on the examination and presentation of MAC address-related data, encompassing aspects such as dataset description, key findings regarding MAC presence patterns, and their interpretation within the study's context. In the subsequent subsection, the focus shifts to the "Machine Learning XGBoost Model," where the details of the XGBoost algorithm practical implications of the model's results are discussed.

5.2.1 MAC Addresses Presence Analysis

The MAC presence analysis was done by considering the probe request frame to better understand MAC address randomization. The total number of days and trips used in the analysis is nine days and thirty-two trips. The total unique MAC addresses on various days are 73,828. Majority of the MAC addresses around 80% are short lived and are not recurring, while some of the MAC addresses are found in only one trip but they are long-lived, no OUI names were found for the long-lived suggesting they could be randomized Mac addresses of devices as suggested by Koç (2022). Table 5-1 shows that most of them are not trip recurring, in addition out of the 73,828, only 454 MAC count recurred for more than 10 trips. Table 5-2 presents similar results but for day recurring. In our data preprocessing, we excluded the MAC addresses that recurred for more than 3 days, these addresses were associated with non-passengers.

Furthermore, Table 5-3 represents number of unique MAC addresses that lived for particular minutes. Most of the MAC addresses are short lived, and only around 106 count lived for more than 20 minutes. The MAC presence analysis improved our prediction by including filters that eliminated the requests that were not associated with the passengers such as Cradle Point which is associated with the bus router and was observed throughout the trip. In addition, filtering data on location/stops were lots of non-passengers MAC addresses were captured which in our case was C.K Steele Plaza bus terminal. Thus, eliminating such noise influenced our model to perform better.

Table 5-1 Unique MAC Addresses Trip Recurring Count

Trip Recurrence Count	Unique MAC Addresses Count
1	70788
2	1283
3	428
4	264
5	153
6	146
7	100
8	86
9	67
10	59

Table 5-2 Unique MAC Addresses Day Recurring Count

Day Recurrence Count	Unique MAC Addresses Count
1	71139
2	1177
3	434
4	262
5	211
6	171
7	157
8	202
9	75

Table 5-3 MAC Addresses Count Present in ‘x’ Minutes

Total Minutes	MAC Addresses Count
1	71889
2	2102
3	624
4	293
5	179
6	128
7	92
8	90
9	83
10	66
11	50
12	49
13	29
14	38
15	31
16	24
17	8
18	17
19	6
20	8

5.2.2 Model Performance Prediction Results

The features were grouped into two groups considering the time intervals of 1,2 and 5 minutes. The first group contained all features consisting of a total of 99 features, which included, hour, minute, average signal strength, total count sample, average duration, signal strength of different types and subtypes, burst counts etc. The second group contained 62 features which were features related to mobile devices such as probe requests, association requests, reassociation

request and others such average signal strength and device counts for signal strength with ranged from -70 dBm to -40 dBm. Table 5-4 represents the evaluation results for the XGBoost regression model's predictive performance. Specifically, it sheds light on how well the model performs in predicting outcomes in different time intervals. For instance, when considering a 1-minute time interval and focusing on features that are directly related to mobile devices (referred to as "feature group 2"), the model achieved an R-squared value of 0.84. This R-squared value indicates that the model described approximately 84% of the variance in the data, suggesting a reasonably good fit. Additionally, the Root Mean Squared Error (RMSE), a measure of prediction accuracy, is found to be low approximately 2.46 for this same feature group and interval.

Moreover, it's important to note that feature group 2 consistently outperformed feature group 1 across different time intervals (1, 2, and 5 minutes). Feature group 2's superior performance was particularly noteworthy because it contains fewer features compared to feature group 1. For instance, when examining the 1-minute time interval, feature group 2 included a total of sixty-two features, including the burst counts. In contrast, feature group 1 comprised of ninety-nine features for the same 1-minute intervals. This difference in the number of features has practical implications for computational efficiency, as working with a smaller set of features requires less computational time.

Table 5-4 Predicted Performance of the Model

Interval group (minute)	Feature group	R2	RMSE	Full data shape
1	1	0.826967	2.58	(1361, 99)
1	**2**	**0.842697**	**2.46**	**(1361, 62)**
2	1	0.734194	3.26	(503, 104)
2	2	0.78012	3.02	(503, 63)
5	1	0.636177	3.76	(128, 108)
5	2	0.740434	3.45	(128, 64)

Figure 5-9 shows the distribution of the prediction error compared to the true values. In the figure we can see that most of the prediction deviation is around zero which indicates the prediction is close to the true values.

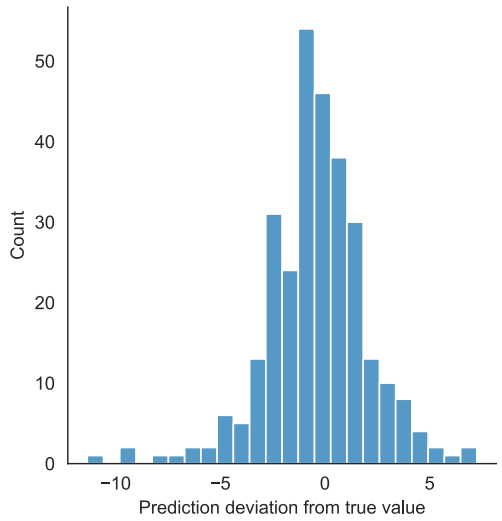


Figure 5-9 Distribution of prediction error

5.2.3 Feature Importance

Feature importance analysis is a crucial component of machine learning regression models that enables us to comprehend the relative importance of input variables (features) in predicting the dependent variable. In our analysis we also analyzed the features to help us answer questions such as: Which features have the greatest influence on the model's predictions? Which features could be regarded less significant and omitted for simplification or computational efficiency? As shown in Figure 5-10, features such as hour, sample counts and average signal strength have more influence in the prediction results of the model, however, features such as association requests and block acknowledgement can be regarded as less significant features. In machine learning regression models, feature importance analysis is a valuable tool that helps us obtain insight into the significance of input features, enhance model interpretability, and potentially improve model performance. Feature importance analysis facilitated the making of knowledgeable decisions regarding feature selection, and model refinement.

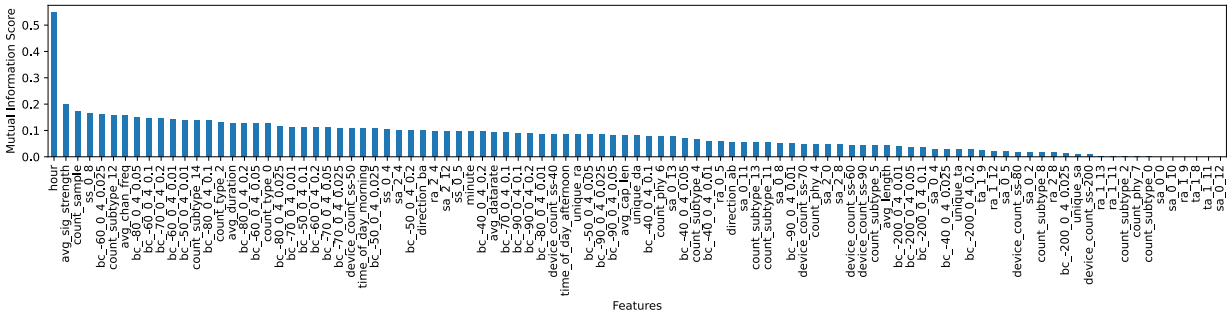


Figure 5-10 Feature importance results

5.3 Lynx

We divide the preliminary results into two segments. First, exploratory data analysis that describes relationships between different feature (i.e., independent) variables such as probe request signal,

signal strength, unique number of MAC addresses etc., and target variable which is ground truth passenger number (i.e., manual passenger count). In addition, a relation between manual passenger count and automatic passenger count (APC) data is also described. Second, the preliminary results from prediction models using machine learning algorithms are described.

5.3.1 Exploratory Data Analysis

The exploratory data analysis is conducted based on inbound trip data on April 7, 2023. A strong similarity between the number of Wi-Fi probe request signals and the unique MAC among these signals is found at different timestamps with 1-minute timestep as shown in Figure 5-11. It is found that the number of probe requests ranges from 4 to 5 times the number of unique MAC addresses.

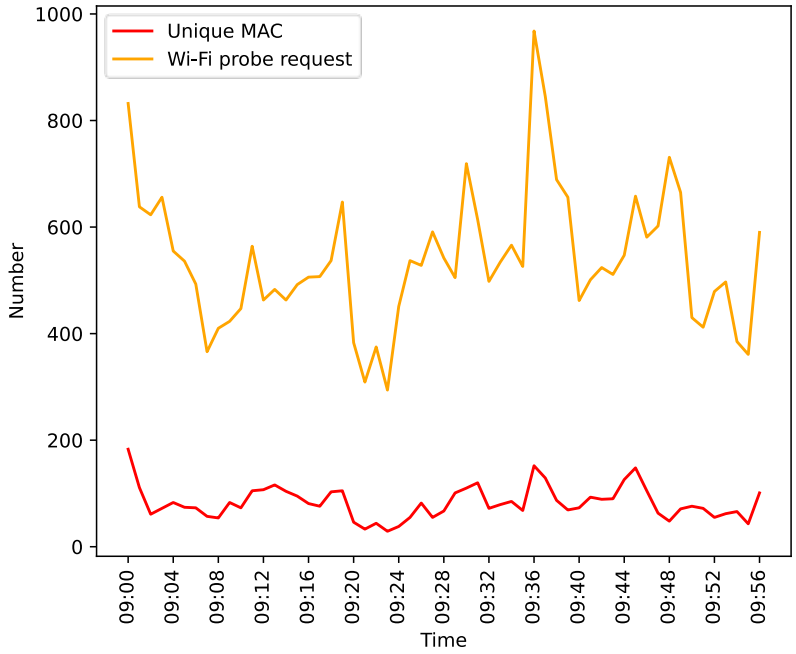


Figure 5-11 Frequency profile of probe request and unique MAC during the inbound trip on April 7, 2023

However, there is an inverse relation between the number of unique MAC addresses and the number of corresponding signal strengths. Figure 5-12 shows that from 9:00 am-9:24 am, a smaller number of unique MAC addresses with stronger signal strength are found compared with a completely inverse relation during the remaining time, 9:24 am-9:56 am.

Figure 5-12 also shows the profiles of the frequency of unique MAC addresses and ground truth passenger count (i.e., manual passenger count) at different timestamps with a 1-minute timestep. At both ends (UCF Superstop and Lynx Central Station), the number of passengers is less, whereas the number of unique MAC addresses is comparatively much higher. As these two bus stops are also bus terminals (buses from other routes start and end their entire trip at this location), foreign MAC addresses from the vicinity were collected and therefore, the number of unique MAC addresses is very high.

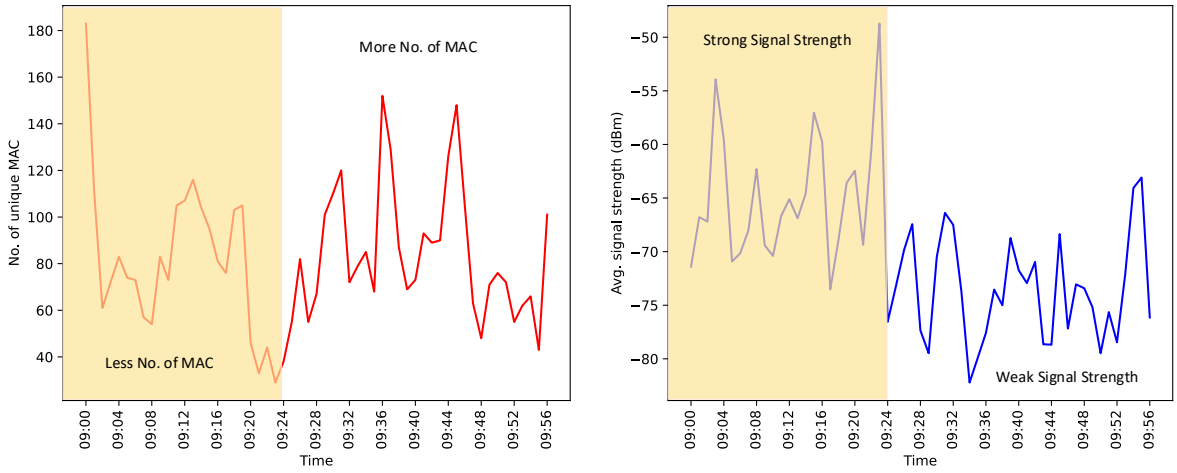


Figure 5-12 Number of MAC addresses (top) and average signal strength (bottom) along every minute of inbound trip time on April 7, 2023

Section numbers “1”, “2” and “3” in Figure 5-13 show similar trends between unique MAC addresses and passenger count. Segment “1” and “2” have a rising trend of passenger count that match with the rising trend of unique MAC number, whereas section number “3” has a falling trend in both unique MAC and passenger count. The trend of other segments does not match between unique MAC and passenger count.

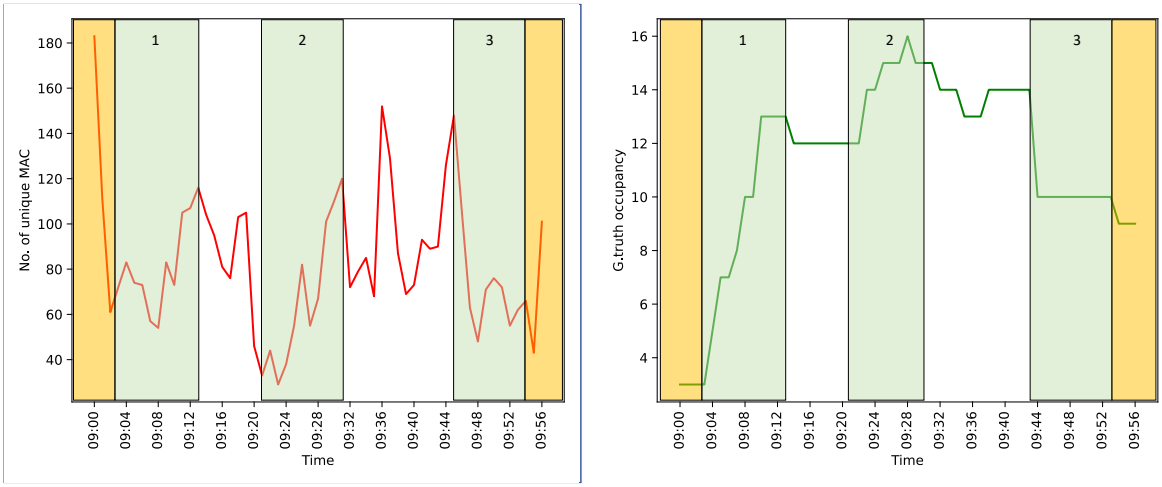


Figure 5-13 Number of unique MAC addresses (left) and ground truth passenger count (right) along every minute of inbound trip time on April 7, 2023

We also conducted a statistical approach to study the relationship among more features which is described at the beginning of the “prediction model results” sections. Based on available APC

data, Figure 5-14 is plotted to represent the relation between manual passenger count and APC data. It shows a similar trend between APC data and manual passenger count data with a positive offset multiplier that has a mean of 1.5 and a standard deviation of 0.5. As the APC data has a similar pattern to the manual count data with a small standard deviation, APC data may be easily transferred to manual passenger count data.

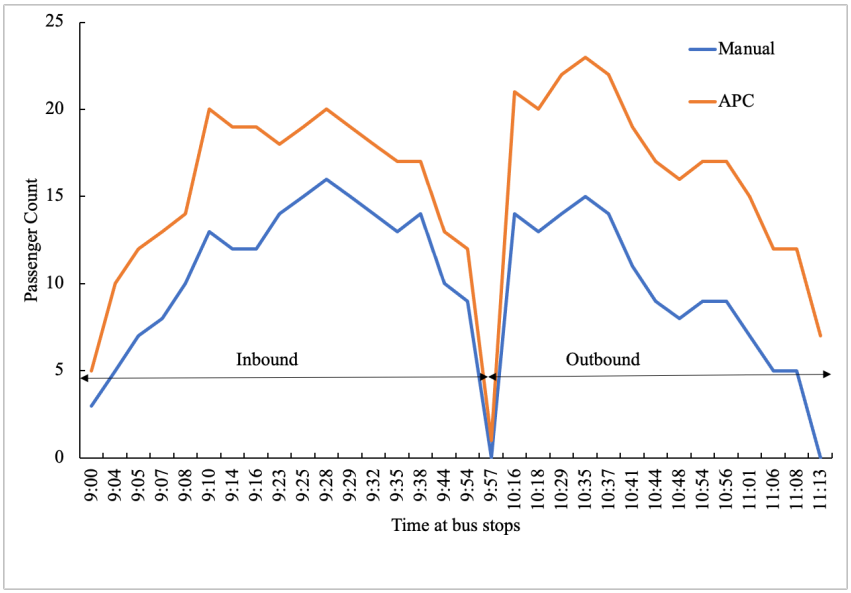


Figure 5-14 Manual passenger count and APC data comparison for inbound and outbound trips on April 7, 2023

5.3.2 Prediction Model Results

A total of 12 features representing radio signal strength, time, signal/frame types, number of recurring MAC addresses within each bound (i.e., single trip recurring MAC) and both bound directions (i.e., both trips recurring MAC), and number of all device MAC addresses are considered independent variables. Ground truth passenger count (i.e., manual passenger count) is considered a dependent variable. A 5-minute timestep is considered. From the correlation matrix, as shown in Figure 5-15, it is found that the feature "freqRecurMACWithinBound" has strong positive collinearity (1) with "stationMAC". In addition, the feature "subtype_ProbeRequest" has strong positive collinearity with "freqRecurMACWithinBound" (0.9) and "stationMAC" (0.9). Therefore, to avoid overfitting in the prediction model, features "freqRecurMACWithinBound" and "stationMAC" are omitted.

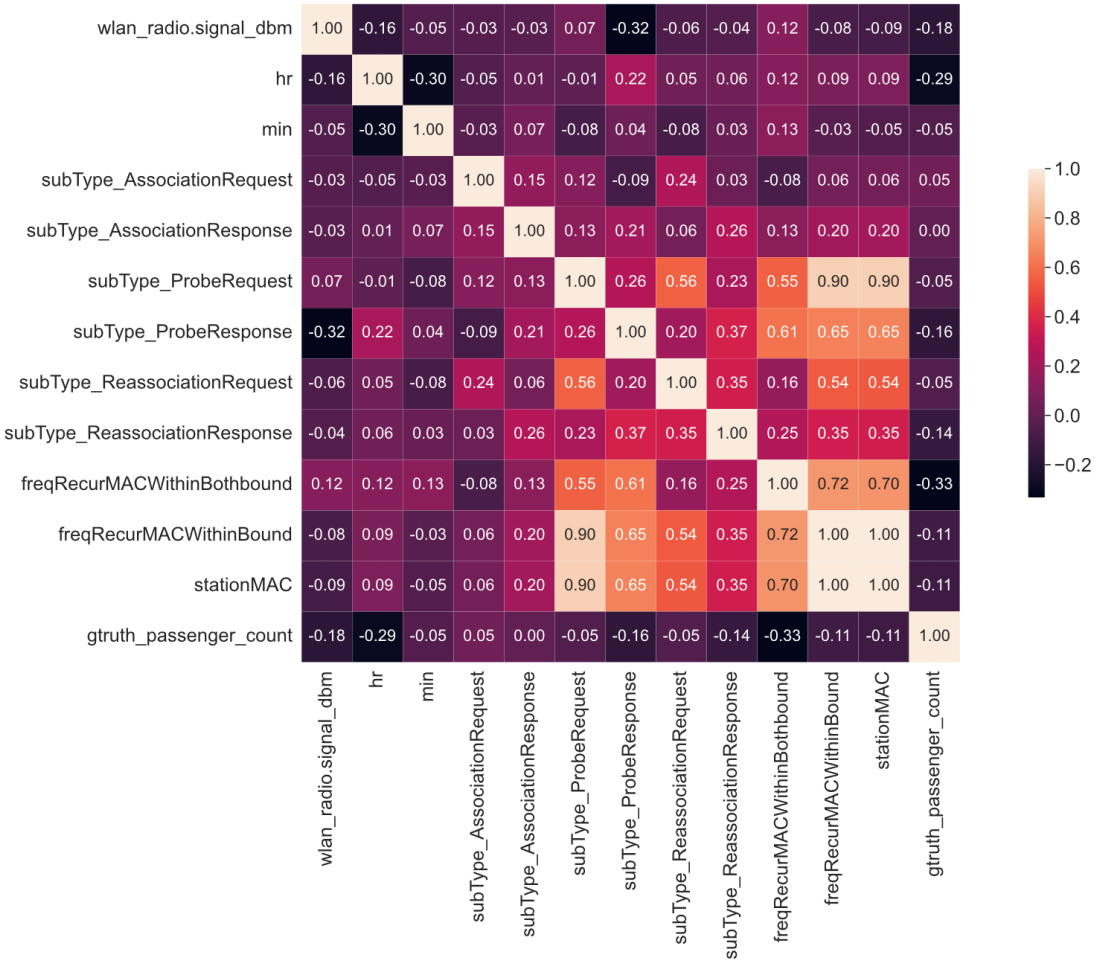


Figure 5-15 Correlation matrix of variables for the prediction model

Using sci-kit learn machine learning library four regression models: linear, polynomial with 2nd order, decision tree and random forest were developed. Besides regression models, two classification models: k-nearest neighbor (kNN) and random forest were also developed considering classification labels for passenger occupancy: almost empty (0-4), low (5-9), moderate (10-14), almost high (15-19), high (20-24) and very high (25-29).

For the performance study of these models, one dataset from March 27 to 29 and April 3 to 6 were accumulated. Considering the modified hold-out validation technique, 90% of this dataset was kept for the training model and the remaining 10% was kept for testing purposes. Another dataset combining March 30 and April 7 data was kept for validation purposes as unseen data.

Two performance metrics parameters root mean square error (RMSE) and R-Square were considered to understand the performance of regression models. On the other hand, accuracy and confusion matrix were parameters to understand the performance of classification models.

Table 5-5 Performance table for regression models

Data Characteristics			Model performance metrics	Regression model			
				Linear	Polynomial	Decision Tree	Random Forest
Train data	Mar. 27-29; Apr. 3-6	90% data	RMSE	3.85	3.10	0.00	1.70
Test data		10% data		3.50	61.62	3.59	3.05
Cross-validation data	Mar. 30 and Apr 7	100% data		3.48	5.89	4.81	3.33
Train data	Mar. 27-29; Apr. 3-6	90% data	R Square	0.26	0.52	1.00	0.86
Test data		10% data		0.17	Neg	0.12	0.37
Cross-validation data	Mar. 30 and Apr 7	100% data		Neg	Neg	Neg	0.06

In terms of RMSE, the polynomial regression model shows very poor performance in test data followed by cross-validation data compared with the remaining models. The decision tree regression model shows large overfitting. The remaining models, the linear regression model and random forest regression model (with tuned hyperparameter: max_depth=61; estimators=11) show better performance as the mean of the target variable is 11.3. However, in terms of R-square, although random forest performs better than linear, the R-Square values from the random forest and other models perform poorly for test data and cross-validation data. In other words, the regression models cannot well-explain the variance of the target variable. The results are presented in Table 5-5.

In terms of accuracy for cross-validation data by classification models, KNN (with tuned hyperparameter: n_neighbors=9) can predict 43% of observations, and the random forest model (with tuned hyperparameter: max_depth=41; n_estimators=41) can predict 51% of observations. However, random forest model shows overfitting as shown in Table 5-6.

Table 5-6 Performance table for classification models

Data Characteristics			Model performance metrics	Classification model	
				kNN	Random Forest
Train data	Mar. 27-29; Apr. 3-6	90% data	Accuracy	50%	100%
Test data		10% data		56%	56%
Cross-validation data	Mar. 30 and Apr 7	100% data		43%	51%

Two confusion matrices for classification models are shown in Table 5-7 and Table 5-8. The darker cell represents the intensity of the wrong prediction of the label. For example, if true label is “low”, and the prediction is either “almost empty” or “moderate”, it means the absolute intensity of wrong classification is 1. If the prediction is “almost high”, the absolute intensity of wrong classification is 2. That’s why the “almost high” has more yellowish cell than either “almost empty” or “moderate”.

The confusion matrix for random forest has 6 light yellowish cells with a total sum of 19 and 1 dark yellowish cell with a value of 3. On the other hand, KNN has 5 light yellowish cells with a total sum of 24 and 2 dark yellowish cells with a sum of 2. By comparison, random forest has better accuracy.

Table 5-7 Confusion matrix 1

kNN		Predicted					
		Almost empty	Low	Moderate	Almost high	High	Very high
TRUE	Almost empty	0	2	2	0	0	0
	Low	0	15	8	1	0	0
	Moderate	0	7	5	6	0	0
	Almost high	0	1	1	1	0	0
	High	0	0	0	0	0	0
	Very high	0	0	0	0	0	0

Table 5-8 Confusion matrix 2

Random Forest		Predicted					
		Almost empty	Low	Moderate	Almost high	High	Very high
TRUE	Almost empty	0	4	0	0	0	0
	Low	1	14	6	3	1	0
	Moderate	0	4	9	5	2	0
	Almost high	0	0	1	2	0	0
	High	0	0	0	0	0	0
	Very high	0	0	0	0	0	0

In both random forest regression and classification model, the number of recurring MAC within both bounds and signal strength have shown large significance in predicting. Among the management frames, probe request and probe response have strong relationships in predicting passenger occupancy as shown in Figure 5-16.

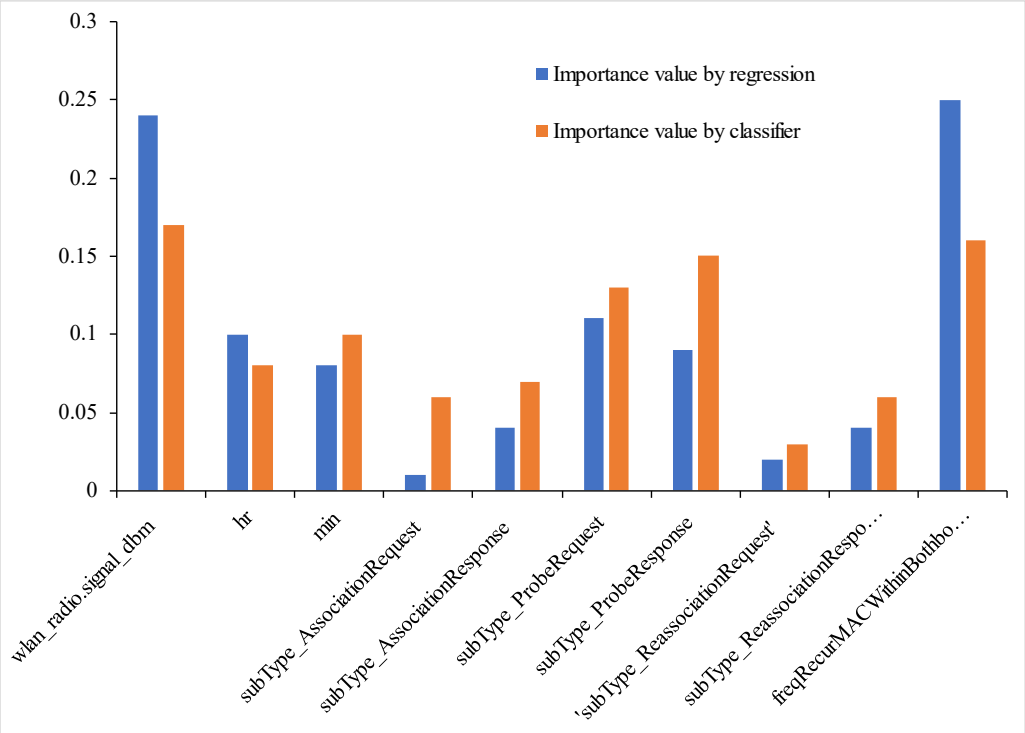


Figure 5-16 Feature importance by random forest model

5.4 Miami-Dade

5.4.1 Machine Learning Results

A series of experiments are performed, starting with Case 1, referred to as the Base Case, where no filtering was applied, and all Wi-Fi packets were used in the machine learning process. Following this, four main filtering conditions were individually applied, labeled as Case 2 to Case 5. For each case, the performance of the three-machine learning algorithm was evaluated, and the results were compared to the Base Case to measure the improvement in predictive accuracy.

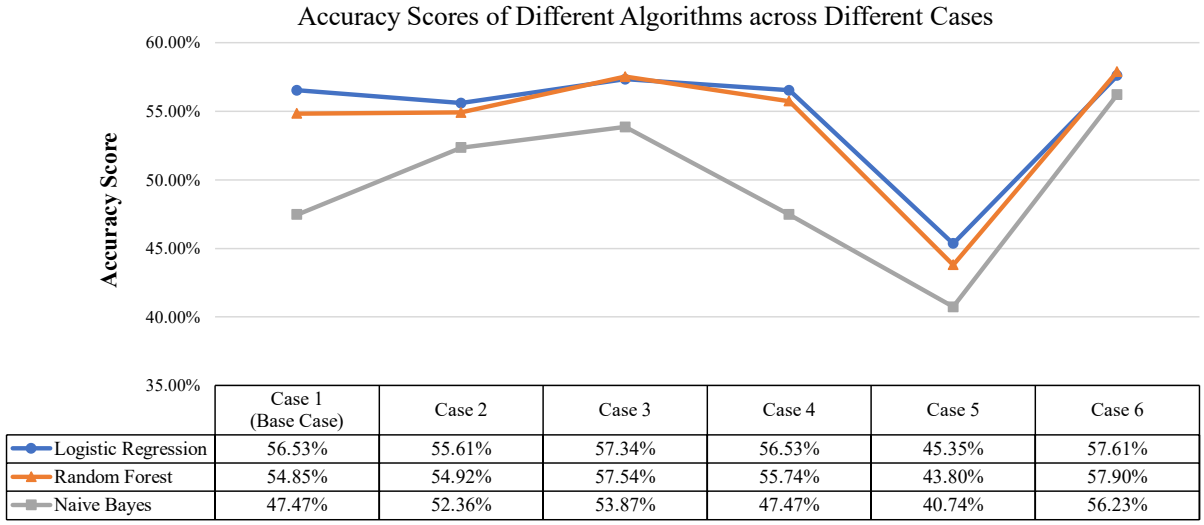
Figure 5-17 presents the results from different machine learning algorithms under various filtering conditions. After applying the filtering condition, the "Number of Features" column represents the number of selected features. "Accuracy Score" shows the average accuracy achieved, and "Standard Deviation" indicates the variability of the accuracy scores across different runs. The "Improvement (Compare to Base Case)" demonstrates how the accuracy score changed compared to the Base Case. Overall, three machine learning algorithms demonstrate a low standard deviation in all cases, with values ranging from 0.71% to 3.26%.

In Case 2, the utilization of only probe requests had varying effects on the different machine learning algorithms. Naive Bayes showed a significant improvement in performance, suggesting that this filtering condition was particularly advantageous for this algorithm. However, Logistic Regression negatively impacted its performance, and Random Forest showed only slight improvements, indicating that the probe request filtering had a limited positive effect on this

model. Among the individual filtering conditions, applying an RSSI threshold (Case 3) consistently showed improvement across all machine learning algorithms. On the other hand, some filtering conditions, like removing MAC addresses associated with "Trip Recurring in a Day" (Case 3) and selecting only packets sent by specific smartphone OUIs (Case 4), did not yield noticeable improvements. According to the individual filtering condition results, using only probe requests (Case 2) and an RSSI threshold (Case 3), which have also been utilized in other relevant studies, demonstrate their effectiveness. Hence, these two filtering conditions were combined into Case 6. The results presented in Figure 5-17 highlight the remarkable improvement achieved from Case 6. As a result, Case 6 will be utilized in the feature selection process to obtain valuable insights into Wi-Fi packets.

			Filtering Conditions				Logistic Regression			Random Forest			Naive Bayes		
Case Number	Number of Features	Dataset Size	Probe Request	RSSI ≥ -55 dBm	Remove MACs Labeled "Trip Recurring in a Day"	Specified Smartphone OUIs	Accuracy Score	Standard Deviation	Improvement (Compare to Base Case)	Accuracy Score	Standard Deviation	Improvement (Compare to Base Case)	Accuracy Score	Standard Deviation	Improvement (Compare to Base Case)
Case 1 (Base Case)	71	3,023					56.53%	2.41%	–	54.85%	2.66%	–	47.47%	3.26%	–
Case 2	32	3,050	✓				55.61%	2.05%	-1.64%	54.92%	2.11%	0.13%	52.36%	1.82%	10.31%
Case 3	68	3,050		✓			57.34%	1.60%	1.43%	57.54%	1.92%	4.91%	53.87%	2.09%	13.48%
Case 4	71	3,023			✓		56.53%	2.41%	0.00%	55.74%	2.87%	1.63%	47.47%	3.26%	0.00%
Case 5	54	3,034				✓	45.35%	2.85%	-19.78%	43.80%	0.86%	-20.13%	40.74%	0.95%	-14.18%
Case 6	31	3,050	✓	✓			57.61%	1.77%	1.90%	57.90%	2.48%	5.57%	56.23%	0.71%	18.46%

Figure 5-17 Machine learning results under different cases



	Case 1 (Base Case)	Case 2	Case 3	Case 4	Case 5	Case 6
Logistic Regression	56.53%	55.61%	57.34%	56.53%	45.35%	57.61%
Random Forest	54.85%	54.92%	57.54%	55.74%	43.80%	57.90%
Naive Bayes	47.47%	52.36%	53.87%	47.47%	40.74%	56.23%

Figure 5-18 Accuracy scores of different algorithms across different cases

5.4.2 Feature Selection

The feature selection process involved high variance filtering and moderate correlation filtering. The aim was to reduce the number of features while preserving the model's performance and predictive accuracy. The results from the feature selection process are illustrated Figure 5-19. To compare the performance of the model before and after feature selection, we first assess the accuracy achieved by using all features from Case 6 without any feature selection. Figure 5-20 shows the comparison of the accuracy across different feature selection methods. Notably, Naive Bayes consistently showed the lowest performance compared to the other algorithms, and the overall standard deviation remains low, ranging from 0.71% to 2.89%.

The feature selection process was conducted based on variation, using two variance thresholds: greater than or equal to 10 and greater than or equal to 1,000. Applying the threshold of 10 resulted in 12 selected features, while the threshold of 1,000 led to a substantial reduction to only 6 features. Surprisingly, despite the reduction from 31 to 12 features, the accuracy only experienced slight changes. However, when using the threshold of 1,000, the accuracy dropped significantly for all algorithms.

Next, the feature selection process based on the correlation coefficient was explored, involving features with a correlation value with the occupancy level greater than or equal to 0.4. The results represent a substantial reduction from the original 31 features to 9. Interestingly, reducing the number of features did not significantly impact the overall accuracy.

Feature Selection	Number of Features	Logistic Regression			Random Forest			Naive Bayes		
		Accuracy Score	Standard Deviation	Improvement Compare to (1)	Accuracy Score	Standard Deviation	Improvement Compare to (1)	Accuracy Score	Standard Deviation	Improvement Compare to (1)
(1) Without Feature Selection	31	57.61%	1.77%	–	57.90%	2.48%	–	56.23%	0.71%	–
(2) Variance ≥ 10	12	57.97%	1.81%	0.63%	57.93%	3.10%	0.06%	55.87%	2.14%	-0.64%
(3) Variance ≥ 1,000	6	56.95%	2.71%	-1.14%	57.25%	2.71%	-1.13%	54.69%	1.79%	-2.74%
(4) Correlation coefficient ≥ 0.4	9	57.74%	1.38%	0.23%	58.00%	2.89%	0.17%	56.10%	2.21%	-0.23%

Figure 5-19 Feature selection results

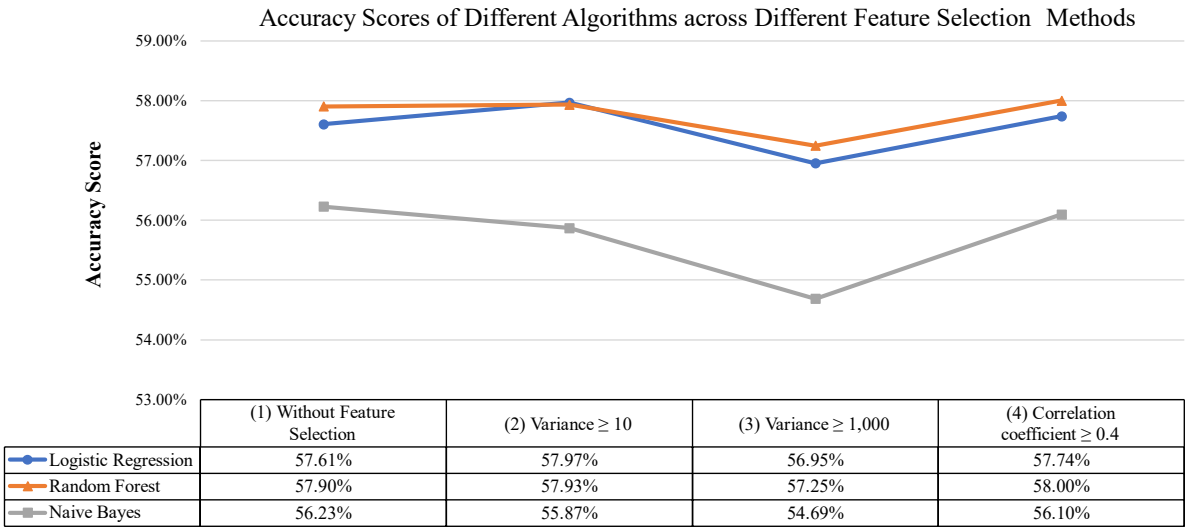


Figure 5-20 Accuracy scores of different algorithms across different feature selection methods

The selected features are listed in Table 5-5. Notably, the nine features that remain after applying correlation coefficient filtering are a subset of the twelve features that were selected through variance filtering. This implies that these nine features exhibit high variance and correlation with occupancy level, influencing the predictive model.

Table 5-9 Selected Features from Variance and Correlation Coefficient Filtering

Feature Name	Description	Feature Selection Methods	
		Variance ≥ 10	Correlation Coefficient≥ 0.4
avg_cap_len	Average capture length	✓	
avg_duration	Average duration	✓	
count_MAC_non_recur	The number of non-recurring MAC addresses	✓	✓
count_probe_request	Number of probe request frames	✓	✓
count_sample	Number of Wi-Fi frames	✓	✓
count_type_0	Number of Wi-Fi frames type 0 (management frame)	✓	✓
sa_0_4	Number of source MAC addresses with type 0, subtype 4 (probe request frame)	✓	✓
sum_sig_strength	Sum of signal strength	✓	✓
SSID_Missing	Number of probe requests contain missing SSID or Wildcard	✓	✓
SSID_Specific	Number of probe requests contain SSID name	✓	
unique_sa	Number of unique source MAC addresses	✓	✓
unique_ta	Number of unique transmitter MAC addresses	✓	✓

To better understand insights into these nine features, a correlation matrix was generated, as presented in Figure 5-25. The matrix analysis reveals strong correlations among these Wi-Fi features, which is expected due to their interrelated nature. The number of devices in the sniffer device's range affects the capture of Wi-Fi packets, resulting in more probe request frames and additional information like the count of MAC addresses and probe requests with missing SSID. Hence, these features show a positive correlation with the occupancy level. However, the sum of signal strength exhibits an inverse correlation with the occupancy level. When more active Wi-Fi devices are present, the sum of signal strength is lower, indicating a higher occupancy in the area. These strong correlations pose a significant challenge of multicollinearity when using Wi-Fi data for prediction. Eliminating redundant variables during data pre-processing might not be appropriate for this specific scenario, given the nature of Wi-Fi data.

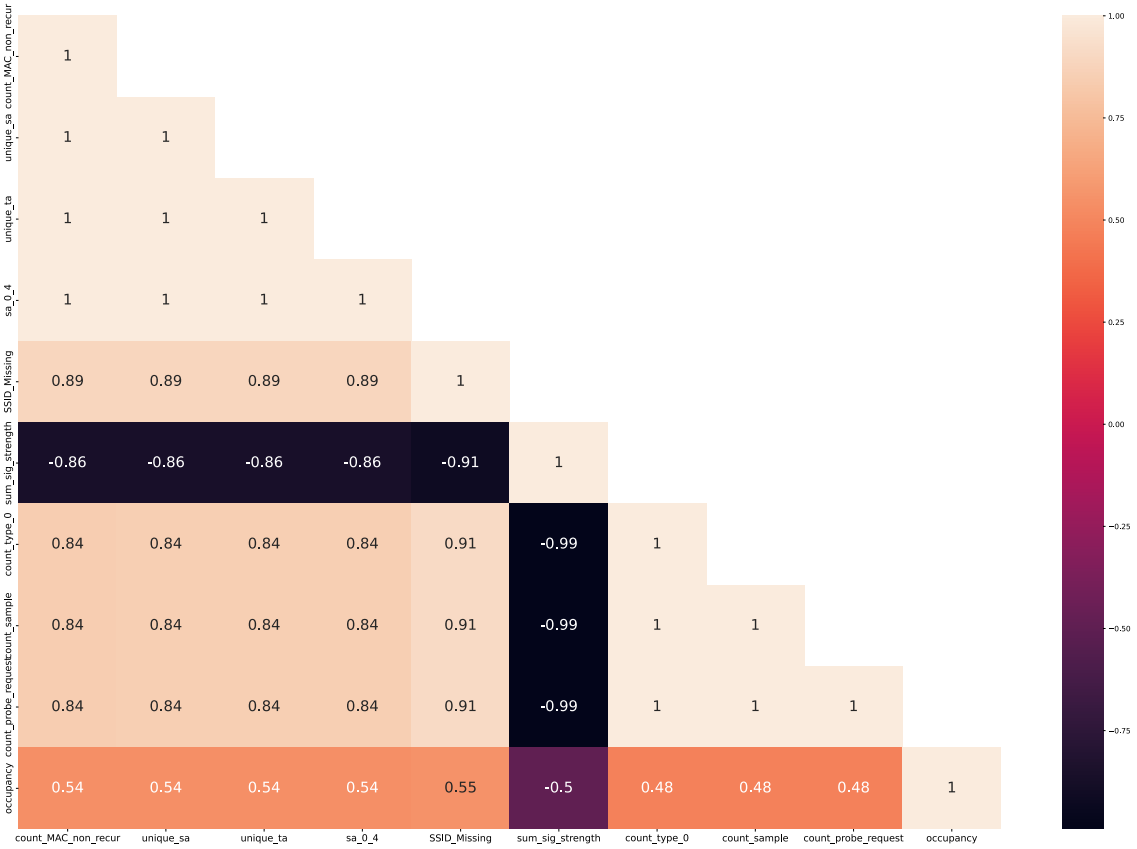


Figure 5-21 Correlation matrix

CHAPTER 6. CONCLUSIONS

This research project has first identified a number of emerging communication and information technologies that can be utilized for the estimation of transit vehicle occupancy. Subsequently, an evaluation was conducted for these technologies, taking into consideration their technical and non-technical capabilities. The comparative assessment of several technologies, including Wi-Fi, cellular data, APC, AFC, and manual survey, has revealed the potential of Wi-Fi probing technologies to estimate occupancy in real time. The utilization of Wi-Fi probing exhibits considerable potential, especially given its minimal hardware requirements and the real-time data availability. Field experiments were undertaken to evaluate the possibilities of Wi-Fi technology, and Wi-Fi technology was used in the pilot studies.

To address the technical challenge arising from MAC address randomization, a data-driven approach was utilized for vehicle occupancy estimation using the collected Wi-Fi frame data. The findings from the bus occupancy estimations in Tallahassee and Orlando Florida, indicated that the maximum R-squared value was 0.84, signifying the model's capability to explain up to 84% of the variance in the data, a noteworthy achievement. Additionally, we can observe a trend that emerges when concentrating exclusively on feature group 2, our model consistently delivers higher accuracy across all examined time intervals in comparison to the inclusion of feature group 1. These outcomes underline the effectiveness of feature selection in enhancing predictive performance. It is important to emphasize that the results obtained through the XGBoost regressor produce promising predictions. Nevertheless, there is room for further enhancement in the regression model's performance, particularly through the exploration of strategies to reduce and manage noise that may have been captured during the data collection process. Such refinements hold the potential to elevate the model's precision.

In Miami-Dade, among the models tested, random forest achieved the highest accuracy score of 58.00% with a standard deviation of 2.89%. Following closely, logistic regression secured the second position with an accuracy score of 57.74% and a standard deviation of 1.38%. Although random forest achieved the highest accuracy, it is worth noting that the accuracy level remains relatively low. Some of these reasons are associated with the operational aspects of Metromover:

- The unique partially bidirectional loop operation of Metromover presents challenges in accurately estimating occupancy levels. Certain useful techniques, like time-to-live (TTL) analysis, commonly used in bus occupancy estimation, face limitations in this scenario, especially when GPS data is unavailable. The bidirectional movement of Metromover cars leads to devices being recounted and timestamped differently at the same station, causing false calculations in the time-to-live of MAC addresses. This issue arises as the device is marked both as "first time seen" and "last time seen" during its journey on the bidirectional section.
- The proximity of neighboring stops in the transit system can lead to the loss of Wi-Fi packets. This loss occurs when the travel time between stations is shorter than the rate at which probe requests are sent, resulting in missing some Wi-Fi packet data.

Other reasons for the relatively lower accuracy that can be attributed to factors beyond the transit system itself are listed below:

- Setting one specific signal strength filtering threshold to identify whether packets are from passengers or non-passengers might be problematic, as noted by (Pu et al., 2021). As this threshold may fluctuate in different environments and timeframes.
- The accuracy of occupancy level predictions can be affected by human error in passenger counting.

Given the complexities of the Metromover system, further research and innovative approaches are needed to address the discussed challenges and limitations effectively.

For future studies, it is recommended to explore the utilization of Bluetooth data, GPS data, or the deployment of multiple sniffer devices on a vehicle to gain deeper insights. Combining these alternative data sources has the potential to enhance the accuracy and comprehensiveness of occupancy level estimation within the Metromover transit system. Such advancements would contribute significantly to the field and provide a more comprehensive understanding of passenger dynamics.

References

Aguilera, V., Allio, S., Benezech, V., Combes, F., and Milion, C. (2013). Using cell phone data to measure quality of service and passenger flows of Paris transit system. *Transport Research Part C: Emerging Technologies,* Vol 43, Part 2, pp. 198-211. DOI: 10.1016/j.trc.2013.11.007.

Algomaiah, M., & Li, Z. (2022). Utilizing Wi-Fi Sensing and an Optimized Radius Algorithm to Count Passengers with Transfers to Enhance Bus Transit OD Matrix. *Journal of Transportation Engineering, Part A: Systems*, *148*(8), 04022049. https://doi.org/10.1061/JTEPBS.0000699.

Apanasevic, T., & Rudmark, D. (2021). Crowdsourcing and Public Transportation: Barriers and Opportunities. https://www.econstor.eu/bitstream/10419/238005/1/Apanasevic-Rudmark.pdf.

Asim, M. A., Kattan, L., & Wirasinghe, S. C. (2022). Smartphone: A Source for Transit Service Planning and Management Using Wifi Sensor Data. In S. Walbridge, M. Nik-Bakht, K. T. W. Ng, M. Shome, M. S. Alam, A. El Damatty, & G. Lovegrove (Eds.), *Proceedings of the Canadian Society of Civil Engineering Annual Conference 2021*, Vol. 250, pp. 305–319. Springer Nature Singapore. https://doi.org/10.1007/978-981-19-1065-4_25.

Barbour, E., Davila, C. C., Gupta, S., Reinhart, C., Kaur, J., & González, M. C. (2019). Planning for sustainable cities by estimating building occupancy with mobile phones. *Nature Communications*, Vol 10, Issue 1, p. 3736. DOI: 10.1038/s41467-019-11685-w.

Basalamah, A. (2016). Sensing the crowds using Bluetooth low energy tags. *IEEE Access*, Vol 4, pp. 4225-4233. DOI: 10.1109/ACCESS.2016.2594210.

Boyle D. K. National Research Council (U.S.). Transportation Research Board United States Federal Transit Administration Transit Development Corporation & Transit Cooperative Research Program. (2008). *Passenger counting systems*. Transportation Research Board.

Candanedo, L. M., & Feldheim, V. (2016). Accurate occupancy detection of an office room from light, temperature, humidity and CO_2 measurements using statistical learning models. *Energy and Buildings,* Vol 112, pp. 28-39.

Chaudhary, M., Bansal, A., Bansal, D., Raman, B., et al. (2016). Finding occupancy in buses using crowdsourced data from smartphones. In *Proceedings of the 17th International Conference on Distributed Computing and Networking,* pp. 1-4. DOI: 10.1145/2833312.2833460.

Chen, C. H., Chang, Y. C., Chen, T. Y., & Wang, D. J. (2008). People counting system for getting in/out of a bus based on video processing. In *2008 Eighth International Conference on Intelligent Systems Design and Applications*,Vol. 3, pp. 565-569. IEEE.

Cui, A. (2006). *Bus passenger origin-destination matrix estimation using automated data collection systems* (Doctoral Dissertation). Massachusetts Institute of Technology, Cambridge, MA.

Darsena, D., Gelli, G., Iudice, I., & Verde, F. (2022). Sensing Technologies for Crowd Management, Adaptation, and Information Dissemination in Public Transportation Systems: A Review. *IEEE Sensors Journal*, Vol 23, Issue 1, pp. 68–87. https://doi.org/10.1109/JSEN.2022.3223297.

Dong, Z., Mokhtarian, P. L., Circella, G., & Allison, J. R. (2015). The estimation of changes in rail ridership through an onboard survey: did free Wi-Fi make a difference to Amtrak's Capitol Corridor service? *Transportation*, Vol 42, pp. 123-142.

Drabicki, A., Kucharski, R., & Cats, O. (2022). Mitigating bus bunching with real-time crowding information. *Transportation*, Vol 50, Issue 3, pp. 1-28. DOI: 10.1007/s11116-022-10270-3.

Dunlap, M., Li, Z., Henrickson, K., & Wang, Y. (2016). Estimation of Origin and Destination Information from Bluetooth and Wi-Fi Sensing for Transit. *Transportation Research Record: Journal of the Transportation Research Board*, Vol 2595, Issue 1, pp. 11–17. https://doi.org/10.3141/2595-02.

E&T. (2020). *Weight sensors to keep trains at safe capacity*. https://eandt.theiet.org/content/articles/2020/06/weight-sensors-to-keep-trains-at-safe-capacity/. Accessed July 15th, 2022.

"802.11 Association Process." (April 8, 2017) https://wifibond.com/2017/04/08/802-11- 34 association-process/ Accessed on March 21st, 2023.

El-Tawab, S., Yorio, Z., Salman, A., Oram, R., & Park, B. B. (2019). Origin-destination tracking analysis of an intelligent transit bus system using internet of things. In *2019 IEEE international conference on pervasive computing and communications workshops (PerCom Workshops)* (pp. 139-144). IEEE. DOI: 10.1109/PERCOMW.2019.8730746.

El-Tawab, S., Arai, I., Salman, A., & Park, B. B. (2020). A framework for transit monitoring system using IoT technology: two case studies. In *2020 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, pp. 1-6. IEEE. DOI: 10.1109/PerComWorkshops48775.2020.9156130.

Freudiger, J. (2015). How talkative is your mobile device? An experimental study of Wi-Fi probe requests. In *Proceedings of the 8th ACM Conference on Security & Privacy in Wireless and Mobile Networks*, pp. 1-6. https://doi.org/10.1145/2766498.2766517

Fukuda, D., H. Kobayashi, W. Nakanishi, Y. Suga, K. Sriroongvikrai, and K. Choocharukul. 2017. Estimation of paratransit passenger boarding/alighting locations using Wi-Fi based monitoring: results of field testing in Krabi City, Thailand. *Journal of the Eastern Asia Society for Transportation Studies*, Vol 12, pp. 2151-2169. https://doi.org/10.11175/easts.12.2151.

Ganz Security. (2021). *What is LiDAR and What is it Used For?* https://ganzsecurity.com/subpage/1312/ Accessed on June 30th, 2022.

Grgurević, I., Juršić, K., & Rajič, V. (2022). Review of Automatic Passenger Counting Systems in Public Urban Transport. In L. Knapčíková, D. Peraković, A. Behúnová, & M. Periša (Eds.), *5th EAI International Conference on Management of Manufacturing Systems*, pp. 1–15. Springer International Publishing. https://doi.org/10.1007/978-3-030-67241-6_1

Gu, J., Jiang, Z., Sun, Y., Zhou, M., Liao, S., & Chen, J. (2021). Spatio-temporal trajectory estimation based on incomplete Wi-Fi probe data in urban rail transit network. *Knowledge-Based Systems*, Vol 211, p. 106528. DOI: 10.1016/j.knosys.2020.106528.

Hakegard, J. E., Myrvoll, T. A., & Skoglund, T. R. (2018, November). Statistical modelling for estimation of OD matrices for public transport using Wi-Fi and APC data. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pp. 1005-1010. IEEE.

Hashimoto, M., Tsuji, A., Nishio, A., & Takahashi, K. (2015). Laser-based tracking of groups of people with sudden changes in motion. In *2015 IEEE International Conference on Industrial Technology (ICIT)*, pp. 315-320. IEEE.

Haywood, L., Koning, M., and Monchambert, G. (2017). Crowding in public transport: Who cares and why?. *Transportation Research Part A: Policy and Practice*, Vol 100, pp. 215-227. DOI: 10.1016/j.tra.2017.04.022.

Hidayat, A., Terabe, S., & Yaginuma, H. (2020). Bus Passenger Volume and Origin-Destination Based on Field Surveys Using a Wi-Fi Scanner. *Transportation Research Procedia*, Vol 48, pp. 1376–1389. https://doi.org/10.1016/j.trpro.2020.08.169.

Hsu, Y. W., Chen, Y. W., & Perng, J. W. (2020). Estimation of the number of passengers in a bus using deep learning. *Sensors*, Vol 20, Issue 8, p. 2178.

Jain, J., Bartle, C., Clayton, W., and Lane, C. (2018). Continuously Connected Customer: WiFi on Trains. *UWE Bristol*. https://uwe-repository.worktribe.com/output/871578/continuously-connected-customer-final-project-report, Accessed on May 27th, 2022.

Jang, W. (2010). Travel Time and Transfer Analysis Using Transit Smart Card Data. *Transportation Research Record: Journal of the Transportation Research Board*, Vol 2144, Issue 1, pp. 142–149. https://doi.org/10.3141/2144-16.

Jenelius, E. (2020). Data-Driven Metro Train Crowding Prediction Based on Real-Time Load Data. *IEEE Transactions on Intelligent Transportation Systems*, Vol 21, Issue 6, pp. 2254–2265. https://doi.org/10.1109/TITS.2019.2914729.

Ji, Y., Zhao, J., Zhang, Z., & Du, Y. (2017). Estimating bus loads and OD flows using location-stamped farebox and Wi-Fi signal data. *Journal of Advanced Transportation*, *2017*.

Jiang, C., Masood, M. K., Soh, Y. C., and Li, H. (2016). "Indoor occupancy estimation from carbon dioxide concentration*." Energy and Buildings,* Vol 131, pp. 132–141. DOI: 10.1016/j.enbuild.2016.09.002.

Jiang, Y., Miao, Y., Alzahrani, B., Barnawi, A., Alotaibi, R., & Hu, L. (2021). Ultra large-scale crowd monitoring system architecture and design issues. *IEEE Internet of Things Journal*, Vol 8, Issue 13, pp. 10356-10366.

Junior, J. C. S. J., Musse, S. R., & Jung, C. R. (2010). Crowd analysis using computer vision techniques. *IEEE Signal Processing Magazine*, Vol 27, Issue 5, pp. 66-77. DOI: 10.1109/MSP.2010.937394.

Junior, M. P. R., and R. A. Medrano. (2018). "32° ANPET." In UAI-FI: Ummétodo baseado em aprendizado de máquina para contagem automática de passageiros utilizando sinais Wi-Fi. Gramado, Brasil: ANPET.

Junior, M. P. R., Medrano, R. M. A., & Almeida, C. F. (2022). UAI-FI: Using artificial intelligence for automatic passenger counting through Wi-Fi and GPS data. *TRANSPORTES*, Vol 30, Issue 2, pp. 2555-2555. https://doi.org/10.14295/transportes.v30i2.2555.

Kang, L., Qi, B., & Banerjee, S. (2016). A Wireless-Based Approach for Transit Analytics. In *Proceedings of the 17th International Workshop on Mobile Computing Systems and Applications*, pp. 75-80. DOI: 10.1145/2873587.2873589.

Kannan, P. G., Venkatagiri, S. P., Chan, M. C., Ananda, A. L., & Peh, L. S. (2012). Low cost crowd counting using audio tones. In *Proceedings of the 10th ACM Conference on Embedded Network Sensor Systems,* pp. 155-168. DOI: 10.1145/2426656.2426673.

Khomchuk, P., Tuladhar, S. R., & Sivananthan, S. (2018). *Predicting passenger loading level on a train car: A Bayesian approach* (arXiv:1808.06962). arXiv. http://arxiv.org/abs/1808.06962.

Kim, J., Preston, J., & Revell, K. (2019). Investigating the effect of train occupancy information. https://publications.ergonomics.org.uk/uploads/Investigating-the-effect-of-train-occupancy-information-.pdf, Accessed May 27th, 2022.

Koç, F. M. (2022). *Occupancy Detection in Indoor Environments Based on Wi-Fi Measurements and Machine Learning Methods*.

Kostakos, V., Camacho, T., & Mantero, C. (2010). Wireless detection of end-to-end passenger trips on public transport buses. In *13th International IEEE Conference on Intelligent Transportation Systems,* pp. 1795-1800, IEEE. DOI: 10.1109/ITSC.2010.5625062.

Kouyoumdjieva, S. T., Danielis, P., & Karlsson, G. (2020). Survey of non-image-based approaches for counting people. *IEEE Communications Surveys & Tutorials*, Vol 22, Issue 2, pp. 1305-1336. DOI: 10.1109/COMST.2019.2902824.

Lazo, L., (2017) "*Start-ups finding footing with crowdsourced bus services in cities with ailing transit."*https://www.washingtonpost.com/local/trafficandcommuting/start-ups-find-footing-with-crowdsourced-bus-service-in-cities-with-ailing-transit/2017/09/02/c8920cca-8c1c-11e7-91d5-ab4e4bb76a3a_story.html, Accessed on May 27th, 2022.

Lengvenis, P., Simutis, R., Vaitkus, V., & Maskeliunas, R. (2013). Application of computer vision systems for passenger counting in public transport. *Elektronika ir Elektrotechnika*, Vol 19, Issue 3, pp. 69-72.

Lesani, A., Nateghinia, E., & Miranda-Moreno, L. F. (2020). Development and evaluation of a real-time pedestrian counting system for high-volume conditions based on 2D LiDAR. *Transportation Research Part C: Emerging Technologies*, Vol 14, pp. 20-35. DOI: 10.1016/j.trc.2020.01.018.

Liu, D., Sheng, B., Hou, F., Rao, W., & Liu, H. (2014). From Wireless Positioning to Mobile Positioning: An overview of Recent Advances. *IEEE Systems Journal*, Vol 8, Issue 4, pp. 1249-1259. DOI: 10.1109/JSYST.2013.2295136.

Li, K., Yuen, C., Kanhere, S. S., Hu, K., Zhang, W., Jiang, F., & Liu, X. (2016). *SenseFlow: An Experimental Study for Tracking People* (arXiv:1606.03713). arXiv. http://arxiv.org/abs/1606.03713.

Liu, G., Yin, Z., Jia, Y., & Xie, Y. (2017). Passenger flow estimation based on convolutional neural network in public transportation system. *Knowledge-Based Systems*, Vol 123, pp. 102-115. DOI: 10.1016/j.knosys.2017.02.016.

Masters, M. R., Healy, R. M., Torres, A. D., & Fraley, R. L. (2003). Passenger Counting and Service Monitoring: A Worldwide Survey of Transportation Agency Practices. Accessed on June 24th, 2022. https://rosap.ntl.bts.gov/view/dot/42145.

Mehmood, U., Moser, I., Jayaraman, P. P., and Banerjee, A. (2019). Occupancy estimation using WiFi: A case study for counting passengers on busses. In *2019 IEEE 5th World Forum on Internet of Things (WF-IoT)*, pp. 165-170. IEEE. DOI: 10.1109/WF-IoT.2019.8767350.

Mikkelsen, L., Buchakchiev, R., Madsen, T., and Schwefel, H. P. (2016). "Public transport occupancy estimation using WLAN probing." *8th International Workshop on Resilient Networks Design and Modeling (RNDM),* pp 302-308*, IEEE.* DOI: 10.1109/RNDM.2016.7608302.

Mishalani, R. G., McCord, M. R., & Reinhold, T. (2016). Use of Mobile Device Wireless Signals to Determine Transit Route-Level Passenger Origin–Destination Flows: Methodology and Empirical Evaluation. *Transportation Research Record: Journal of the*

Transportation Research Board, Vol 2544, Issue 1, pp. 123–130. https://doi.org/10.3141/2544-14.

Moovit (2021) Avoiding the Crowd: Know Before You Go with the Moovit App's Crowdedness Feature, https://moovit.com/blog/avoiding-the-crowd-know-before-you-go-with-the-moovit-apps-crowdedness-feature/ Accessed on May 27th, 2022.

Moser, I., McCarthy, C., Jayaraman, P. P., Ghaderi, H., Dia, H., Li, R., ... and Fuss, F. K. (2019). A methodology for empirically evaluating passenger counting technologies in public transport. In *41st Australasian Transport Research Forum (ATRF)*. https://www.australasiantransportresearchforum.org.au/papers/2019.

Mukheja, P., Kiran K, M., Velaga, N. R., and Sharmila, R. B. (2017). Smartphone-based crowdsourcing for position estimation of public transport vehicles. *IET Intelligent Transport Systems*, Vol 11, Issue 9, pp. 588-595. DOI: 10.1049/iet-its.2016.0247.

Murdan, A. P., Bucktowar, V., Oree, V., & Enoch, M. P. (2020). Low-cost bus seating information technology system. *IET Intelligent Transport Systems*, Vol 14, Issue 10, pp. 1303–1310. https://doi.org/10.1049/iet-its.2019.0529.

Myrvoll, T. A., Hakegard, J. E., Matsui, T., & Septier, F. (2017). Counting public transport passenger using WiFi signatures of mobile devices. 2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC), pp. 1–6. https://doi.org/10.1109/ITSC.2017.8317687.

Nielsen, B. F., Frølich, L., Nielsen, O. A., & Filges, D. (2014). Estimating passenger numbers in trains using existing weighing capabilities. *Transportmetrica A: Transport Science*, *10*(6), 502–517. https://doi.org/10.1080/23249935.2013.795199.

Nitti, Pinna, F., Pintor, L., Pilloni, V., and Barabino, B. (2020). "iABACUS: A Wi-Fi-Based Automatic Bus Passenger Counting System." *Energies (Basel),* Vol 13, Issue 6, p. 1446. DOI: 10.3390/en13061446.

Noursalehi, P., Koutsopoulos, H. N., & Zhao, J. (2021). Predictive decision support platform and its application in crowding prediction and passenger information generation. *Transportation Research Part C: Emerging Technologies,* Vol 129, p. 103139.

Oliveira, L., Schneider, D., De Souza, J., & Shen, W. (2019). Mobile Device Detection Through WiFi Probe Request Analysis. *IEEE Access*, Vol *7*, pp. 98579–98588. https://doi.org/10.1109/ACCESS.2019.2925406.

Oransirikul, T., Piumarta, I., and Takada, H. (2019). Classifying passenger and non-passenger signals in public transportation by analysing mobile device Wi-Fi activity. *Journal of Information Processing*, Vol 27, pp. 25-32. DOI: 10.2197/ipsjjip.27.25.

Özgün, K., Günay, M., Başaran, B. D., & Ledet, J. W. (2022). Estimation of Alighting Counts for Public Transportation Vehicle Occupancy Levels Using Reverse Direction Boarding. *Available at SSRN 4113026*.

Paradeda, D. B., Junior, W. K., & Carlson, R. C. (2019). Bus passenger counts using Wi-Fi signals: some cautionary findings. *Transportes*, Vol 27, Issue 3, pp. 115-130.

Paradeda, D. B., Kraus Jr, W., Castelan Carlson, R., & Seman, L. O. (2023). Bayesian Estimation of Passenger Boardings at Bus Stops Using Wi-Fi Probe Requests. *Journal of Transportation Engineering, Part A: Systems*, Vol 149, Issue 6, pp. 04023045.

Pattanusorn, W., Nilkhamhang, I., Kittipiyakul, S., Ekkachai, K., & Takahashi, A. (2016). Passenger estimation system using Wi-Fi probe request. In *2016 7th International*

Conference of Information and Communication Technology for Embedded Systems (IC-ICTES), pp. 67-72. IEEE. DOI: 10.1109/ICTEmSys.2016.7467124.

Pelletier, M. P., Trépanier, M., & Morency, C. (2011). Smart card data use in public transit: A literature review. *Transportation Research Part C: Emerging Technologies*, Vol 19, Issue 4, pp. 557-568. DOI: 10.1016/j.trc.2010.12.003.

Pu, Z., Cui, Z., Zhu, M., & Wang, Y. (2019). *Mining Public Transit Ridership Flow and Origin-Destination Information from Wi-Fi and Bluetooth Sensing Data*.

Pu, Z., Zhu, M., Li, W., Cui, Z., Guo, X., & Wang, Y. (2021). Monitoring Public Transit Ridership Flow by Passively Sensing Wi-Fi and Bluetooth Mobile Devices. *IEEE Internet of Things Journal*, Vol. 8, Issue 1, pp. 474–486. https://doi.org/10.1109/JIOT.2020.3007373.

Rahman, S., Wong, J., & Brakewood, C. (2016). Use of Mobile Ticketing Data to Estimate an Origin–Destination Matrix for New York City Ferry Service. *Transportation Research Record*, vol 2544, Issue 1, pp. 1-9. DOI: 10.3141/2544-01.

Rakebrandt, A. (2007), "Transit tracking: automatic passenger counting systems and tracking ridership." *Mass Transit*. Vol 33, Issue 1, pp. 28–34.

Ramachandran, J. (2013). *U.S. Patent No. 8,442,807*. Washington, DC: U.S. Patent and Trademark Office.

Rusca, R., Sansoldo, F., Casetti, C., & Giaccone, P. (2023). What WiFi Probe Requests can tell you. *2023 IEEE 20th Consumer Communications & Networking Conference (CCNC)*, pp. 1086–1091. https://doi.org/10.1109/CCNC51644.2023.10060447.

Ryu, S., Park, B. B., & El-Tawab, S. (2020). WiFi sensing system for monitoring public transportation ridership: a case study. *KSCE Journal of Civil Engineering*, Vol 24, Issue 10, pp. 3092-3104.

Shibata, K., & Yamamoto, H. (2019). People crowd density estimation system using deep learning for radio wave sensing of cellular communication. In *2019 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)* (pp. 143-148). IEEE. DOI: 10.1109/ICAIIC.2019.8669071.

Sipetas, C., Keklikoglou, A., & Gonzales, E. J. (2020). Estimation of left behind subway passengers through archived data and video image processing. *Transportation Research Part C: Emerging Technologies*, Vol 118, p. 102727. DOI: 10.1016/j.trc.2020.102727.

Sørensen, A. Ø., Olsson, N. O., Akhtar, M. M., and Bull-Berg, H. (2019). Approaches, technologies and importance of analysis of the number of train travelers. *Urban, Planning and Transport Research,* Vol 7, Issue 1, pp. 1-18. DOI: 10.1080/21650020.2019.1566022.

Stirling, D. (2012). "Crowdsourced Information Is Key to Tiramisu Transit App." Accessed on May 27th, 2022. https://ischool.syr.edu/crowdsourced-information-is-key-to-tiramisu-transit-app/.

Sun, Y., & Schonfeld, P. M. (2016). Schedule-Based Rail Transit Path-Choice Estimation using Automatic Fare Collection Data. *Journal of Transportation Engineering*, Vol. 142, Issue 1, p. 04015037. https://doi.org/10.1061/(ASCE)TE.1943-5436.0000812.

Tang, C., Li, W., Vishwakarma, S., Chetty, K., Julier, S., and Woodbridge, K. (2020). Occupancy detection and people counting using WiFi passive radar. *2020 IEEE Radar Conference (RadarConf20)* , pp. 1-6. IEEE. DOI: 10.1109/RadarConf2043947.2020.9266493.

Tejas, N.A., Zain Y., and Phanimozhi K. (2015). "Cloud-Based Public Transport Information System for Android Devices Using Crowd Sourcing." *International Journal of Innovative Research in Computer and Communication Engineering,* Vol 3, pp. 4503-4509. DOI: 10.15680/ijircce.2015.0305073.

Texas A&M Transportation Institute (2022). "Crowdsourcing." Public Engagement and Communications. https://mobility.tamu.edu/mip/strategies.php, Accessed on May 27th, 2022.

Traunmueller, M. W., Johnson, N., Malik, A., & Kontokosta, C. E. (2018). Digital footprints: Using WiFi probe and locational data to analyze human mobility trajectories in cities. *Computers, Environment and Urban Systems*, Vol 72, pp. 4–12. https://doi.org/10.1016/j.compenvurbsys.2018.07.006.

United States Department of Energy the Office of Vehicle Technologies. Improving Mobility, Affordability, and Energy Efficiency Through Transit: Fiscal year 2020. https://www.energy.gov/sites/default/files/2020/07/f76/FY20_VTO_2197_selections_table-for_release.pdf.

Vanhoef, M., Matte, C., Cunche, M., Cardoso, L. S., & Piessens, F. (2016). Why MAC address randomization is not enough: An analysis of Wi-Fi network discovery mechanisms. In Proceedings of the 11th ACM on Asia conference on computer and communications security, pp. 413-424.

Vanhoef, M., & Piessens, F. (2016). Key reinstallation attacks: Forcing nonce reuse in WPA2. In Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security , pp. 1313-1328.

Vemula A., Patil N., Paharia V., et al (2015). "Improving Public Transportation Through Crowd-Sourcing." *International Conference on Communication Systems and Networks (COMSNETS)*. IEEE, pp. 1–6. DOI: 10.1109/COMSNETS.2015.7098724.

Videa, A., & Wang, Y. (2021). Inference of Transit Passenger Counts and Waiting Time Using Wi-Fi Signals Final Report. (Doctoral dissertation, Montana State University-Bozeman, College of Engineering).

Vieira, T., Almeida, P., Meireles, M., and Ribeiro, R. (2020). Public Transport Occupancy Estimation using WLAN Probing and Mathematical Modeling. *Transportation Research Procedia*, Vol 48, pp. 3299-3309. DOI: 10.1016/j.trpro.2020.08.122.

Wanek-Libman, M. (2020). Occupancy data: A real crowd-pleaser. https://www.masstransitmag.com/technology/passenger-info/article/21154973/occupancy-data-a-real-crowd-pleaser, Accessed on May 27th, 2022.

Wei, X., Zhang, Y., Wei, Y., Hu, Y., Tong, S., Huang, W., & Cao, J. (2021). Metro passenger-flow representation via dynamic mode decomposition and its application. *IEEE Transactions on Neural Networks and Learning Systems*. DOI: 10.1109/TNNLS.2021.3090695.

Weppner, J., Lukowicz, P., Blanke, U., & Tröster, G. (2014). Participatory Bluetooth Scans Serving as Urban Crowd Probes. *IEEE Sensors Journal*, Vol 14, Issue 12, pp. 4196-4206. DOI: 10.1109/JSEN.2014.2360123.

Yang, X., Xue, Q., Ding, M., Wu, J., & Gao, Z. (2021). Short-term prediction of passenger volume for urban rail systems: A deep learning approach based on smart-card data. *International Journal of Production Economics*, Vol 231, p. 107920.

Yu, H., He, Z., & Liu, J. (2007). A vision-based method to estimate passenger flow in bus. In *2007 International Symposium on Intelligent Signal Processing and Communication Systems* (pp. 654-657). IEEE. DOI: 10.1109/ISPACS.2007.4445972.

Zhao, J., Zhang, L., Ye, K., Ye, J., Zhang, J., Zhang, F., & Xu, C. (2022). GLTC: A Metro Passenger Identification Method Across AFC Data and Sparse Wi-Fi Data. *IEEE Transactions on Intelligent Transportation Systems*. DOI**:** 10.1109/TITS.2022.3171332.