

Inference of Transit Passenger Counts and Waiting Time Using Wi-Fi Signals

Final Report

by

Aldo Videa

Graduate Research Assistant

Yiyi Wang, Ph.D.

Assistant Professor

Western Transportation Institute

College of Engineering

Montana State University

A report prepared for the

Small Urban, Rural and Tribal Center on Mobility

and

United States Department of Transportation

August 16, 2021

DISCLAIMER

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated in the interest of information exchange. The report is funded, partially or entirely, by a grant from the U.S. Department of Transportation's University Transportation Centers Program. However, the U.S. Government assumes no liability for the contents or use thereof.

AVAILABILITY OF DATASET

The dataset is available at the following link: <https://doi.org/10.5061/dryad.tht76hf00>

TABLE OF CONTENTS

1. Introduction.....	1
1.1. Background	1
1.2. Project Goals	1
1.3. Report Organization	1
2. Literature reviews	3
2.1. Preliminary Literature Review	3
2.2. Project Literature Review.....	3
3. Datasets	18
3.1. Study Area.....	18
3.2. Hardware and Software.....	25
3.3. Surveys	27
3.4. Manual Counts	37
3.5. Smart Station Data	62
4. Methodologies.....	73
4.1. Ridership	73
4.2. Wait Time.....	82
4.3. OD Flow Characteristics	86
4.4. Estimated Time of Arrival.....	87
5. Results and Analysis	90
5.1. Ridership	90
5.2. OD Flow Characteristics	122
5.3. Wait Time.....	126
5.4. Travel Times	131
6. Conclusions.....	136
6.1. Ridership	136
6.2. OD Flow Characteristics	137
6.3. Wait Time.....	137
6.4. Travel Time	138
6.5. Limitations of Research	138
6.6. Future Research.....	140
7. Appendix A: Sample Survey Implemented in the Pilot Study (BlueLine)	142

8.	Appendix B: Origins and Destinations During Manual Counts	145
9.	Appendix C: Description of the Attributes Collected by the Smart Station	151
10.	Appendix D: Estimation of Number of Clusters for the K-means Algorithm	157
11.	References.....	160

LIST OF TABLES

Table 1: Obstacle severity on wireless signals (Harwood, 2011).....	9
Table 2: Characteristics of Internet of Things (ITU, 2013)	11
Table 3: Bluetooth versus AFC and Survey Methods (Kostakos, Camacho, and Mantero, 2013)	14
Table 4: Demographics of Bozeman, Gallatin County and United States (Taunya Fagan, 2019)	18
Table 5: Description of the surveys conducted with passengers	28
Table 6: Schedule of survey implementation and data collection	28
Table 7: OD matrix of Blueline from the survey data	30
Table 8: OD matrix of Greenline from the survey data	31
Table 9: OD Matrix of Orangeline from the survey data.....	32
Table 10: OD Matrix of Redline from the survey data	33
Table 11: OD matrix of Yellowline from the survey data	35
Table 12: Travel times recorded from the manual collection in the Blueline	50
Table 13: Stopped times recorded from the manual collection in the Blueline.....	51
Table 14: Travel times recorded from the manual collection in the Greenline	53
Table 15: Stopped times from the manual collection in the Greenline.....	54
Table 16: Travel times recorded from the manual collection in the Orangeline	55
Table 17: Stopped times recorded from the manual collection in the Orangeline.....	56
Table 18: Travel times recorded from the manual collection in the Redline.....	57
Table 19: Stopped times recorded from the manual collection of the Redline.....	58
Table 20: Travel times recorded from the manual collection in the Yellowline	60
Table 21: Stopped times recorded from the manual collection in the Yellowline.....	61
Table 22: Summary Statistics of the Wi-Fi data (no. of observations = 147,141)	68
Table 23: Summary Statistics of the GPS data (no. of observations = 145,620).....	70
Table 24: Rule-based Method	74
Table 25: Enhanced Rule-based method.....	75
Table 26: Variables and categories per variable used in wait time experiment.....	83
Table 27: Number of signals after implementation of Rule-based method for the Blueline	90
Table 28: Number of signals after implementation of Rule-based method for the Greenline.....	90
Table 29: Number of signals after implementation of Rule-based method for the Orangeline....	91
Table 30: Number of signals after implementation of Rule-based method for the Redline	91
Table 31: Number of signals after implementation of Rule-based method for the Yellowline....	91

Table 32: Coefficients of the Poisson regression for the rule-based method	94
Table 33: MSE and APE of the Blueline after implementation of rule-based method.....	95
Table 34: MSE and APE of the Greenline after implementation of rule-based method.	96
Table 35: MSE and APE of the Orangeline after implementation of rule-based method.	96
Table 36: MSE and APE of the Redline after implementation of rule-based method.....	96
Table 37: MSE and APE of the Yellowline before and after Poisson correction.....	97
Table 38: Coefficients of the Poisson regression for the enhanced rule-based method	102
Table 39: MSE and APE of the Blueline after implementation of enhanced rule-based method	104
Table 40: MSE and APE of the Greenline after implementation of enhanced rule-based method	104
Table 41: MSE and APE of the Orangeline after implementation of enhanced rule-based method	104
Table 42: MSE and APE of the Redline after implementation of enhanced rule-based method	105
Table 43: MSE and APE of the Yellowline after implementation of enhanced rule-based method	105
Table 44: Results of cluster analysis on the Blueline datasets.....	111
Table 45: Results of cluster analysis on the Greenline datasets	112
Table 46: Results of cluster analysis on the Orangeline datasets	112
Table 47: Results of cluster analysis on the Redline datasets.....	112
Table 48: Results of cluster analysis on the Yellowline datasets	113
Table 49: Coefficients of the Poisson regression for the K-means clustering method.....	114
Table 50: MSE and APE of the Blueline after implementation of machine learning method....	116
Table 51: MSE and APE of the Greenline after implementation of machine learning method .	117
Table 52: MSE and APE of the Orangeline after implementation of machine learning method	117
Table 53: MSE and APE of the Redline after implementation of machine learning method.....	117
Table 54: MSE and APE of the Yellowline after implementation of machine learning method	118
Table 55: Comparison of the accuracy of the ridership estimation methods.....	122
Table 56: Coefficients of the most complicated model (response variable: error).....	128
Table 57: Coefficients of the model with no interactions (response variable: error)	129
Table 58: Coefficients of the model without device as predictor (response variable: error).....	129
Table 59: Coefficients of the model with actual time as the predictor (response variable: error)	129
Table 60: Analysis of variance table (response: travel time difference)	134

Table 61: Number of clusters by the elbow, silhouette, and gap statistic methods	158
--	-----

LIST OF FIGURES

Figure 1: Structure of a MAC Address (Abedi, Bhaskar, & Chung, 2013).....	5
Figure 2: Detection of a phone's probe by a Raspberry Pi	7
Figure 3: Wi-Fi communication between two devices and a Transmission Medium (Cisco Press, 2017)	8
Figure 4: Overview of the internet of things (ITU, 2013)	10
Figure 5: Types of devices and their relationship with physical things (ITU, 2013)	11
Figure 6: Population growth in Bozeman	20
Figure 7: Streamline Blueline transit route	22
Figure 8: Streamline Greenline transit route	23
Figure 9: Streamline Orangeline transit route	23
Figure 10: Streamline Redline transit route	24
Figure 11: Streamline Yellowline transit route	24
Figure 12: Streamline daytime service route map (HRDC, 2019).....	25
Figure 13: Components of the Smart Station.....	26
Figure 14: Placement of Smart Station on the buses	27
Figure 15: Percent Distribution of Devices by Manufacturer.....	29
Figure 16: Feedback provided by passengers	37
Figure 17: Number of passengers observed during the pilot study	38
Figure 18: Origins and destinations of the Blueline from the pilot study	40
Figure 19: Origins and destinations of the Greenline from the pilot study	41
Figure 20: Origins and destinations of the Orangeline from the pilot study	42
Figure 21: Origins and destinations of the Redline from the pilot study	43
Figure 22: Origins and destinations of the Yellowline from the pilot study	44
Figure 23: Sum of passenger counts per line	45
Figure 24: Blueline passenger counts	46
Figure 25: Greenline passenger counts	46
Figure 26: Orangeline passenger counts	47
Figure 27: Redline passenger counts	47
Figure 28: Yellowline passenger counts	48
Figure 29: Interface of the remote-control program	63
Figure 30: Total number of unique Wi-Fi networks detected.....	65
Figure 31: Percentages of networks by type	66

Figure 32: Percentages of networks by channel	67
Figure 33: Percentages of networks by the rate of data transfer	67
Figure 34: Histograms of the numeric variables of the Wi-Fi data	69
Figure 35: Histogram of the detection time of the Wi-Fi data.....	70
Figure 36: Histograms of the coordinates of the GPS data.....	71
Figure 37: Histogram of the speed of the GPS data.....	72
Figure 38: Methodological approach of the research	73
Figure 39: K-means algorithm (Piech, 2013)	79
Figure 40: Algorithm to estimate the wait time	83
Figure 41: OD flows and matrix representation (Rodrigue et al., 2017)	87
Figure 42: Plots of rule-based estimated values vs. counts by bus lines	92
Figure 43: Plot of rule-based estimated values vs. counts for all lines.....	93
Figure 44: Plot of corrected ridership estimates vs. counts of the rule-based	95
Figure 45: Counted, estimated, and corrected passengers with the rule-based method of the Blueline.....	98
Figure 46: Counted, estimated, and corrected passengers with the rule-based method of the Greenline.....	98
Figure 47: Counted, estimated, and corrected passengers with the rule-based method of the Orangeline.....	99
Figure 48: Counted, estimated, and corrected passengers with the rule-based method of the Redline	99
Figure 49: Counted, estimated, and corrected passengers with the rule-based method of the Yellowline.....	100
Figure 50: Plots of enhanced rule-based estimated values vs. counts by bus lines	101
Figure 51: Plot of enhanced rule-based estimated values vs. counts for all lines.....	102
Figure 52: Plot of corrected ridership estimates vs. counts of the enhanced rule-based method.	103
Figure 53: Counted, estimated, and corrected passengers with the enhanced rule-based method of the Blueline.	106
Figure 54: Counted, estimated, and corrected passengers with the enhanced rule-based method of the Greenline.	106
Figure 55: Counted, estimated, and corrected passengers with the enhanced rule-based method of the Orangeline.	107
Figure 56: Counted, estimated, and corrected passengers with the enhanced rule-based method of the Redline.	107

Figure 57: Counted, estimated, and corrected passengers with the enhanced rule-based method of the Yellowline.....	108
Figure 58: Cluster plot of the Blueline Wi-Fi data on January 14, 2019.....	109
Figure 59: Elbow method for determination of number of clusters.....	110
Figure 60: Silhouette method for determination of number of clusters.....	111
Figure 61: Gap statistic method for determination of number of clusters	111
Figure 62: Plot of unsupervised machine learning estimated values vs. counts for all lines.....	114
Figure 63: Plot of corrected ridership estimates vs. counts of unsupervised machine learning .	116
Figure 64: Counted, estimated, and corrected passengers with the clustering method of the Blueline	119
Figure 65: Counted, estimated, and corrected passengers with the clustering method of the Greenline.....	119
Figure 66: Counted, estimated, and corrected passengers with the clustering method of the Orangeline.....	120
Figure 67: Counted, estimated, and corrected passengers with the clustering method of the Orangeline.....	120
Figure 68: Counted, estimated, and corrected passengers with the clustering method of the Yellowline.....	121
Figure 69: Estimated OD matrix of the Blueline	123
Figure 70: Estimated OD matrix of the Greenline.....	123
Figure 71: Estimated OD matrix of the Orangeline.....	124
Figure 72: Estimated OD matrix of the Redline	124
Figure 73: Estimated OD matrix of the Yellowline.....	125
Figure 74: Smart station detection time versus actual time devices were detectable	127
Figure 75: Boxplots of Smart Station time error by distance	127
Figure 76: Boxplots of Smart Station time error by device type	128
Figure 77: Histograms of the observed and estimated wait times	130
Figure 78: Mean observed and estimated wait time (seconds)	131
Figure 79: Dispersion and probability histogram of the travel time differences (n=1,292)	132
Figure 80: Histogram of the travel time differences (n=1,292)	133
Figure 81: TukeyHSD plot for all groups	134
Figure 82: Plot of Cooks' distance standardized residuals vs leverage	135
Figure 83: Origins and destinations of the Blueline from the manual counts	146
Figure 84: Origins and destinations of the Greenline from the manual counts	147

Figure 85: Origins and destinations of the Orangeline from the manual counts	148
Figure 86: Origins and destinations of the Redline from the manual counts.....	149
Figure 87: Origins and destinations of the Yellowline from the manual counts	150

ABBREVIATIONS

<u>Abbreviation</u>	<u>Definition</u>
AFC.....	Automated Fare Collection
ANN.....	Artificial Neural Network
ANOVA	Analysis of Variance
APC.....	Automatic Passenger Counter
APE	Absolute Percentage Error
CDR	Call Detail Records
CI.....	Confidence Interval
CLT	Central Limit Theorem
dBm.....	Decibels-Milliwatts
FHWA.....	Federal Highway Administration
GHz	Gigahertz
GPS	Global Positioning System
GSM.....	Global System for Mobile
HRDC	Human Resource Development Councils
IEEE.....	Institute of Electrical and Electronic Engineers
IoT.....	Internet of Things
ITU.....	International Telecommunication Union
MAC	Media Access Control
Mbps	Megabits per Second
MHz	Megahertz
MIT	Massachusetts Institute of Technology
MLR.....	Multiple Linear Regression
mph	Miles per Hour
MSE	Mean Squared Error
MSU.....	Montana State University
mW.....	Milliwatt
NMEA.....	National Marine Electronics Association
OCR	Optical Character Recognition
OD.....	Origin-Destination
PCA.....	Principal Component Analysis

RSSI	Received Signal Strength Indicator
SLR	Simple Linear Regression
SS	Smart Station
SSE.....	Sum of Squared Errors
SSID.....	Service Set Identifier
UTC.....	Coordinated Universal Time
UTM.....	Universal Transverse Mercator
Wi-Fi	Wireless Fidelity
XLAN	Wireless Local Area Network
XML.....	Extensible Markup Language

EXECUTIVE SUMMARY

Passenger data such as real-time origin-destination (OD) flows and waiting times are central to planning public transportation services and improving visitor experience. This project explored the use of Internet of Things (IoT) Technology to infer transit ridership and waiting time at bus stops. Specifically, this study explored the use of Raspberry Pi computers, which are small and inexpensive sets of hardware, to scan the Wi-Fi networks of passengers' smartphones. The process was used to infer passenger counts and obtain information on passenger trajectories based on Global Positioning System (GPS) data. The research was conducted as a case study of the Streamline Bus System in Bozeman, Montana. To evaluate the reliability of the data collected with the Raspberry Pi computers, the study conducted technology-based estimation of ridership, OD flows, wait time, and travel time for a comparison with ground truth data (passenger surveys, manual data counts, and bus travel times).

This study introduced the use of a wireless Wi-Fi scanning device for transit data collection, called a Smart Station. It combines an innovative set of hardware and software to create a non-intrusive and passive data collection mechanism. Through the field testing and comparison evaluation with ground truth data, the Smart Station produced accurate estimates of ridership, origin-destination characteristics, wait times, and travel times.

Ridership data has traditionally been collected through a combination of manual surveys and Automatic Passenger Counter (APC) systems, which can be time-consuming and expensive, with limited capabilities to produce real-time data. The Smart Station shows promise as an accurate and cost-effective alternative. The advantages of using Smart Station over traditional data collection methods include the following: (1) Wireless, automated data collection and retrieval, (2) Real-time observation of passenger behavior, (3) Negligible maintenance after programming and installing the hardware, (4) Low costs of hardware, software, and installation, and (5) Simple and short programming and installation time. If further validated through additional research and development, the device could help transit systems facilitate data collection for route optimization, trip planning tools, and traveler information systems.

1. INTRODUCTION

1.1. Background

Transportation agencies use passenger information, such as real-time origin-destination (O-D) flows and waiting times, to plan public transportation services (e.g. identify new service needs) and improve user experience (e.g. enhance trip planning tools with real-time traveler information). However, this type of information is often unavailable because it requires continuous monitoring of transit networks (e.g. where and when passengers board and disembark). Some transit buses utilize automatic passenger counters (APCs), which count every time a passenger boards and deboards the bus, but these devices cannot provide individual trajectories because they can't determine *when and where* passengers board or de-board. Statistical methods like the Iterative Proportional Fitting (IPF) is often used to estimate passenger O-D flows (McCord & Mishalani, 2016), but they do not represent actual passenger behavior. Furthermore McCord et al. noted that APC and other passenger count methods are prone to overestimation (5% more than ground truth values).

With a high market penetration of smartphones, an opportunity for transportation research emerged. Researchers have used Bluetooth technology, for example, to infer passengers' transit information from their smartphones, but this technology is very limited due to the small number of people who have Bluetooth activated on their phones (El-Tawab, Oram, Garcia, Johns, & Park, 2017). It is possible to collect Wi-Fi signals emitted from personal devices and correlate them with the number of passengers (using certain adjustment factors that reflect ownership rate of traceable devices). This method has been used in Virginia (El-Tawab, Oram, Garcia, Johns, & Park, 2017), Washington (Langston, 2016), and Ohio (McCord & Mishalani, 2016) to infer the number of passengers, along with their real-time locations.

1.2. Project Goals

The Small Urban, Rural, Tribal Center on Mobility (a University Transportation Center authorized by the US Department of Transportation) initiated this project to explore new tools for collecting passenger information. The goal of the project was to apply existing Internet of Things (IoT) technology to infer transit ridership and waiting times at bus stops.

The research was conducted in collaboration with the Streamline bus service in Bozeman, Montana (population ~48,000). The study focused on weekday service, which consists of six lines that serve the city. In addition to advancing understanding of IoT technologies, the results of the research may be useful for Streamline to improve the efficiency of track fleet operation, identify new service needs, and evaluate quality of service for current passengers. Furthermore, Streamline could add the real-time traveler information to its current mobile app to help riders plan trips, reduce their waiting time, and improve overall user experience.

1.3. Report Organization

This final report synthesizes documents produced during the life of the project, including the literature review conducted for the proposal and project summaries developed for the sponsor agency at USDOT. A large portion of the new research formed the basis for the graduate research and thesis of author Aldo Alejandro Videa Martinez. It is a primary source of the content of this final report (Videa Martinez, 2019).

Chapter 2 summarizes the preliminary literature review conducted prior to the project, as well as the in-depth review conducted as part of the research effort. Chapter 3 describes the datasets collected, and Chapter 4 discusses the research methodologies applied. Chapter 5 describes the results and analysis, and Chapter 6 discusses conclusions and recommendations for future research.

2. LITERATURE REVIEWS

2.1. Preliminary Literature Review

The research team conducted a preliminary literature review to establish the need for the research and develop the proposed scope of the project.

In 2016, a team of researchers from the University of Virginia and James Madison University conducted a study to estimate passenger waiting time at bus stops using computers programmed to detect Media Access Control (MAC) addresses, which are unique identification numbers assigned to personal devices (like cellphones and tablets) as they access a Wi-Fi or cellular network. A few experiments were conducted to test the accuracy of the method under various conditions. These test conditions included a single bus station; two adjacent bus stations to test whether the algorithm can match MAC addresses to the correct bus stop; an evacuation drill that produced a large amount of confounding signals from evacuees standing by the bus stops; and long observation time (6.5 hours) for the two bus stations to test durability of the batteries. Subsequently, the study collected three variables: arrival time of each detected device, signal strength of the devices (measured by RSSI, which is an indicator of the power level being received), and MAC addresses. The results showed that this Internet of Things (IoT) technology can detect traceable Wi-Fi signals from devices and calculate waiting time at bus stops with a high level of accuracy (El-Tawab, Oram, Garcia, Johns, & Park, 2017).

A team from the University of California, Berkeley and The Ohio State University developed a method to reconstruct and track transit information at the resolution of passenger trajectories. The research targeted a metro line in San Francisco. Team members matched the locations of the travelers that carried traceable smartphones to bus locations through the automatic vehicle location (AVL) technology imbedded in their bus fleets, with the goal of developing a personalized scheduler for transit riders. This methodology required each passenger to download a survey app and allow access to their location. The challenge with this methodology is that passengers must download the app and consent to provide access to their location. The team was successful in gathering passenger locations during trips and matching them with transit locations (Carrell, Lau, Mishalani, Sengupta, & Walker, 2015).

2.2. Project Literature Review

Prior to conducting the field study, the research team prepared a thorough synthesis of research works that have utilized wireless technologies to estimate parameters from passengers. Additionally, the review includes relevant background information on Wi-Fi networks.

This literature review was prepared for the dual purposes of this research report and the master's thesis of one of the principal investigators/authors (Videa Martinez, 2019).

2.2.1. Wireless Communication Technologies

For a better understanding of how Wi-Fi signals can be used for ridership estimation, some technical considerations will be explained. Furthermore, it is important to know how and under which situations passengers can be detected and counted. This section provides a detailed overview of the technical aspects.

2.2.1.1. Wi-Fi

Wi-Fi wireless is a communication standard that was designed to establish connections for a wireless local area network (WLAN). Wi-Fi is based on Institute of Electrical and Electronics Engineers (IEEE) 802.11 standard (Han, et al., 2012). Almost all smartphones, notebook computers, TVs and other pocket electronic devices have the option to establish a Wi-Fi connection.

Wi-Fi devices use two simple operation modes to connect with each other. The infrastructure mode utilizes an existing network infrastructure, which could be a router (or beacon) to connect several Wi-Fi devices with each other. The other mode is an ad hoc connection between devices. Devices in these modes will send beacon messages to expose the presence of an established network. The beacon message sends information related to the network, like the service set identifier (SSID) and the capability of the network (Böhm, 2016).

The Wi-Fi interfaces for mobile phones work in the ad hoc mode to seek a possible network and send beacon messages while they operate in this mode. Wi-Fi devices scan wireless channels of 2.4 and 5 GHz to find other devices that are sending beacon messages. There are two modes of scanning: active and passive (Han, et al., 2012). The passive scanning mode refers to the devices that are listening to beacon messages and switching between the two wireless channels. In the active scanning mode, the devices actively seek other devices, send probe and probe request messages, and wait for other devices to respond.

In most cases, sensors use the active mode to detect other devices in the nearby area (Abbot-Jard, Shah, & Bhaskar, 2013). However, the Raspberry Pi works in the passive mode because it does not respond to beacon messages. The typical time that a Wi-Fi enabled device is discovered by a sensor is 1 second (Duflot, Kwiatkowska, Norman, & Parker, 2006), which is less than the time a Bluetooth sensor takes to discover a new device. Both Bluetooth and Wi-Fi operate in the same radio frequency band; the reason for their difference is the architectural design of the hardware.

2.2.1.2. MAC Address

Media Access Control (MAC) addresses are unrepeatable identifiers that are used for the majority of IEEE 802 network technologies, such as Bluetooth and Wi-Fi (Abedi, Bhaskar, & Chung, 2013). Figure 1 shows the architecture of a MAC address which constitutes six bytes. Most smartphones possess both Bluetooth and Wi-Fi capabilities. The MAC address is a unique identifier; therefore, it can be tracked, and this specific feature has been the motivation of various studies. For example, capturing Bluetooth devices on a highway and in urban areas has been used to estimate vehicles' travel time.

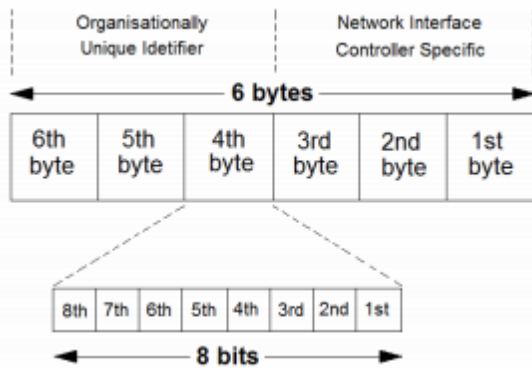


Figure 1: Structure of a MAC Address (Abedi, Bhaskar, & Chung, 2013)

There are some factors that might affect the quality of MAC address data collection. These factors can be associated with the software and the hardware implemented. The antenna type is one of these contributing factors. Some characteristics of the antennas like gain and polarization must correspond to specific applications (Porter, Kim, Magaña, Poocharoen, & Arriaga, 2013). For Wi-Fi scanning, the antennas must be able to be set in monitor mode to scan the Wi-Fi signals with the implemented software. Some devices are designed to only work as Bluetooth scanners, Wi-Fi only or both. The MAC address scanners are set to work with a synchronized clock.

2.2.1.3.

Comparison of Bluetooth and Wi-Fi Technologies

There are some factors that vary between Bluetooth and Wi-Fi technologies. These factors make it more convenient for some researchers to work with, depending on the objectives of their studies. Below, some of the most important differences are presented.

Architecture. Both Wi-Fi and Bluetooth technologies manage the traffic by a central unit called Master in Bluetooth and AP in Wi-Fi. These central units are responsible for directing packets between the devices. Bluetooth can possess a maximum of 7 slave units, while Wi-Fi has a maximum of 2,007. The nominal range of Bluetooth is 10 meters for mobile devices. On the other hand, Wi-Fi technologies have a nominal range of 35 meters indoors and up to 100 meters outdoors (Ferro & Potorti, 2004).

Discovery Time. Collecting accurate discovery time is a key component in collecting efficient data of short duration. Bluetooth discovery time is approximately 10.21 seconds, while the Wi-Fi discovery time is around 1 second. This means that the discovery time for MAC addresses using Bluetooth technology is around 10 times higher than with the Wi-Fi technology.

Usage Popularity. In some experiments, results have indicated that Wi-Fi sensors can detect more than 90% of all MAC addresses present in different environments. Nevertheless, Bluetooth consistently scanned less than 10% of all the MAC addresses that were present in all the environments. This is an indication that Wi-Fi technology is more successful at detecting MAC address in many places like offices, malls, and universities where the experiments were done (Abedi, Bhaskar, & Chung, 2013).

Signal Strength. Bluetooth and Wi-Fi signals can be measured in decibelsmilliwatts (dBm). The environment and the antenna type have an impact on the signal strength detected by the Bluetooth and Wi-Fi technologies ((Abedi, Bhaskar, & Chung, 2013); (Dimitrova, Alyafawi, & Braun, 2012)).

2.2.1.4. Principles of Wi-Fi Tracking

Wi-Fi tracking is sustained by the circumstances that smartphones and tablets are currently present everywhere. Even in developing countries, consumption is high as people are consumers and not producers (Internet Society Organization, 2017). Tracking Wi-Fi enabled devices is based on the idea that a Wi-Fi access point (referred to as a Wi-Fi scanner) can discover a network (Curran et al., 2011).

A network connection can be established in two ways, a device passively waiting for other signals or a device actively seeking an access point. For mobile devices, it is usually more effective to actively seek Wi-Fi scanners to establish a connection. Under this paradigm, mobile devices send probe requests periodically even if they have not established a connection.

Probe requests contain a large amount of information including the MAC address of the sending device. The MAC address, as it is a unique code, can be used as an identifier for a device.

Wi-Fi tracking works by the following process. A Wi-Fi scanner receives a probe request from a mobile device with an identifier (MAC address) and a specific time. Ordering the information in chronological order provides a series of detected devices. If the location of the scanner is known, there can be an estimation of the location of the mobile device, over a specific time frame. This is the basic idea of Wi-Fi tracking. However, in real world applications, many factors can influence the accuracy of the detections as will be explained in this chapter.

Wi-Fi Scanners Technical Background. A Wi-Fi scanner can be as simple as a home router with a modified software that would enable it to record nearby Wi-Fi frames. This same outcome can be achieved by a laptop; nevertheless, routers have a lower price.

Almost any device that has a Wi-Fi interface can function as a Wi-Fi scanner. The only requirement is that the device can be set up in monitor mode as defined in the IEEE 802.11 standard (IEEE Standards Association, 2012). The monitor mode allows a device to capture Wi-Fi frames from other devices without the need for an association with the devices. Wi-Fi scanning is performed by routers or devices that have been specifically programmed for that task.

A Raspberry Pi computer can be used to scan Wi-Fi networks. This device has been used to detect wait time of passengers of a transportation system based on the time stamps of MAC addresses (El-Tawab et al., 2017). The process by which a Wi-Fi network is detected by a Raspberry Pi is shown in Figure 2.



Figure 2: Detection of a phone's probe by a Raspberry Pi

2.2.1.5. Common Issues of Wi-Fi Tracking

A perfect scenario would be one in which a mobile device's location is accurately known in addition to its specific time stamps over a time interval. This signifies that there are no moments when the location of the device is not being registered. Unfortunately, collecting perfect datasets through Wi-Fi tracking is a difficult task. There are various sources of errors. Some of these sources will be discussed here:

Scanner Malfunctioning. Some errors are generated by the scanner device, and usually these are easier to identify and fix. For instance, the scanner can shut down and fail to detect networks, which will generate an inconsistency with the density of the detections over time.

Limitations of Radio-based Detections. Wi-Fi utilizes the air as its data transmission medium, which is shown in Figure 3. This medium is unreliable (Salyers, Striegel, & Poellabauer, 2008). One example of such unreliability is that most Wi-Fi devices claim to have a 100-meter transmission range in favorable conditions. Nevertheless, these specifications cannot be trusted due to sources of impairment in the transmission of wireless communications, like attenuation distortion, loss of free space, the noise of other signals, atmospheric absorption, multiple paths that the information can go to, and refraction (Beard & Stallinds, 2016).

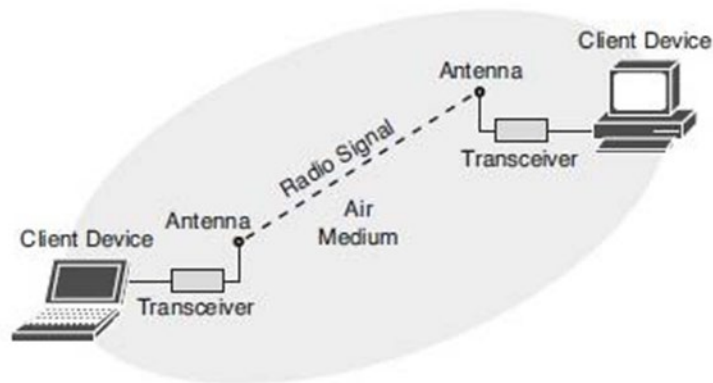


Figure 3: Wi-Fi communication between two devices and a Transmission Medium (Cisco Press, 2017)

As a result of these limitations, it has been shown that the transmission range is increased in tunnels. In the same way, buildings and people are known to hamper the transmissions. This means that the shape and area in which Wi-Fi signals can be received are irregular and circumstantial depending on the location and surrounding objects. For this reason, some phones are not detected immediately when they board a bus.

Limitations of RSSI. The received signal strength indicator (RSSI) trilateration method is theoretically able to provide the location coordinates of a device (H. Liu, Darabi, Banerjee, & Liu, 2007). However, there are some problems to be discussed. First, RSSI measurements, when taken by scanners, are not standardized and their values can fluctuate between different types of scanners. Second, the signal strength, can likewise dramatically differ between different manufacturers and even for different devices of the same model. Solutions for this have been introduced (Y. Kim, Shin, & Cha, 2012), but only when the mobile device is being used as a scanner. Researchers have introduced an experimental evaluation to illustrate these difficulties of RSSI (Zanca, Zorzi, Zanella, & Zorzi, 2008).

Timing Errors. When the Wi-Fi scanning devices are not properly synchronized, they introduce a massive error in the detection times. This is also true for all other types of scanners (Petre, Chilipirea, & Baratchi, 2016). Even when the scanners are synchronized, there are difficulties because some probe requests are not detected immediately when they enter the scanners' region of detection.

MAC Address Issues. MAC addresses used to be detected and utilized as reliable unique identifiers for a device. However, this is not true anymore because some devices change their MAC addresses randomly. Such randomizations occur at different time intervals (Musa & Eriksson, 2012). In fact, the randomization has millions of random possibilities that make it untraceable to the original device (Martin et al., 2017).

In summary, there are several limitations that make Wi-Fi tracking a challenging task. This research requires the correct inference about the passengers who use a transit system. Estimation of the number of riders is difficult due to the randomization of MAC addresses by some devices, people carrying more than one Wi-Fi enabled device, or people who do not carry devices at all. In addition, the devices' rates of data transfer have enormous variations. Seemingly, this phenomenon is caused by random behavior. This behavior varies by device type (Cunche, 2014). The combined effect of these limitations, the large number of noise sources and the unreliability of the

transmission medium can make the movement of a device look erratic. To end this section, the effects of environmental objects on wireless communication are shown in Table 1.

Table 1: Obstacle severity on wireless signals (Harwood, 2011)

Obstruction	Obstacle Severity	Example Use
Wood	Low	Inside a wall or hollow door
Drywall	Low	Inside walls
Furniture	Low	Couches or office partitions
Clear glass	Low	Windows
Tinted glass	Medium	Windows
People	Medium	High-volume traffic areas that have considerable pedestrian traffic
Ceramic tile	Medium	Walls
Concrete blocks	Medium/high	Outer wall construction
Mirrors	High	Mirror or reflective glass
Metals	High	Metal office partitions, doors, metal office furniture
Water	High	Aquariums, rain, fountains

2.2.2. Internet of Things Technologies

Internet of Things (IoT) represents the interconnection of the sensors that can utilize various technologies for connection (e.g., RFID, Bluetooth, Wi-Fi, LTE, 3G) (Ezechina, Okwara, & Ugboaja, 2015). IoT-enabled devices share information about their conditions with the surrounding environment where other machines with a software system can retrieve the information and make programmed decisions. Subsequently, the information could be sent anywhere in the world through the Internet. The Internet has impacted education, communication, business, science, government, and all human activities. While there are hundreds of specific uses of the Internet of Things, industry groups these uses into two main categories:

- Category one: this encompasses the vision of millions of heterogeneous and interconnected devices with unique IDs that are interacting. In this category, devices are “aware” of the information they send and the devices to which they intend to send such data.
- Category two: this incorporates the analysis of the data collected by smart devices with sensing and connectivity capability. This category deals with data mining, it is generally used in the commercial and marketing industry.

IoT technologies have become a popular topic of research, partly because smartphone usage and speed of data connection are increasing. Although the monetary cost of data connection has decreased, the technology has not been applied in its full potential for transportation applications.

Meanwhile, roadway traffic density continues to be a principal problem today. Policy makers acknowledge that a good transportation system positively affects the health of the people (Lee & Sener, 2016). Therefore, efficient transportation systems are a priority for city planners. However, most practitioners have traditionally focused on the vehicle-centric approach rather than focusing on new technologies that would benefit the transportation system.

Emerging technologies that have permeated society can be used as a frame of data collection to obtain high-resolution real-time passenger parameters indirectly. IoT technology has become a tremendous tool that has changed the way humans interact. IoT has allowed communication between people and objects, machines and equipment (Aldein Mohammed & Ali Ahmed, 2017).

This research applies the category two approach. The data was collected through portable computers that were programmed to retrieve Wi-Fi information. Later, the data was processed to obtain information on passengers and vehicles' characteristics. Statistical analyses were performed to test if the results are similar to the ground truth observations that were collected from passenger counts and travel times.

The key component to achieving IoT communication is the ability to establish a connection between different devices, so they are able to communicate. This property is indispensable when labeling a device as an IoT device. The way the communication is performed is not essential.

The following properties are also important for performing the communication process: sensing, maneuvering, capturing, storing and processing data (Bude & Kervefors, 2015). There are different ways devices communicate with other devices: devices communicate through the communication network with a gateway (case a), without a gateway (case b) or without any intermediary (case c).

Additionally, it is possible to find combinations of cases a and c, and b and c as well; in other words, devices can share information with other devices using direct communication through a local network. Figure 4 shows these interactions.

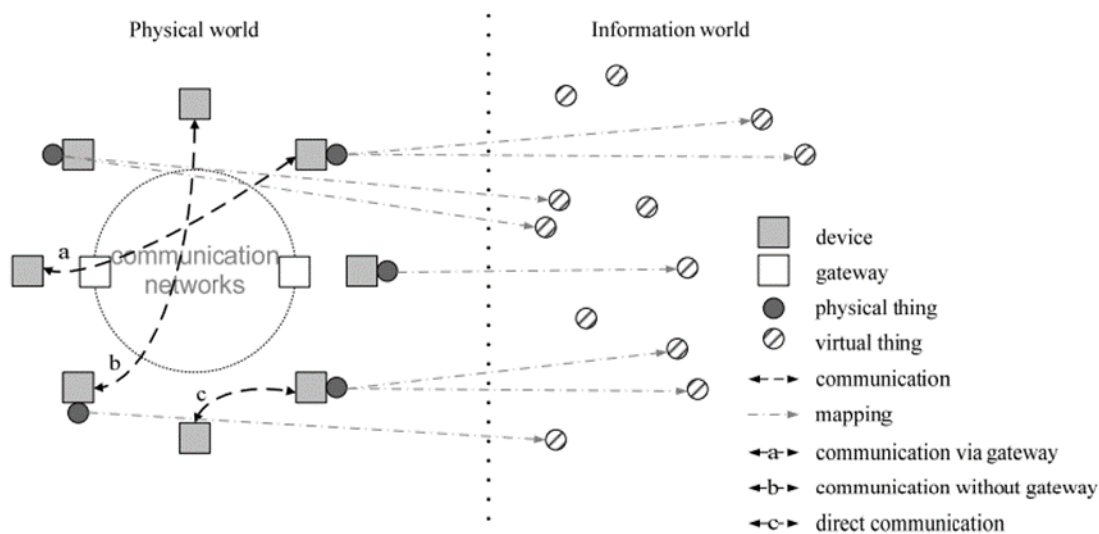


Figure 4: Overview of the internet of things (ITU, 2013)

The sets of communication networks contain data that is gathered by devices and redirected to applications and other devices. The networks offer capabilities for unfailing and efficient data transfer. The IoT can be accomplished via existing networks or packetbased networks (International Telecommunication Union, 2013). Figure 5 shows the different kinds of devices and their relationship between physical things.

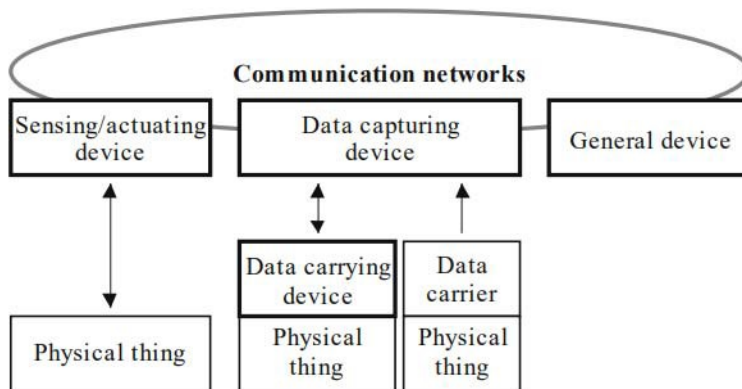


Figure 5: Types of devices and their relationship with physical things (ITU, 2013)

The differences between an IoT device and other non-IoT devices are shown in Table 2. These characteristics provide a good description of the differences between IoT devices.

Table 2: Characteristics of Internet of Things (ITU, 2013)

Characteristics	Description
Interconnectivity	Everything can be connected to the global information and communication infrastructure.
Things-related services	Provides things-related services within the constraints of things such as privacy and semantic consistency between physical and virtual things.
Heterogeneity	Devices within IoT have different hardware and use different networks but they can still interact with other devices through different networks.
Dynamic Changes	The state of a device can change dynamically, thus the number of devices can vary. (Device states: connected, disconnected, waking up, and sleeping)
Enormous scale	The number of devices operating and communicating will be larger than the number of devices in the current Internet. Most of this communication will be device to device instead of human to device.

2.2.3. Methods of Origin-Destination Data Collection

Transit agencies have a significant interest in collecting data on passengers and vehicles, as is shown by their continuous investments in Automated Passenger Counter (APC) systems. There are several ways to collect data to estimate transit ridership and OD information. Numerous studies have been conducted that varied in data collection methods, as well as data analysis. The following studies have been summarized because they possess similar objectives and provided valuable insight useful for this study.

2.2.3.1. Automated Passenger Counter Data

Automated Passenger Counter (APC) technology is adopted mainly for bus services. APC systems offer passenger boarding and de-boarding counts. This technology is relatively inexpensive. The system counts riders entering and exiting the bus, which removes the need for drivers or technicians who are employed usually by transit agencies to count the passengers on board. This reduces the workload of drivers and allows for greater comfort of their drive, which translates into a smoother ride for the passengers as well (Bongiorno, Bosurgi, Pellegrino, & Sollazzo, 2017).

The APC system eliminates the error from manual passenger counting. The different methods of APC devices include treadle mats, infrared beams, passive thermal, digital cameras, ultrasound, and light beams. The treadle mats count the passengers when they traverse the bus steps. Infrared beams, passive thermal, and ultrasound systems detect the presence of passengers as they pass between two columns and count them. The digital cameras detect passenger movement; however, some errors exist. Other objects and poor lighting can impact data analysis (Lefloch, Cheikh, Hardeberg, Gouton, & Picot-Clemente, 2008).

The passengers' boarding and de-boarding counts cannot be linked. Yet, information regarding OD flows can be inferred. Boarding and de-boarding counts provide indirect information that can be used to estimate the OD flow patterns. The error that is associated with APC systems is random and systematic. The main sources of APC systems errors are mechanical problems, environmental factors, passenger behavior, and data processing (Mishalani, Ji, & McCord, 2011).

2.2.3.2. Automated Fare Collection System

Automated Fare Collection (AFC) systems provide a reliable and inexpensive alternative to the common manual fare collection procedures. AFC data can be used to create OD matrices to assess the performance of the system and contributes to the planning and management of the system. One of the disadvantages of this approach is that the passenger boarding location is recorded at the bus-route level. This makes it difficult to acquire data on the specific bus stops in which the passengers boarded (Hora, Dias, Camanho, & Sobral, 2017).

Another disadvantage of this system is its inability to track passengers' movements through the transit system like arrival time at a bus station or transferring to another bus route. Additional disadvantages include high energy consumption, the high cost of the initial investment, and privacy concerns (Heydt-Benjamin, Chae, Defend, & Fu, 2006). Nevertheless, this method is known for eliminating the response bias, the low response rates from specific demographics (Barry, Newhouser, Rahbee, & Sayeda, 2002).

2.2.3.3. Survey Data

Some transit agencies send their personnel to conduct passenger counts or onboard surveys in order to estimate OD data. Beyond collecting ridership counts and passengers' movements, the onboard survey methods provide opportunities for the agencies to interact with passengers and hear their needs through a face-to-face experience. These methods, however, are time-consuming and labor-intensive. Other concerns are privacy and response bias. More recently, onboard surveys are combined with interviews using web-enabled devices and online surveys (Chow, 2014).

2.2.3.4. Analysis by Population

The Center for Urban Transportation Research in Tampa, Florida has implemented different methods in an attempt to improve public transit systems (Perk & Kamp, 2003). One method estimates the number of passengers by the service area according to the population per square mile. This method typically over or underestimates passenger statistics. Another method to collect passenger data estimated passenger miles by sampling different collections of transit passengers but requires a great deal of time and money. Neither method provided the information necessary to accomplish a massive improvement to the public transit systems.

2.2.3.5. Analysis by Mobile Signaling

The Massachusetts Institute of Technology (MIT) conducted a study using mobile signals detected by a service provider. Researchers not only looked at the number of mobile signals from smartphones, but they also used a GPS-based method to measure the speed and direction of any one individual from the smartphone and used Call Detail Records (CDR) to look at a computer record of a telephone exchange. This data collection method is only valid for people using Global System for Mobile (GSM). That is why MIT also utilized GPS and CDRs to have a more complete collection of data. The study used CDRs to organize trip information into a user ID, the origin of travel, the destination of travel and start and end times. Since the research extended to private vehicles, public transit, and pedestrian transportation, each smartphone was placed into its respective category based on travel times over distances of travel (H. Wang, Calabrese, Di Lorenzo, & Ratti, 2010). The smartphones were detected by the GSM Communications (Rouse, 2007).

The study used speed to infer travel modes via a k-means unsupervised clustering algorithm. Each smartphone was considered as one anonymous data point. The Wi-Fi scanner recorded the locations (latitude and longitude), with three locations for every second of each smartphone (i.e., an anonymous data point). K-means unsupervised clustering algorithm was used to group consecutive general location measurements to define the origins or destinations for each group. To avoid overlaps between any two groups, the researchers split any overlapping groups into subgroups based on the initial distance and then computed the average travel time of each subgroup. Then the error of transportation mode (walking, public transit or driving) was measured as the average of the differences between k-means based average travel time and GPS-based average travel time. The travel movement led to the conclusion that public transit travel times were a function GPS data schedule which was provided by Google Maps. The combination of mobile signaling, GPS and CDRs provided a much higher rate of accuracy than any one method would have provided on its own. Although this research included a variety of transportation modes, its dependence on CDRs to obtain the data results in a coarse-grained dataset and lower detail of travel information.

2.2.3.6. Bluetooth Data

Studies have inferred OD matrices from Bluetooth data (Michau et al., 2015). One source of error is that some passengers may carry more than one device. Nevertheless, it is often assumed that each passenger carries only one device (Dunlap, Li, Henrickson, & Wang, 2016). In addition, errors could happen when the devices fail to function correctly for unidentified reasons or low battery power. In these cases, no data is collected (Purser, 2016).

Previously, mobile phone tracking has been utilized to measure passenger flows between different cities. Nevertheless, the results have yielded low-quality spatial data. Therefore, these methods have been more suited for long distance trips. However, this technology is very useful to capture individual trips. Table 3 shows a comparison of the Bluetooth methods and other popular OD estimation data collection methods.

2.2.3.7. Analysis by Wi-Fi Signaling

Wi-Fi signaling uses the collection of longitudinal data about human mobility through the wireless data shared by mobile devices and is actually a common practice (Lazer et al., 2009). By utilizing a large set of sensors, it can detect Wi-Fi access points and GPS to generate a high-resolution map of ridership patterns (Ferris, Hähnel, & Fox, 2006; Lim, Wan, Ng, & See, 2007). The combination of Wi-Fi's access points and GPS information improves the accuracy of the positioning data.

Table 3: Bluetooth versus AFC and Survey Methods (Kostakos, Camacho, and Mantero, 2013)

	Method of OD estimation		
	Bluetooth detection	AFC	Survey
Sample Size	~10%	>50%	~3%
Spatial accuracy of destination data	High	Relies on inferencing (which introduces bias)	High (explicitly stated by respondent)
Representativeness and sample bias	Demographic bias on technology adoption	Bias if all passengers do not swipe ticket	Bias due to sampling technique, human memory, and self-selection of respondents
Passenger effort	Enable Bluetooth	Swipe ticket	Answer questionnaire

The method of scanning all network traffic in an area is known as wardriving. This technique, however, is not very well known and only a few studies are found in the literature (Letchner, Fox, & LaMarca, 2005; Rekimoto, Miyaki, & Ishizawa, 2007). Indisputably, Wi-Fi networks, which were first intended for communication, can also serve as an infrastructure to track the location of the users. These networks are present everywhere and mobile devices are constantly sending probes. Generally, almost all phone applications require sharing information such as location to function. This information requires the consent of the user to install a specific application. These

applications are demonstrated to be accurate in the detection of users' mobility. However, information detection through wardriving is also a possibility.

The advantages of wardriving are that it is a low-cost technique and it does not require a vast technical knowledge in computer hardware. The initial purpose of wardriving was to detect access points associated with a specific location. Nevertheless, through passive detection, a device can listen to all networks, even the cloaked ones, which include devices' probes. These devices can be associated with individuals through their unique MAC addresses (Etter, 2002).

Researchers have attempted to use the smartphone Wi-Fi signaling data collection method to improve safety and mobility in the United States through innovative research and data collection (El-Tawab et al., 2017). This research proposed a method for scanning MAC addresses using Raspberry Pi computers and for estimating the wait time of passengers at a specific bus station. The researchers performed four experiments to test the functioning of their equipment. They tested their devices and were able to successfully scan MAC addresses, which was the main scope of the project. To estimate a passenger's waiting time, they subtracted the first time of detection from the last time of detection.

2.2.4. Methods of Travel Time Estimation

Travel time data had been very hard to obtain until recently. Travel time is of primary importance in user information systems because it is natural for drivers to understand this value. The total investment of Dynamic Message Signs hardware alone in the U.S. surpassed 330 million dollars in 2005. The Federal Highway Administration (FHWA) recommends providing estimated travel times to popular destinations on major highways (Meehan, 2005). The quality of this system depends a lot on the precision of the travel times estimations. Imprecise travel time estimates can have a negative effect on the system because passengers will not rely upon the information that is provided by the transportation agencies, resulting in vague data for the system management. Consequently, it is important to understand the accuracy of the travel times estimates. The FHWA recommends a maximum error of ± 20 percent, with the ideal maximum value of ± 10 percent error (Wang et al., 2011). This section describes the current travel time data collection methods that exist in the literature.

2.2.4.1. Probe Vehicle-based Travel Time Analysis

Probe vehicle-based analysis depends on a driver that will utilize a car and drive on the highway at the pace of other drivers. Usually, GPS-equipped vehicles are used to provide coordinates and times. This method has traditionally been considered expensive because it requires the use of special vehicles and hired drivers. However, with the increased use of GPS in vehicles, as well as the ability to buy GPS data from routing service providers like Google, the method has become more affordable. One disadvantage of this method is that GPS probe vehicles provide small sample sizes (Wang & Yan, 2002).

2.2.4.2. License Plate Reader-based Travel Time Analysis

This method is based on the utilization of software that recognizes the license plate numbers at a location and then matches the numbers at another location. Optical Character Recognition (OCR) is the method by which the plates are recorded. With properly installed cameras, this method can yield a detection rate of up to 98% (Yasin, Karim, & Abdullah, 2010). This approach has a very

high accuracy because the detection zone is small. However, the OCR may malfunction, which results in erroneous data. The error has been reported to be around 8 percent (Pokrajac et al., 2009).

The main disadvantage of this method is that it is very expensive. Despite the advantages in accuracy, this method has only been used a few times due to the high expenses associated with buying, installing and managing the sensors.

2.2.4.3. Estimation of Travel Time Based on Historical Data

This method relies on the utilization of sensors embedded in the ground, particularly loop data. These sensors are very popular to estimate travel time. Speeds are obtained from the loops based on the average vehicle length; the travel times are compared against historical data. This method, however, may not be available for all corridors since not all roads have sensors or historical data. This method is sensitive to errors due to special events. In such cases, the error can be beyond that recommended by FHWA (Monsere & Breakstone, 2006).

2.2.4.4. MAC Address-based Travel Time Analysis

The increasing use of electronic devices in everyday life, in combination with the need for those devices to communicate with each other, has generated a considerable flow of information that is found everywhere in human societies. Businesses that contain high numbers of people like commercial centers have exploited that flow of information to determine travel patterns of individuals (Bullock, Haseman, Wasson, & Spitler, 2010).

Of the many methods to track the MAC address, Bluetooth has been indisputably the most popular method. Transportation researchers have thoroughly investigated Bluetooth tracking, especially for travel time (Ahmed, El-Dariby, Morgan, & Abdulhai, 2008; Bhaskar, Qu, & Chung, 2015; Erkan & Hastemoglu, 2016; Haghani, Hamedi, Sadabadi, Young, & Tarnoff, 2010; Haseman, Wasson, & Bullock, 2010; Quayle, Koonce, DePencier, & Bullock, 2010; Wasson, Sturdevant, & Bullock, 2008).

Bluetooth tracking is inexpensive, and the data collection is easy to execute. These characteristics partly explain the ubiquity of this approach. Additionally, the results yielded by this method are very accurate. However, some issues associated with this method include fluctuation in the detection time accuracy, spatial errors originated by utilizing different brands and software, and the great number of noise sources of MAC addresses. Errors have been modeled in order to calibrate the sensors for data collection. Nevertheless, these calibrations are hard to execute because the models are complex (Chen & Hung, 2011).

2.2.5. Machine Learning in Transportation Engineering

Machine learning is a branch of artificial intelligence that permits available computers (with specific capabilities) to perform complex tasks without being veraciously trained. Machine learning algorithms are divided into three major categories: supervised learning, unsupervised learning, and reinforcement learning. The main objective of machine learning is to identify patterns in the data for applications in different fields (Praveena & Jaiganesh, 2017).

The first instance in which machine learning was used in transportation engineering was to investigate different driving situations for autonomous vehicles (Pomerleau, 1991). Another pioneering study was in urban rail control for optimization of travel time, energy consumption, and passenger comfort (Arciszewski, Khasnabis, Khurshidulhoda, & Ziarko, 1994). Since then,

artificial intelligence has been used in various areas of transportation engineering like traffic (Bazzan, 2009; Yanyun Li, Li, & Yoshie, 2014), intersection analysis (Abdulhai, Pringle, & Karakoulas, 2003; Arel, Liu, Urbanik, & Kohls, 2010; Wiering, 2000), autonomous vehicles (Buluswar & Draper, 1998; Kuderer, Gulati, & Burgard, 2015; Litman, 2017), and more recently transit systems (Chuan Ding, Wang, Ma, & Li, 2016; Hu, Legara, Lee, Hung, & Monterola, 2016; Julio, Giesen, & Lizana, 2016).

Artificial neural networks (ANN) have been a popular approach in transportation (Chan, Dillon, Singh, & Chang, 2012; G. L. Chang & Su, 1995). Previous studies have confirmed the potential of ANNs to accurately predict traffic conditions on freeways such as traffic volumes, travel times, and speeds (Dougherty, Kirby, & Boule, 1993; H. Zhang, Ritchie, & Lo, 2007). In addition, traffic conditions on urban streets were obtained such as OD flows and bus schedule deviation (Kalaputapu & Demetsky, 1995; Toqué, Côme, Mahrsi, & Oukhellou, 2016).

All these applications show that machine learning is a powerful tool to analyze data and predict various characteristics of transportation engineering for different facilities and conditions. Machine learning will continue to be an important tool for modern research and combined with statistical analysis, will be part of transportation studies for the improvement of transportation processes, transportation infrastructure, traffic engineering, and transit systems.

2.2.6. Summary of Project Literature Review

Transit agencies have an increasing interest in investing in new technologies to better manage their transportation infrastructure network. Wireless communication technologies provide a way for information to be transferred between various devices. Different methods, like Bluetooth and Wi-Fi, have their advantages and disadvantages. The interconnectivity between these devices is called the Internet of Things. Raspberry Pi computers can be modified as scanners to collect data from this interconnectivity of devices. The data collected can be used to estimate ridership, OD matrices and travel times. Wi-Fi technology is among the least studied technology to gather data for transportation transit systems. This literature review offers preliminary information toward the research methodology, data sets, and data analysis.

3. DATASETS

In this study, all the data were collected by three methods: (1) surveys conducted onboard, in which the riders were asked about their boarding and alighting information, and the usage of Wi-Fi enabled smartphones, (2) manual counts of characteristics onboard; these characteristics are the number of passengers, their OD information, and the travel time of the buses which constitute the ground truth values used for statistical comparisons, and (3) information collected with the Raspberry Pi computers, which consists of the raw Wi-Fi and GPS data. This chapter presents descriptions of the study area, hardware, and software used and a general overview of each of the three methods used for data collection and post-processing. Additionally, the statistics and graphic representation of the quantitative variables are discussed.

3.1. Study Area

This section provides insight into the study area, bus lines, and data collection. The description presents an overview, existing conditions and current situation of the city.

3.1.1. Existing Demographics of the City of Bozeman

The study area includes the City of Bozeman. The city is the county seat of Gallatin County, which is in the southwestern part of the state of Montana. Montana State University (MSU) is in Bozeman and is the largest source of trip generation, employment, and economic activity (Cambridge Systematics Inc., 2013).

The estimated population of Bozeman in 2019 is 49,000, which makes it the fourth most populous city in Montana. The annual growth in 2017 was 3.73%, and the city is expected to grow steadily in the future (World Population, 2018). Figure 6 shows the annual growth rate of the city every decade since 1900 and every year since 2010. Table 4 shows the general demographics of the city compared to Gallatin County and the United States.

Table 4: Demographics of Bozeman, Gallatin County and United States (Taunya Fagan, 2019)

Demographics	Bozeman, MT	Gallatin County, MT	United States
Population	49,000	110,000	328,440,000
Population density/square mile	102.1	34.4	81.6
Percent male	52.60%	52.10%	50.20%
Percent female	47.40%	48.60%	49.80%
Median age	32.6	32.9	38.8
People per household	2.3	2.36	2.39
Median household income	\$44,455	\$52,833	\$46,230
Average income per capita	\$25,087	\$28,939	\$25,373

The transportation needs of the city are based on the land use and development of its socio-economic activities. Currently, the city seeks redevelopment and enhancements in the historical

Downtown and East Main Street. This activity has made Bozeman a rapidly growing city, with much of the growth related to commercial development (Robert Peccia & Associates & Alta Planning + Design, 2017).

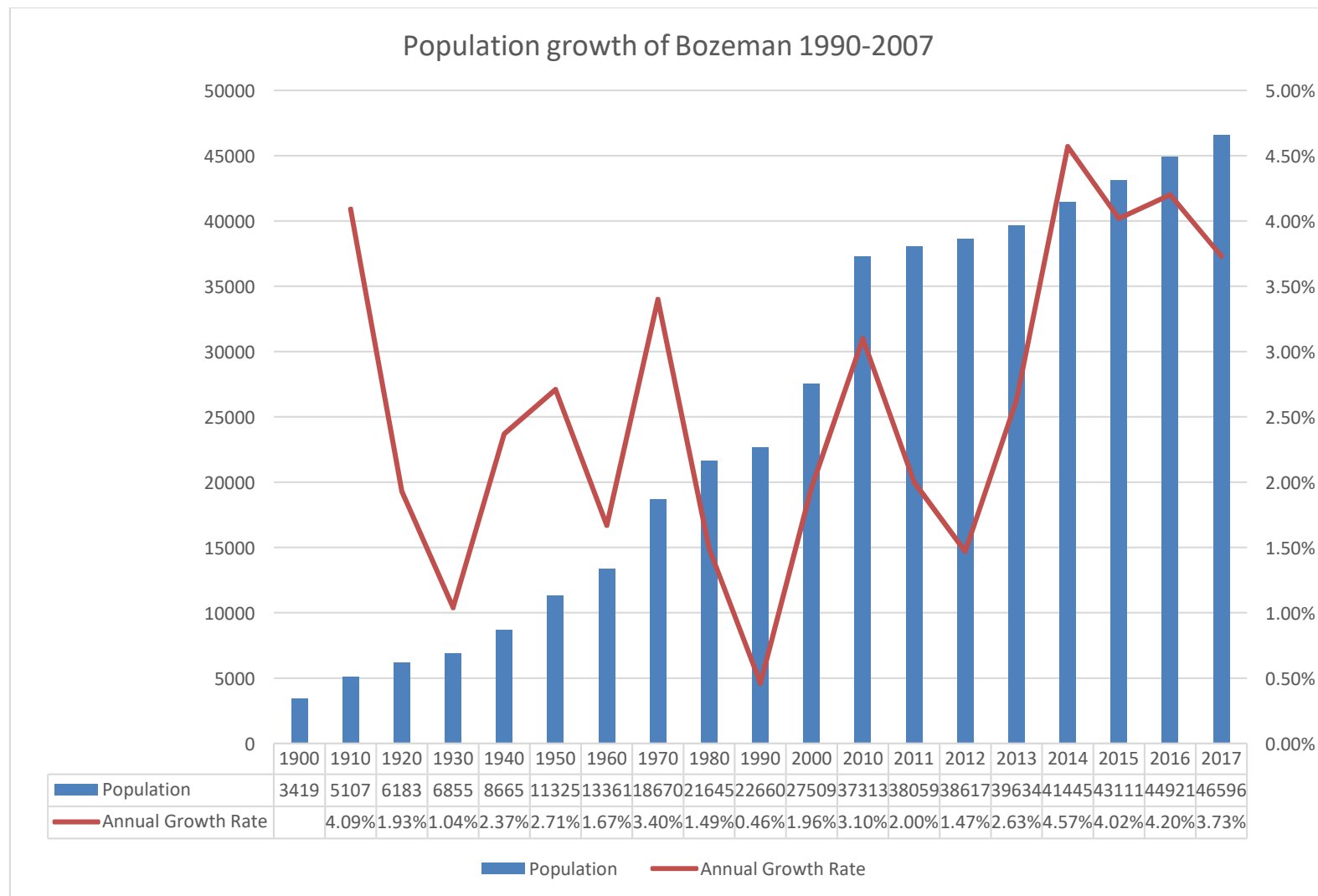


Figure 6: Population growth in Bozeman

In the future, the city intends to migrate from auto-oriented development planning to more mixed-use developments that aim to increase the density of the urban population. Center-based commercial development is still being pursued (Bozeman City Commission, 2009).

3.1.2. Overview of Streamline

Streamline provides fare-free fixed route public transportation in Bozeman, Belgrade, and Livingston, and seasonal routes to the Bridger Bowl ski area. The routes provide mobility during all seven days of the week; however, the frequency during weekends is reduced. The buses run a loop departing from Montana State University every hour and coming back to the same bus stop. During peak hours, there are two buses running, doing loops every half hour. The peak times are from 7:00 AM to 9:00 AM and from 4:00 PM to 6:00 PM. Riders are largely MSU students, faculty, and staff. During the month of December, the ridership decreases, partly due to school holidays. For the period between 2007 and 2015, the ridership rate has averaged a 7.5 percent increase annually. For Fiscal Year 2017, the Streamline budget was 1.6 million dollars (HRDC, 2019).

This study focuses on the daytime weekday service. This service runs from Monday to Friday starting at 6:30 AM for some lines and finishing at 7:15 PM at MSU. There are five lines that run in this service: blue, green, orange, red and yellow. Below, the characteristics of each line are presented.

- **Blueline Retail.** The blue route provides service for Montana State University, Downtown, Bridger Peaks and the Gallatin Center. It is designed to provide access to retailers, grocers, and businesses on the north side of town. The duration of the loop, from the time it departs and comes back to MSU, is 54 minutes. It serves 38 bus stops. Figure 7 shows the route traveled by the Blueline buses.
- **Greenline Express.** The green route provides service for Montana State University, Gallatin Valley Mall, Four Corners, and Belgrade. It is designed to meet the rising demand for a public commuter service between Bozeman and Belgrade. The duration of the loop, from the time it departs and comes back to MSU, is 61 minutes. It serves 26 bus stops. Figure 8 shows the route traveled by the Greenline buses.
- **Orangeline University.** The orange route provides service for Montana State University, Downtown, the Public Library, and Bozeman Deaconess Hospital. This route provides access to the eastern areas of Bozeman and the city hospital. The duration of the loop, from the time it departs and comes back to MSU, is 26 minutes. It serves 18 bus stops. Figure 9 shows the route traveled by the Orangeline buses.
- **Redline Downtown.** The red route provides service for Montana State University, Downtown, Gallatin Valley Mall and Bozeman High School. It is the main route to go east and west. Redline is designed for Downtown access and connecting intersecting routes. The duration of the loop, from the time it departs and comes back to MSU, is 50 minutes. It serves 39 bus stops. Figure 10 shows the route traveled by the Redline buses.
- **Yellowline University.** The yellow route provides service for Montana State University, Gallatin Valley Mall and Valley Commons. It is designed to provide access to the far west side of town. The duration of the loop, from the time it departs and comes back to MSU, is

All the bus lines are shown in Figure 12. The weekday schedule is a fixed route service. All the routes depart from Montana State University. There are three transfer stations across the system: Montana State University is a point in common for all five lines; Gallatin Valley Mall is a connection for the green, red and yellow lines; and the Downtown Transfer Station connects the blue, orange and red lines.

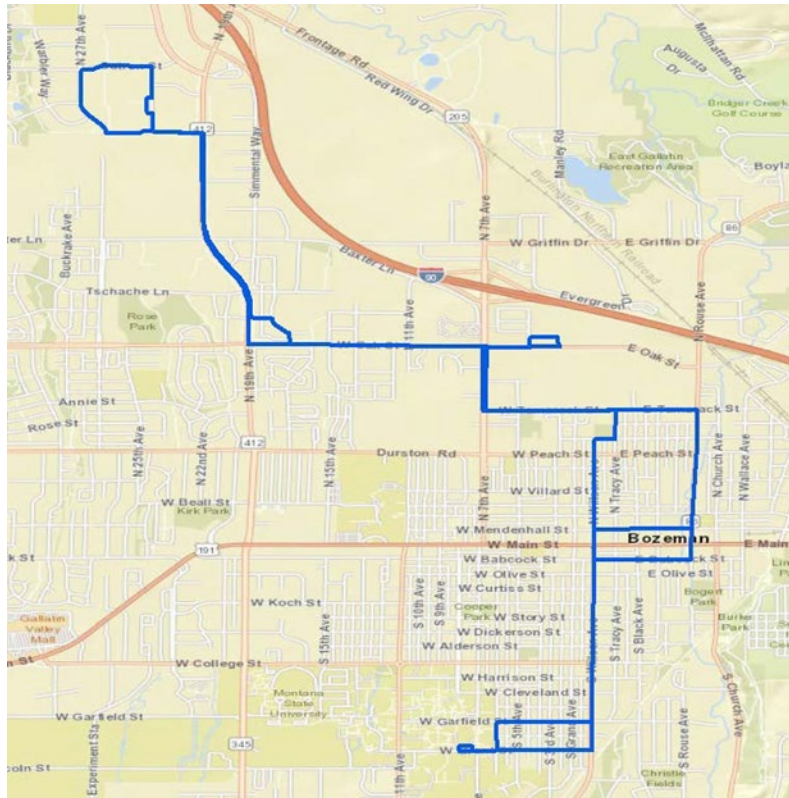


Figure 7: Streamline Blueline transit route

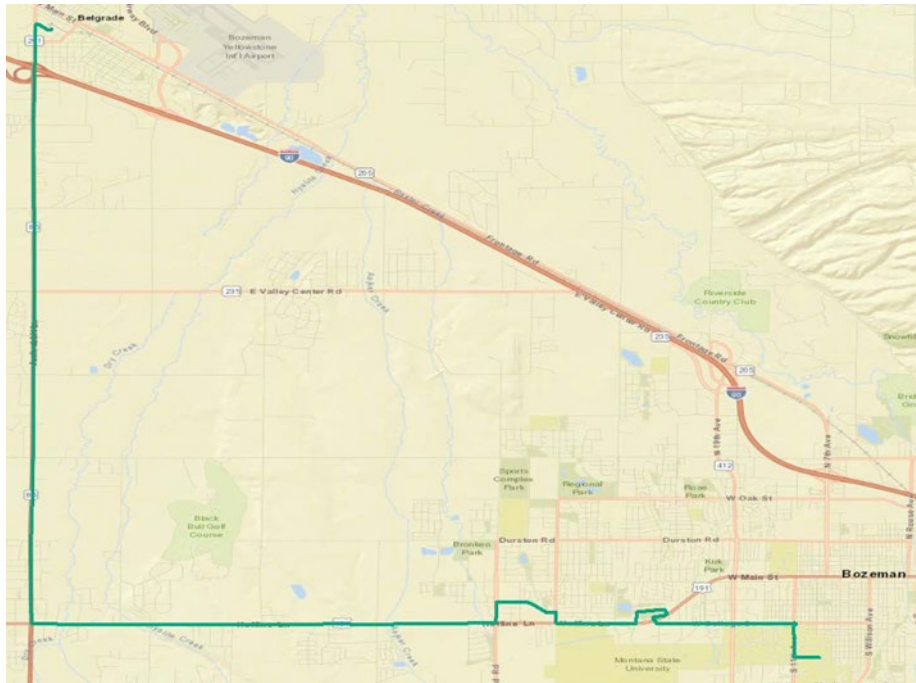


Figure 8: Streamline Greenline transit route

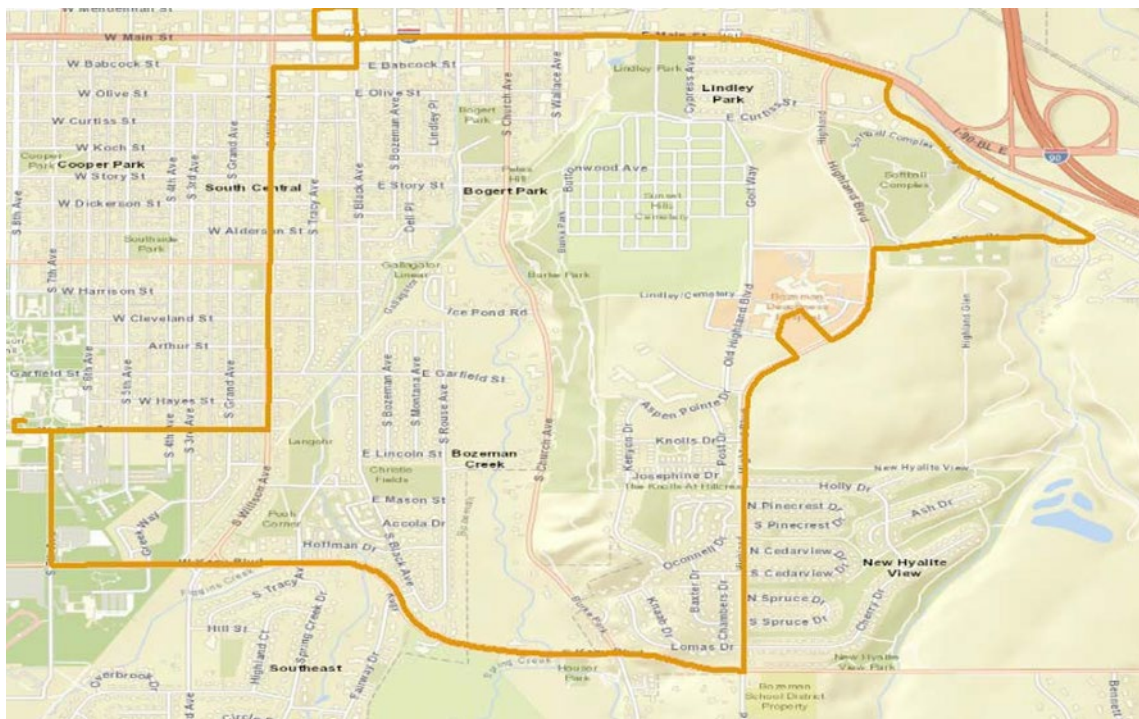


Figure 9: Streamline Orangeline transit route

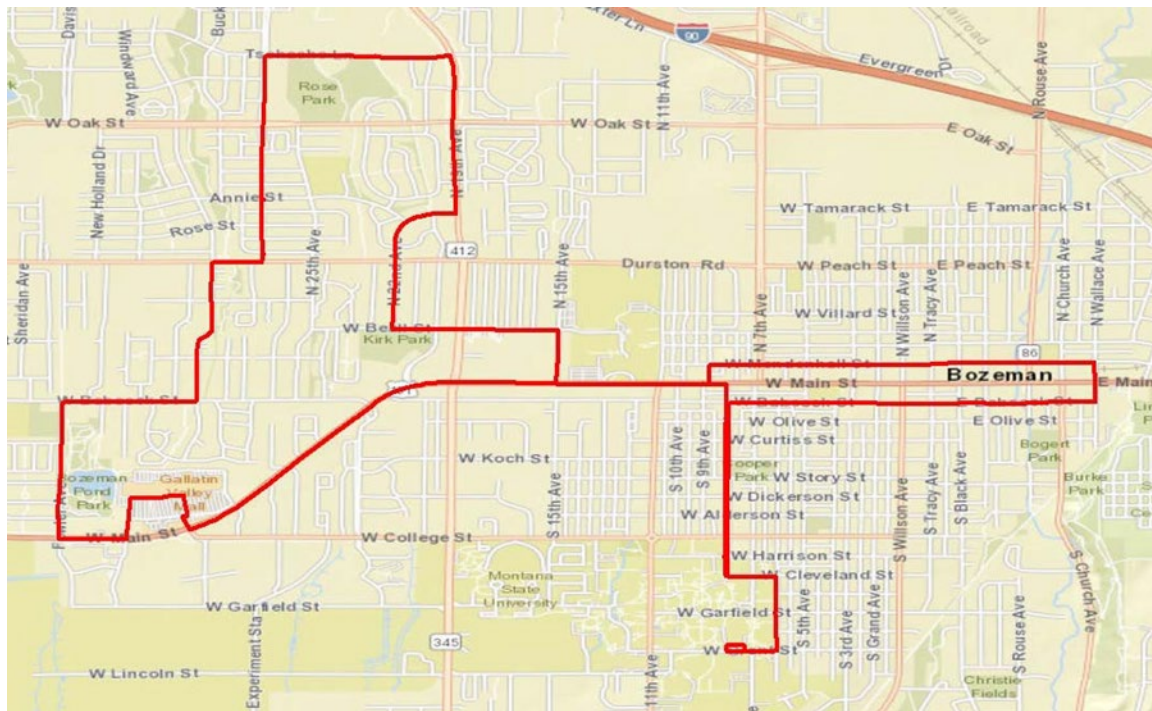


Figure 10: Streamline Redline transit route

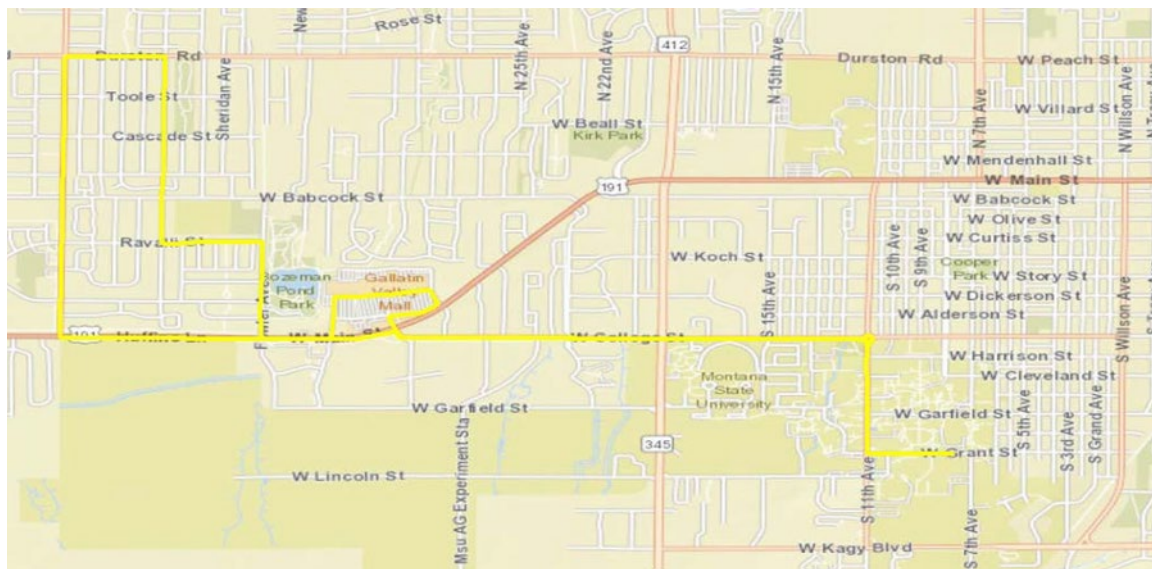


Figure 11: Streamline Yellowline transit route

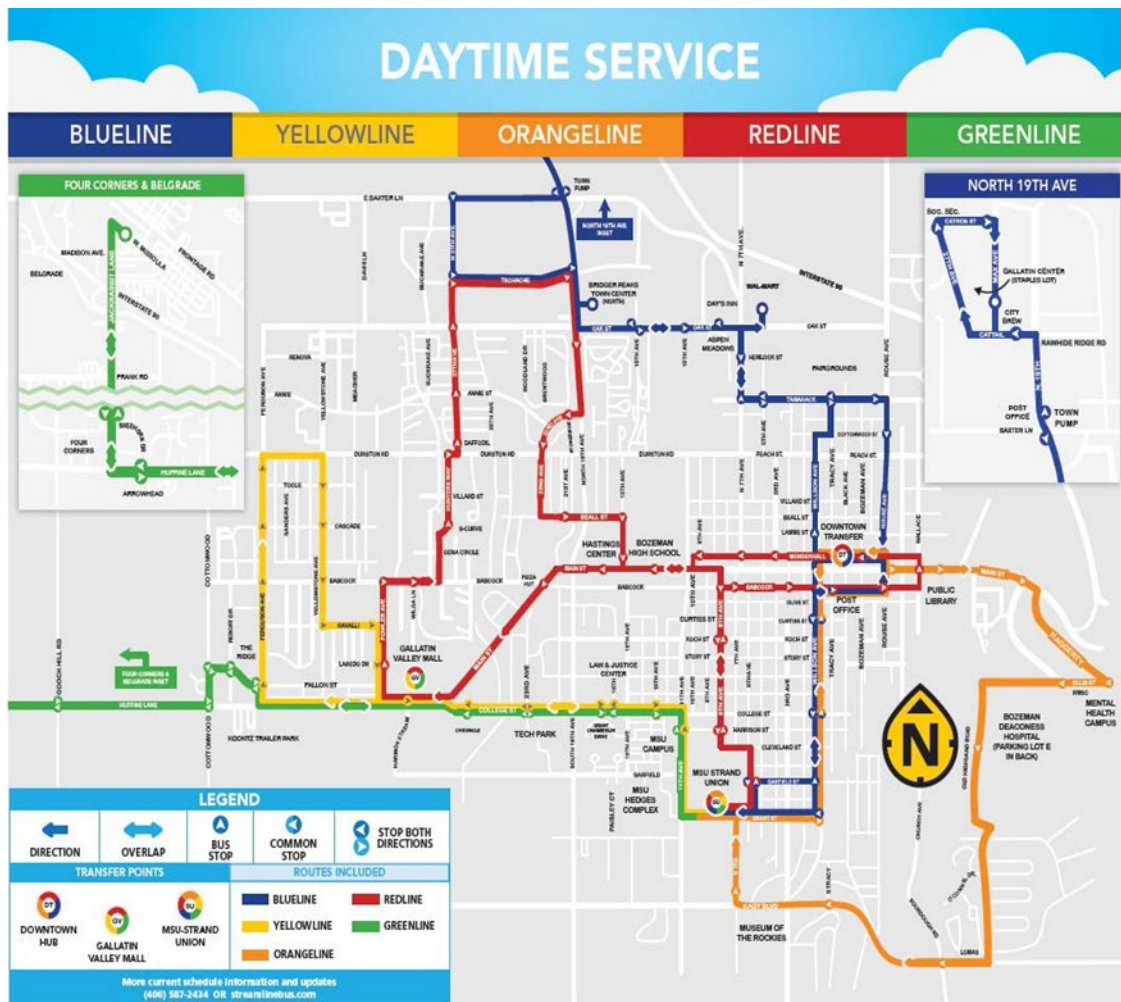


Figure 12: Streamline daytime service route map (HRDC, 2019)

3.2. Hardware and Software

This study employed sensors to collect Wi-Fi signals and track bus GPS. An innovative set of hardware and software was utilized in the data collection, and the set was called Smart Station (SS). This section covers the approach utilized for developing a reliable, non-intrusive and passive mechanism for data collection. The hardware and software technical details are explained below.

3.2.1. Hardware

Given the requirements of sensors to be deployed onboard the buses and that they had to function independently, the Raspberry Pi (a single-board Linux computer) with a portable power source was an ideal option. In order to sense Wi-Fi devices, a series of components were integrated together as listed below:

- Raspberry Pi. The sensor hardware. This is a 900MHz quad-core ARM Cortex-A7 CPU. Model 3B was used throughout this research. The CPU unit can process data with low

power consumption and is operated through Linux. It requires a micro SD card to store the operating system and the software used to detect Wi-Fi signals.

- Wi-Fi Adapter. This device receives wireless signals. It was adapted in monitor mode to become a de facto router and detect Wi-Fi signals. The Ralink 5370 Chipset allows this mode, and this was the chipset used in all the Wi-Fi adapters.
- Power Supply. A 5V/2A portable battery was used to provide electric current to the Raspberry Pi and all the other components that were connected to it.
- GPS Receiver. This device receives the signals sent by the GPS satellites to obtain the location of the Smart Station. It consists of a GPS module with a chip microcomputer and an active antenna.
- Internet Provider. The internet was provided through a mobile hotspot that rode on the buses next to the Smart Station. Another way to establish an internet connection is by connecting to a router that provides internet, which was done whenever possible. Via the web, the data collected could be stored online and retrieved from another device.

Figure 13 illustrates the components used for data collection. Figure 14 shows the placement of the Smart Station inside the buses.

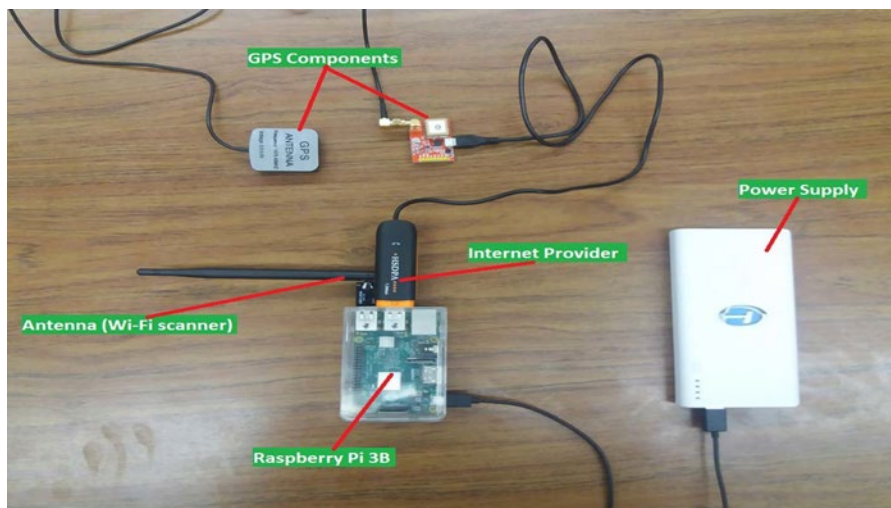


Figure 13: Components of the Smart Station

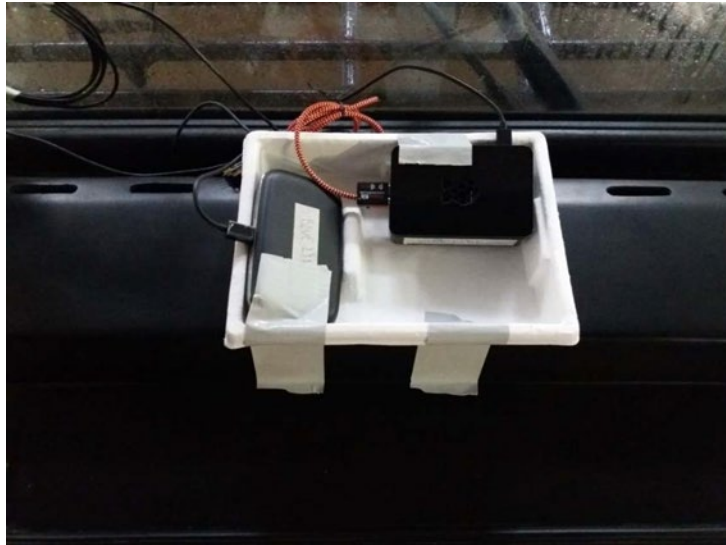


Figure 14: Placement of Smart Station on the buses

3.2.2. Software

The Raspbian operating system was installed in the Raspberry Pi computers, which is a Linux-based, open-source operating system. Raspbian is the officially supported operating system for Raspberry Pi Computers (Raspberry Pi Foundation, 2019). The software used to scan for Wi-Fi signals was Kismet Wireless, which is a program that passively detects Wi-Fi (IEEE 802.11) networks. This software presents information about the source device that is generating the Wi-Fi signal, including manufacturer, average estimated GPS coordinates, network channel, and MAC address (Kismet Wireless, 2019). The GPS program used in processing the information obtained by the receiver was gpsd, which is an open-source, Linux compatible GPS logger (Cathedral and the Bazaar, 2019).

The computers were programmed to run Wi-Fi detection software and to collect GPS data after rebooting. Wi-Fi signals may be sensitive to physical barriers and weather (Bai, Ireson, Mazumdar, & Ciravegna, 2017), and they do not directly equate to the number of passengers. This is partially associated with some people carrying many Wi-Fi enabled devices or no devices at all. Additionally, there are other sources of noise like routers, pedestrians, and people nearby who carry devices and are driving other vehicles. Additionally, because of the design of Wi-Fi detection itself, the results usually under or overestimate the actual number of people.

This research aimed to determine if a reasonable estimate of the transit ridership can be obtained using the hardware and software previously explained. Strategies, like rule-based algorithms and machine learning techniques, were introduced to overcome these issues. The details on dates and equipment deployment are explained in the following sections.

3.3. Surveys

The purpose of the manual surveys was to provide ground truth data for passengers' OD flows characteristics, the manufacturer of the devices that passengers carried, and smartphone ownership rates. Additionally, passengers could respond to an extra question in order to give feedback to the

Streamline bus service. The structural content of this survey is shown in Table 5. A survey sample is shown in Appendix A.

Table 5: Description of the surveys conducted with passengers

Question	Description
1	Asked if passengers were carrying any Wi-Fi enabled device like smartphones, tablets, and others. Passengers could also indicate the number of devices they carried.
2	Asked what the brand names of the devices were.
3	Asked the boarding and alighting stops. Passengers only had to check the bus stop name. This question was different for each line.
4	Asked if passengers had any feedback to improve the Streamline service.

The surveys were implemented during a period of two weeks, from April 2 to April 13 of 2018. For every line, five days were surveyed. On every survey day, the surveys were distributed to all the passengers who agreed to participate and were at the buses from the departure of the MSU stop until it completed the loop and came back to the bus stop. A total of 25 loops were surveyed. The surveys were implemented for a total of 1,075 minutes, which is nearly 18 hours. Table 6 shows a detailed representation of the schedule.

During this same time, Wi-Fi data were being scanned by the Smart Stations. However, details about the procedure and the data will be presented in the Smart Station Data section of this chapter.

Table 6: Schedule of survey implementation and data collection

Line	Day 1	Day 2	Day 3	Day 4	Day 5
Blue	April 2, 2018 (3:21 PM to 4:15 PM)	April 3, 2018 (4:21 PM to 5:15 PM)	April 4, 2018 (7:51 AM to 8:45 AM)	April 5, 2018 (9:21 AM to 10:15 AM)	April 6, 2018 (8:51 AM to 9:45 AM)
Green	April 2, 2018 (12:10 PM to 1:15 PM)	April 10, 2018 (12:10 PM to 1:15 PM)	April 11, 2018 (7:15 AM to 8:16 PM)	April 12, 2018 (5:15 PM to 6:20 PM)	April 13, 2018 (7:15 AM to 8:16 PM)
Orange	April 4, 2018 (4:45 PM to 5:11 PM)	April 5, 2018 (12:45 PM to 1:11 PM)	April 6, 2018 (4:45 PM to 5:11 PM)	April 9, 2018 (4:45 PM to 5:11 PM)	April 10, 2018 (1:45 PM to 2:11 PM)
Red	April 3, 2018 (3:49 PM to 4:39 PM)	April 4, 2018 (12:19 PM to 1:09 PM)	April 5, 2018 (2:19 PM to 3:09 PM)	April 6, 2018 (7:19 AM to 8:09 AM)	April 9, 2018 (3:19 PM to 4:09 PM)
Yellow	April 9, 2018 (7:15 AM to 7:39 AM)	April 10, 2018 (3:15 PM to 3:39 PM)	April 11, 2018 (4:15 PM to 4:39 PM)	April 12, 2018 (12:15 PM to 12:39 PM)	April 13, 2018 (4:15 PM to 4:39 PM)

During the 25 loops, a total of 394 people was observed to board the buses. Among them, 263 agreed to respond to the surveys, resulting in a response rate of 67 percent. A total number of 248 people among those surveyed, indicated that they were carrying a Wi-Fi enabled device, indicating a penetration rate of almost 95 percent. This penetration rate is higher than the smartphone penetration rate of the United States, which is estimated to be around 71 percent in 2019 (Statista, 2019). This could be attributable to the fact that most Streamline users are MSU students, faculty and staff, and are expected to have more access to mobile device technology. This penetration rate of 95% is convenient for this research study. Figure 15 shows the percent distribution of the devices by the manufacturer as reported by the passengers.

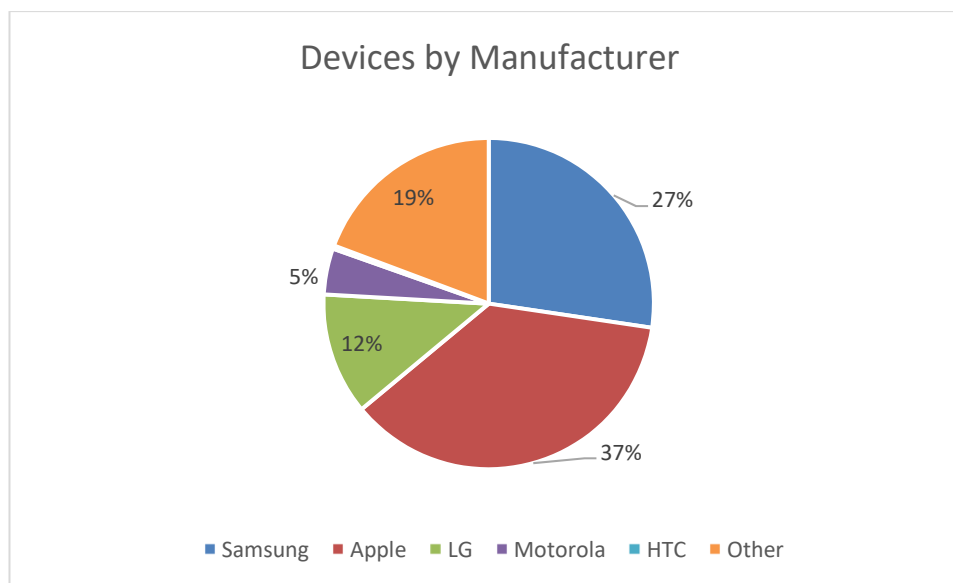


Figure 15: Percent Distribution of Devices by Manufacturer

The smartphone distribution by brand indicates that the most frequent manufacturer is Apple. This represents a challenge for data analysis because Apple devices randomize their MAC addresses. For other brands, it is possible to directly observe the devices and the OD characteristics of passengers that carry those devices.

In the following tables, the OD information is presented by bus lines because the different lines have different stops. The origins and destinations reported by the passengers during all five days were added and are presented in the following tables. Table 7 shows the OD matrix for the Blue line. Table 8 shows the OD matrix for the Green line. Table 9 shows the OD matrix for the Orange line. Table 10 shows the OD matrix for the Red line. Table 11 shows the OD matrix for the Yellow line. In these tables, the rows represent the bus stops as the place where passengers originally boarded. The columns represent the bus stops as destinations for passengers. T_i represents the total number of passengers that boarded at a bus stop. T_j denotes the total number of passengers that alighted at a bus stop.

Table 7: OD matrix of Blueline from the survey data

Blue	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	Ti
1	0	0	0	2	0	2	2	2	0	0	0	0	6	0	1	1	0	0	2	1	1	0	1	0	0	0	0	0	0	0	0	1	0	0	1	23
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
5	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
6	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	5
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	2
23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 7 (continued)

Blue	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	Ti
27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	
28	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
29	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
30	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	
31	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
32	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	
33	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
34	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
35	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Tj	0	0	0	2	0	2	2	3	1	0	0	0	6	1	1	4	0	0	4	2	2	0	1	0	0	0	0	0	0	4	1	1	0	0	9	46

Table 8: OD matrix of Greenline from the survey data

Green	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	Ti
1	0	0	3	2	0	1	0	0	1	1	0	2	2	0	0	0	0	0	0	0	0	0	0	0	0	12
2	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
3	0	0	0	0	0	0	0	0	0	0	0	0	2	1	0	0	0	0	0	0	0	0	0	0	0	3
4	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	0	1	0	0	2	0	0	0	0	0	0	0	0	0	1	4
7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1
13	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1	1	0	1	1	4	11

Table 8 (continued)

Green	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	Ti
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Tj	0	0	4	2	1	1	0	0	1	1	0	4	5	1	4	1	0	0	0	1	1	0	1	1	9	38

Table 9: OD Matrix of Orangeline from the survey data

Orange	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	Ti
1	0	0	2	0	2	1	0	0	0	3	1	2	1	0	0	0	0	0	0	12
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1
6	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	3	5
7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1
9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1

Table 9 (continued)

Orange	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	Ti
11	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	2
13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Tj	0	0	2	0	3	2	0	0	0	4	2	2	1	1	0	0	0	0	6	23

Table 10: OD Matrix of Redline from the survey data

Red	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	Ti	
1	0	0	0	0	2	1	1	2	1	2	4	0	0	0	0	4	3	3	0	0	2	5	0	1	0	0	2	0	1	1	1	0	0	0	0	0	0	0	0	0	36
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
7	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	
8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7	
12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	

Table 10 (continued)

Red	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	Ti	
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	1	0	0	0	2	6	
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	2	4	
22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	2	
23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	
24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	2	
28	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
29	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2	
30	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
31	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
32	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	
33	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	
34	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	
35	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	
36	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
37	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
38	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
39	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Tj	0	0	0	0	2	1	1	2	1	2	6	0	1	0	0	11	4	7	0	0	2	5	0	1	0	0	2	0	4	1	1	0	1	0	2	0	0	3	11	71	

Table 11: OD matrix of Yellowline from the survey data

Yellow	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	Ti
1	0	0	6	7	0	3	1	2	1	1	3	1	2	0	0	0	0	0	0	0	0	27
2	0	0	0	0	0	0	0	1	0	0	2	0	0	0	1	0	0	0	0	0	0	4
3	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	5	7
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	2
5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	1	3
7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	2
9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	2
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	2
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Tj	0	0	6	7	0	3	4	3	1	1	6	1	2	0	1	1	0	0	0	0	17	53

In the tables above, it can be noted that the Blueline has three bus stops less, Greenline has one bus stop less, and Orangeline has one bus stop less than mentioned in the Overview of Streamline section of this chapter. This difference is because more bus stops were added to those lines in August 2018.

For the Blueline, MSU (bus stop 1) is the major generator of passengers. Also, Wal-Mart (bus stop 13) is another main generator of passengers. The other bus stops evenly add passengers to this bus line. The main destinations of this line are MSU (bus stop 35), Wal-Mart (bus stop 13), Bridger Peaks Town Center (bus stop 16), the Social Security Office (bus stop 19) and Bozeman Clinic (bus stop 30). This coincides with the purpose of this line to serve commercial areas of the city.

For the Greenline, MSU (bus stops 1 and 25) and Smith & Missoula in Belgrade (bus stop 13) are the major generators and receivers of passengers. This aligns with the intention of this line to connect the Bozeman urban area and the neighboring urban area of Belgrade, which is approximately 10 miles from Bozeman.

The Orangeline's OD matrix shows that the main contributors to passengers' boarding are MSU (bus stop 1) and Downtown Transfer Station (bus stop 6). The principal destinations are Ellis & Haggerty (bus stop 10) and MSU (bus stop 19). Ellis & Haggerty is a residential area in the eastern part of Bozeman. These characteristics match the purpose of the line to connect downtown and other connections in the eastern areas of Bozeman. It is important to mention that, although the orange line is the only one that serves the city hospital, this stop does not generate most of the passengers' movements. However, this stop is important because it provides a connection to the only hospital in the city.

On the Redline, MSU (bus stop 1) singlehandedly represents around 50% of the origins. Downtown (bus stop 11) is another major generator of passengers. The third largest generator is Gallatin Valley Mall (bus stop 18). For the destinations, the bus stops evenly receive passengers over the entire system. Nevertheless, three bus stops were reported to receive more passengers: MSU (bus stop 39), Hasting Center on Main Street (bus stop 16) and Gallatin Valley Mall (bus stop 18).

The Yellowline reported having two main generators of riders: MSU (bus stop 1) and College & 16th (bus stop 3). Bus stop 3 serves the Family and Graduate Housing of the university. Most passengers alighted the line at MSU (bus stop 21). Also, College & 16th (bus stop 3), College & 23rd (bus stop 4) and Yellowstone & Toole (bus stop 11) are principal destinations. This passenger behavior reflects the purpose of the line to serve passengers on the west side of the city.

In conclusion, Streamline's weekday service seems to overwhelmingly serve the MSU population. The other transfer stations: Downtown and Gallatin Valley Mall, also show high boarding and alighting numbers.

Regarding the recommendations that passengers provided for Streamline, it can be noted that the major concern they have was the accuracy of the information provided on the company's webpage or mobile app. Additionally, some passengers suggested an extension of service times, adding more routes and destinations, improving the bus stops, cleaning the buses or improving them, and improving the punctuality of the service. It is worthy to mention that a lot of passengers praised the system and felt they were grateful for it. Figure 16 shows the main recommendations made by all the passengers who added comments in the feedback section. In total, 78 passengers provided feedback for the service.

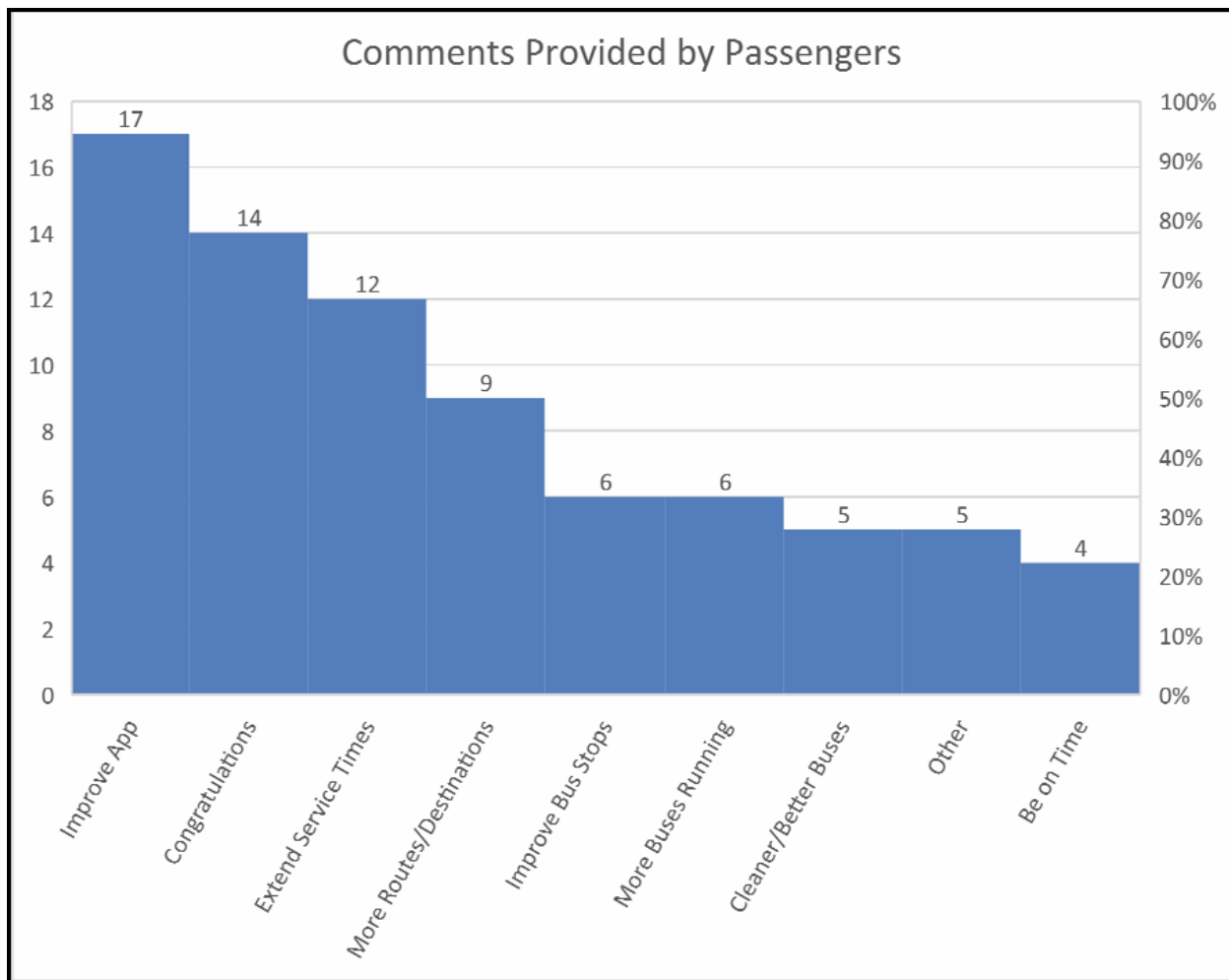


Figure 16: Feedback provided by passengers

3.4. Manual Counts

Manual counts were performed at the same time the Smart Station (SS) was collecting data. The purpose of the onboard observations was to provide a comparison frame for the data collected with the innovative approach of the Smart Stations. During the manual counts, the researchers collected three variables: (1) the number of passengers that boarded and alighted at each bus stop and the associated timestamps, (2) the travel times between the bus stops, and (3) time durations while a bus was stopped at a bus stop. This section describes the instruments used to collect these data and their statistical characteristics, such as averages and standard deviations.

3.4.1. Number of Passengers

The number of passengers was counted by a surveyor that was onboard. The surveyor counted the initial total of passengers at the bus stop when the counts were started at the MSU bus stop. Later, the surveyor counted the number of passengers that boarded and alighted at each bus stop where

the buses made a stop. These data were collected for two different periods: first, when the pilot study was made in April at the same time the surveys were performed and secondly when the data were collected to make the statistical studies.

3.4.1.1. Pilot Data

The pilot data consisted of counting the passengers at each of the bus stops. On this occasion, the travel times were not obtained. The total number of passengers that used the buses during this time is reported in Figure 17. It is observed that the most popular lines are red and yellow, with a total of 144 and 86 passengers, respectively.

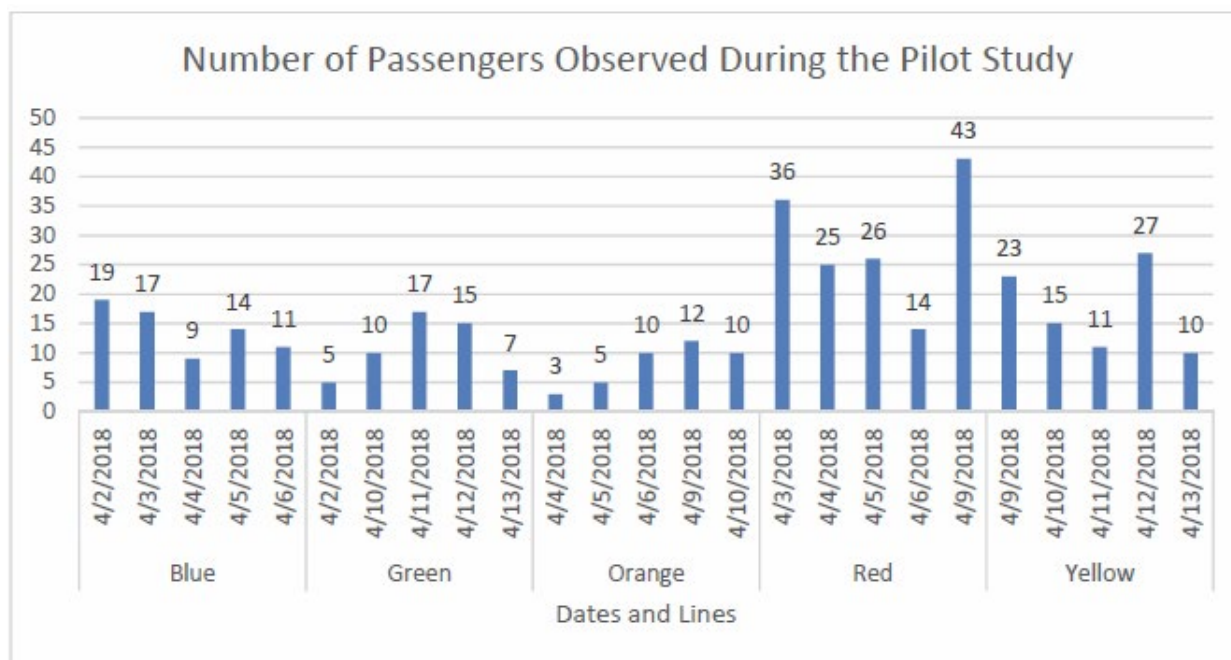


Figure 17: Number of passengers observed during the pilot study

The information on boarding and alighting of the passengers is shown in Figure 18 for the Blueline, Figure 19 for the Greenline, Figure 20 for the Orangeline, Figure 21 for the Redline and Figure 22 for the Yellowline.

For the Blueline, it can be noted that the bus stop that generates and receives the most passengers is the MSU bus stop. Additionally, the Wal-Mart bus stop is another popular origin and destination for passengers. The bus stops near the Downtown area are also popular among passengers. This is consistent with the results obtained from the survey response.

The Greenline shows that MSU and Smith & Missoula bus stops are the major generators and receivers of passengers. Additionally, the housing areas near the MSU campus are areas that show high numbers of passenger movements. These results are consistent with the OD matrix obtained from the surveys.

On the other hand, the Orangeline seems to have passengers using the Bozeman Public Library and the Bozeman Deaconess bus stops. This was not reflected in the surveys. This could be due to

the smaller sample size used in the surveys, or that passengers going to the hospital and the library did not have time to respond or preferred not to do it for other reasons. In the latter case, there could be a response bias. However, the MSU bus stop is consistently a popular stop.

The Redline shows that Hunter's Way is a popular stop that was not reflected in the OD matrix. For the rest, the MSU bus stop and the businesses on Main Street show a high influx and efflux of passengers.

The Yellowline observed data is consistent with the OD flow characteristics obtained from the surveys. The main bus stops are MSU and its surroundings. The Yellowstone & Toole bus stop seems to be very popular as well.

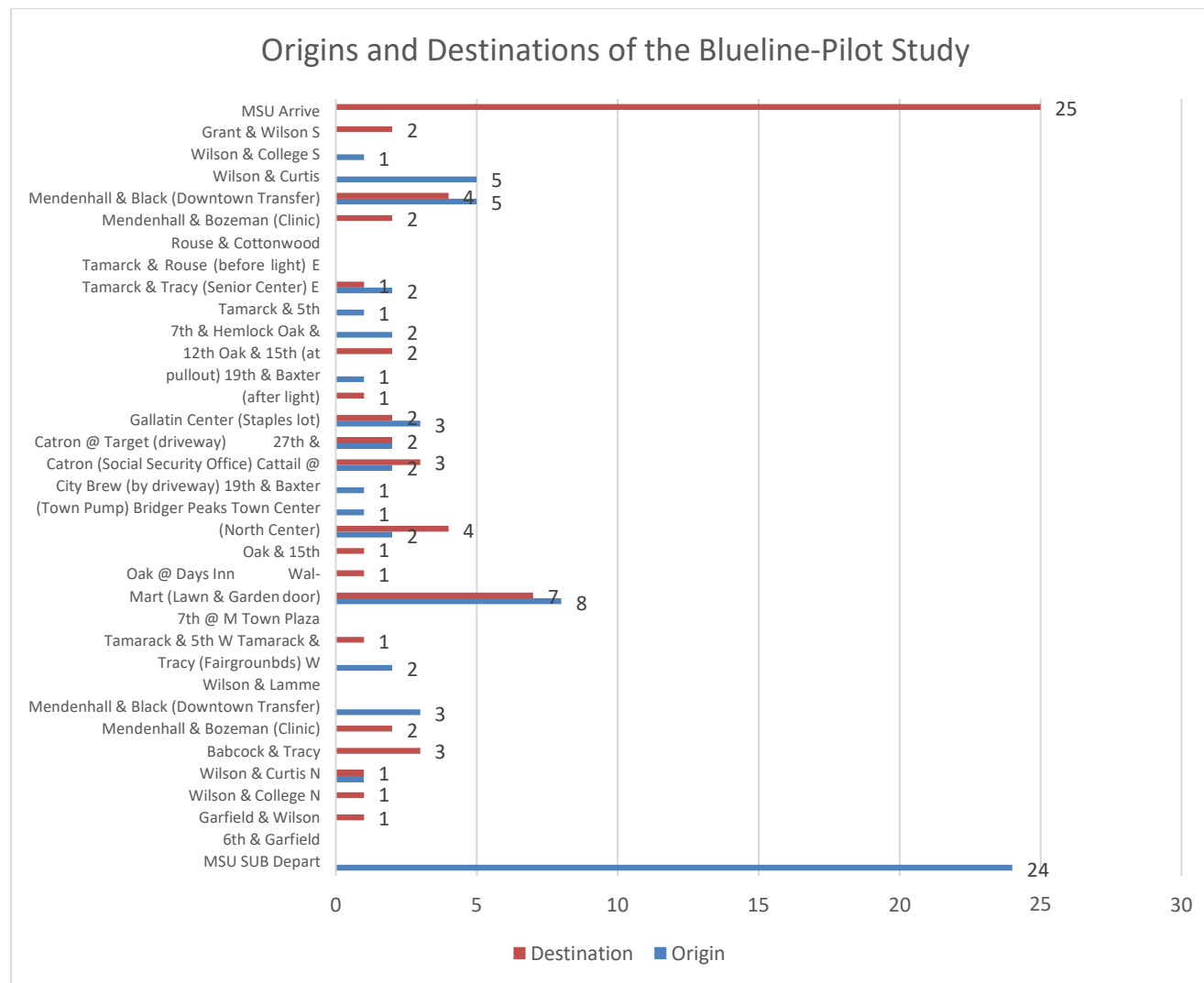


Figure 18: Origins and destinations of the Blueline from the pilot study

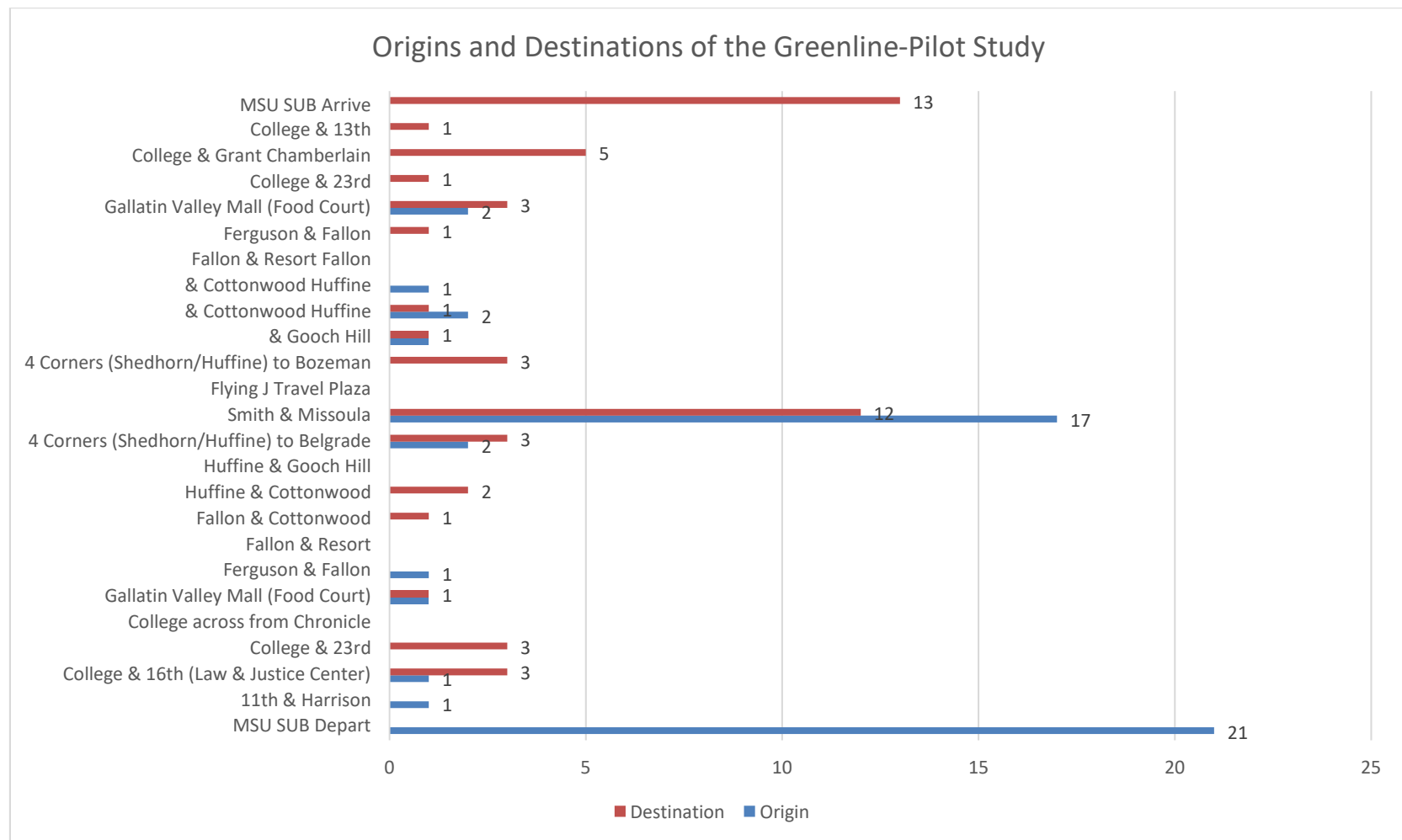


Figure 19: Origins and destinations of the Greenline from the pilot study

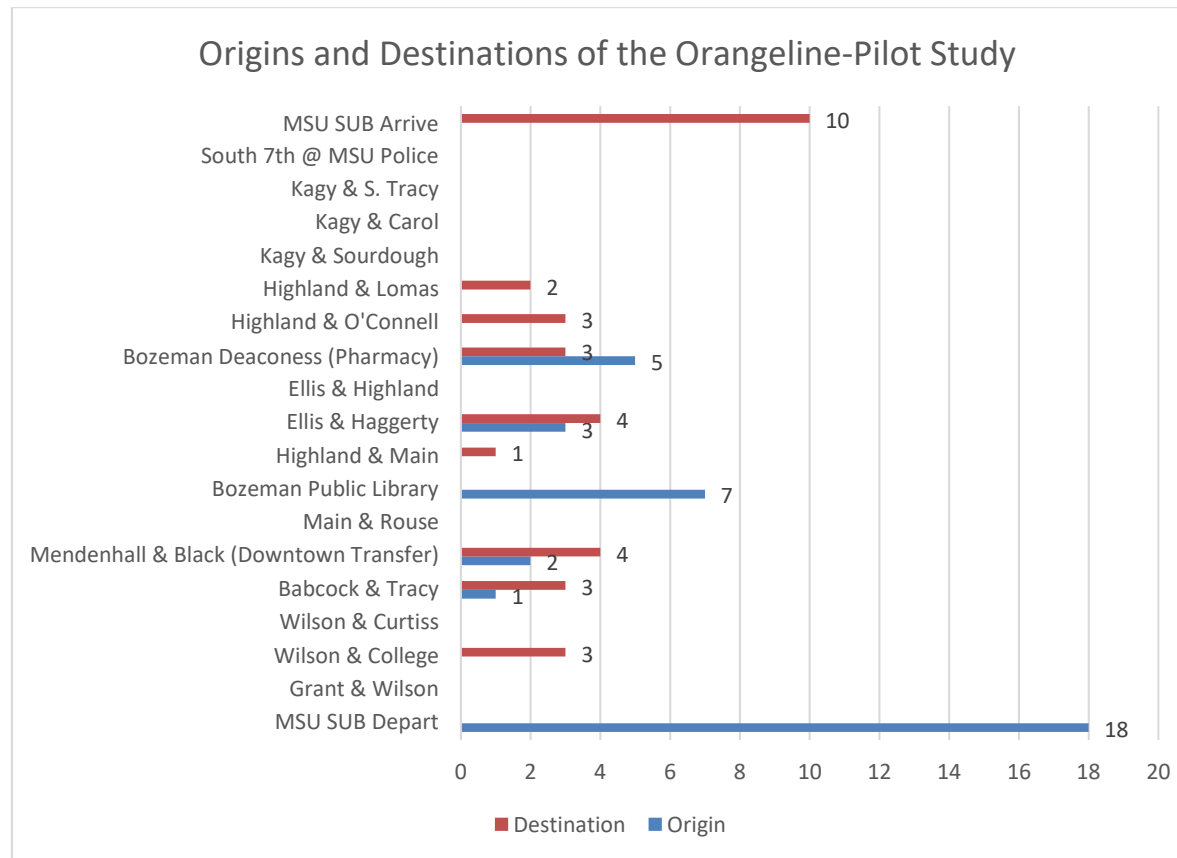


Figure 20: Origins and destinations of the Orangeline from the pilot study

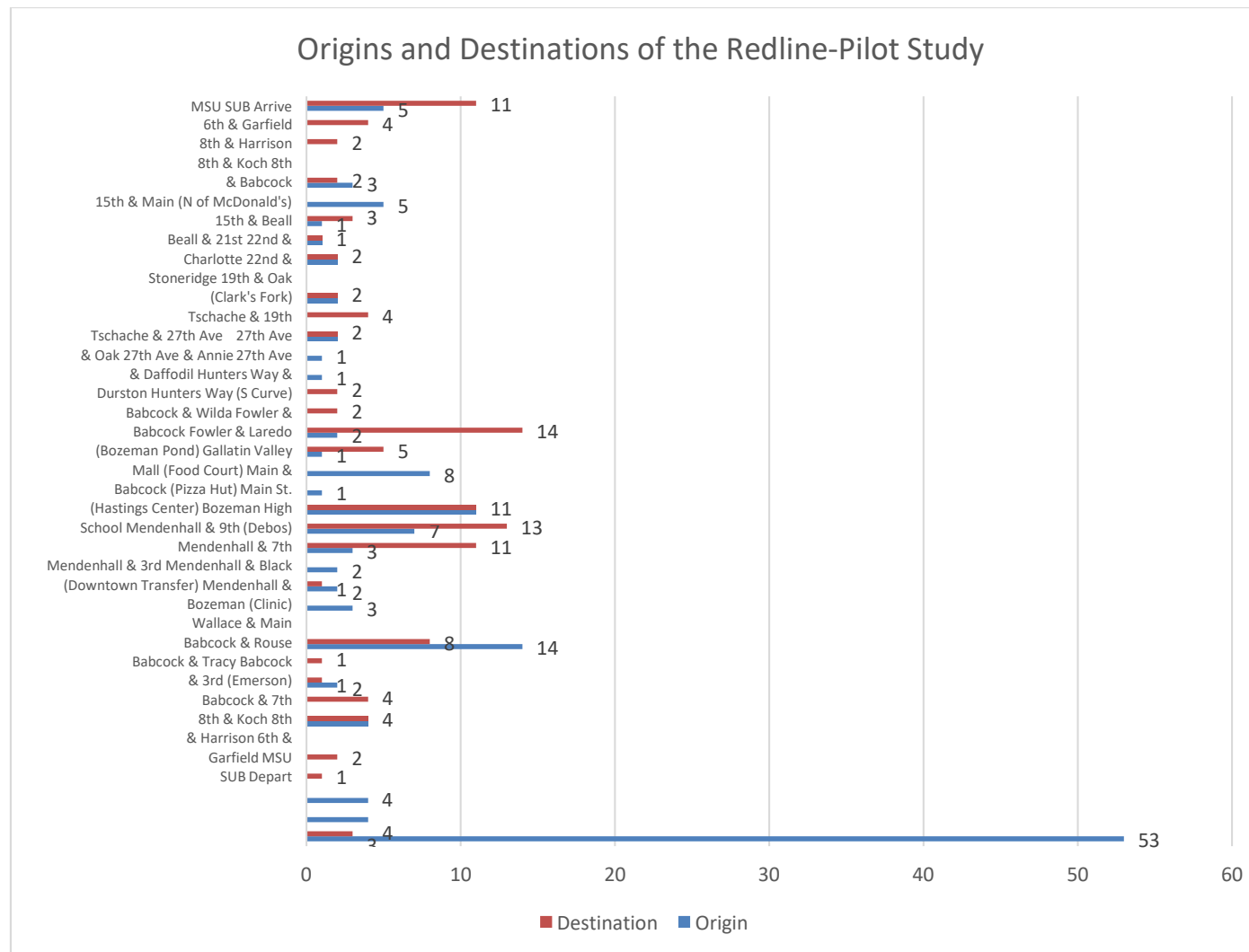


Figure 21: Origins and destinations of the Redline from the pilot study

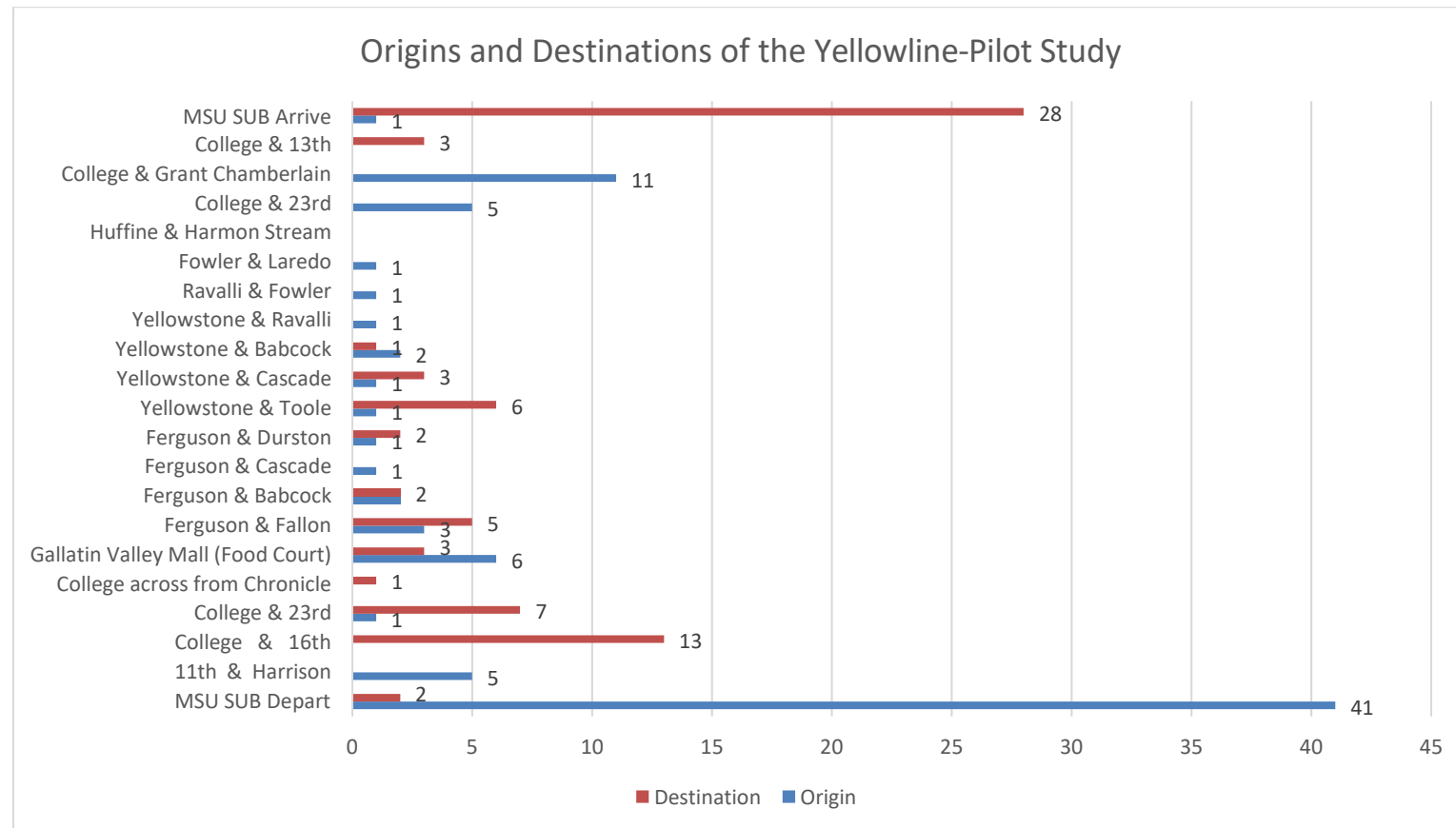


Figure 22: Origins and destinations of the Yellowline from the pilot study

3.4.1.2. Data Collection

The number of passengers was collected in the same manner as in the pilot study. These data were collected in the months of October 2018 for two weeks and January 2019 for three weeks. On this occasion, the travel times were obtained and will be discussed later in this chapter. The total numbers of passengers observed per line are shown in Figure 23.

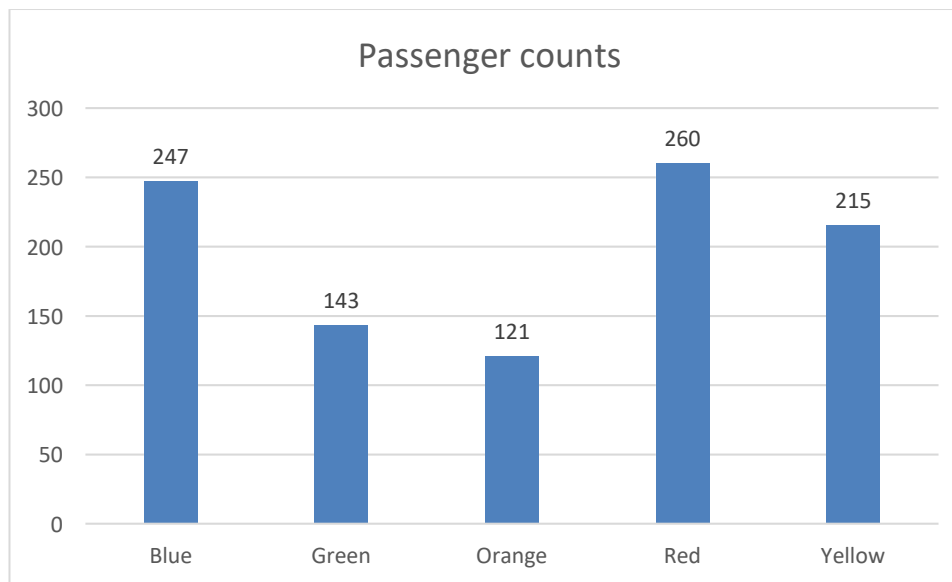


Figure 23: Sum of passenger counts per line

The number of passengers per day is shown in Figure 24 for the Blue line, Figure 25 for the Green line, Figure 26 for the Orange line, Figure 27 for the Red line and Figure 28 for the Yellow line. It is vital to note that each day represents a loop, meaning a bus departing MSU and returning there. Therefore, the counts are not for a full day of activity.

On average, the Blue line shows that approximately 25 passengers boarded the buses for each loop. The standard deviation is around 9 passengers. This shows a big change relative to the mean. Therefore, the buses' occupancy varies substantially by days and times.

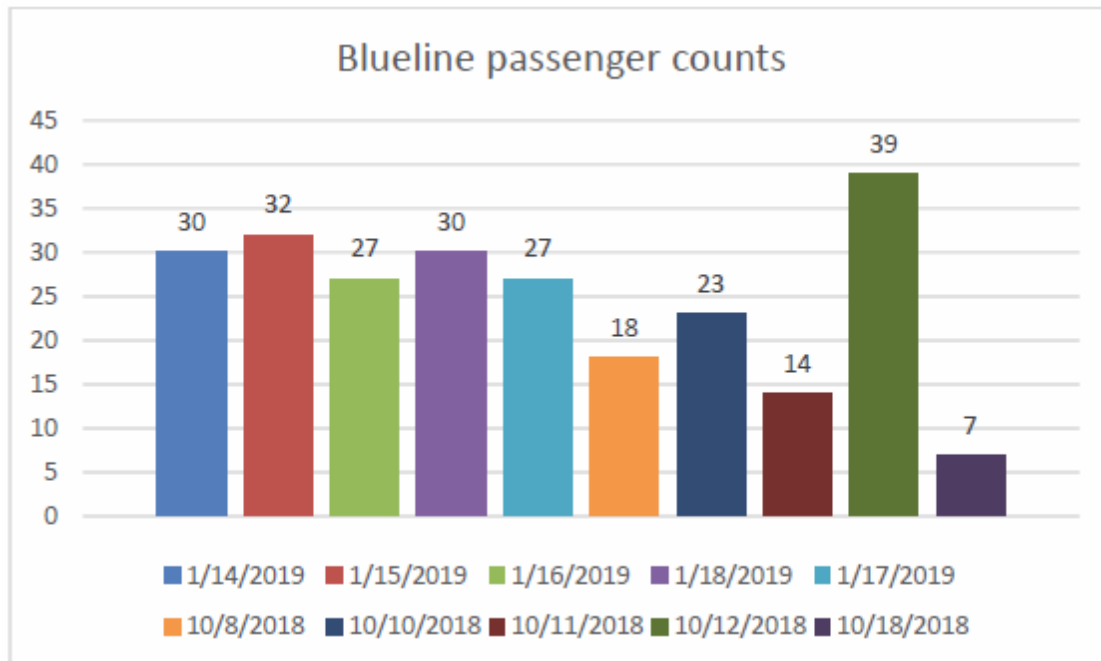


Figure 24: BlueLine passenger counts

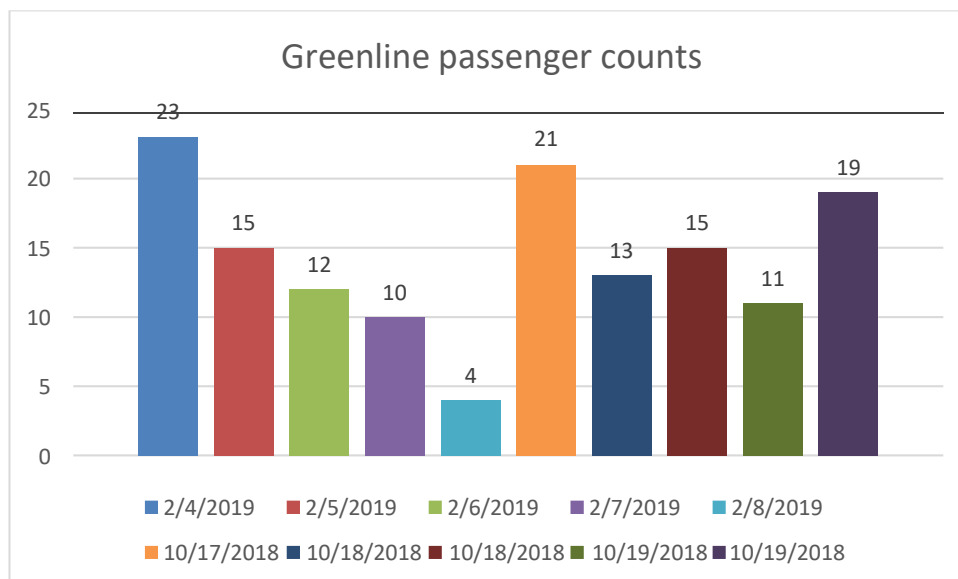


Figure 25: Greenline passenger counts

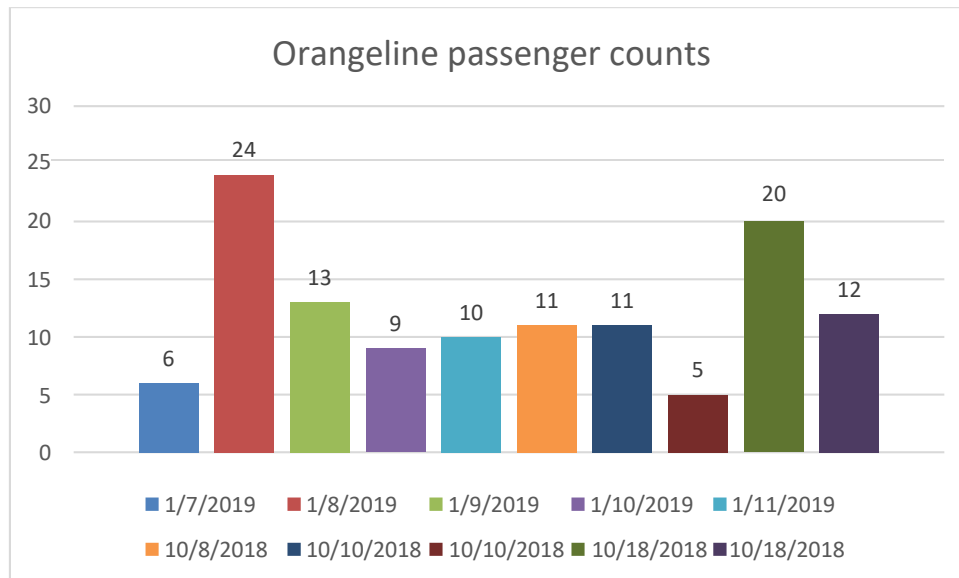


Figure 26: Orangeline passenger counts

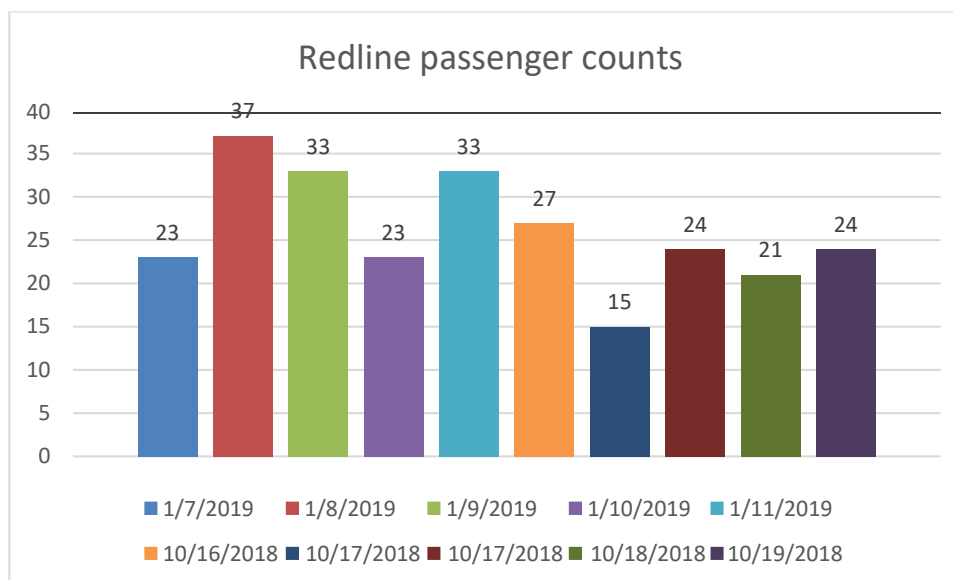


Figure 27: Redline passenger counts

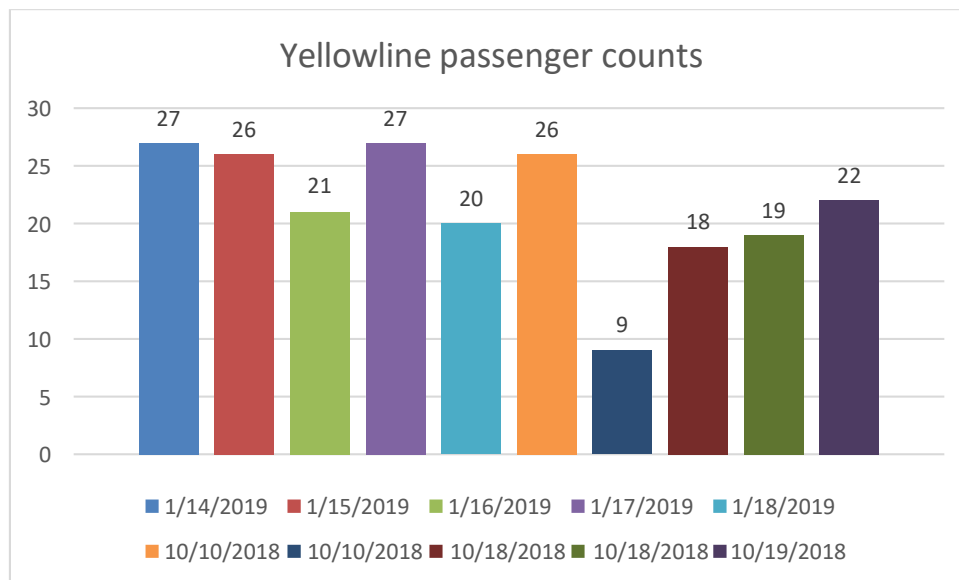


Figure 28: Yellowline passenger counts

The average numbers of passengers are 14 for the Greenline, 12 for the Orangeline, 26 for the Redline and 22 for the Yellowline. Consistently, the blue, red and yellow lines present the greater number of passengers. The standard deviations are 5.6 for the Greenline, 5.9 for the Orangeline, 6.6 for the Redline and 5.6 for the Yellowline. Standard deviations do not vary greatly for these lines in magnitude. While passenger demand varies substantially for different lines, the variances remain relatively constant for each line.

Regarding the origin and destination information from the manual counts, no different patterns were obtained from the manual counts performed during the pilot study. The number of passengers who boarded and alighted the bus is shown in Appendix B in Figures 83-87. These results match the boarding and alighting patterns shown for the survey data.

3.4.2. Travel Time and Stopped Time

The travel times between bus stops and the time a bus was parked at a bus stop (hereinafter referred to as stopped times) were recorded by a surveyor onboard. The surveyor utilized a digital chronometer. The time between when a bus leaves a bus stop and when the bus arrives at the following bus stop represents the travel time. The time between when a bus arrives at a bus stop until the bus starts to move again represents the stopped time. These data were collected for every bus stop and for all the bus lines during the months of October 2018 and January 2019, at the same time the manual counts of passengers were being performed.

The travel times in seconds at each bus stop for each day of data collection, and average and standard deviation for the Blue line are shown in Table 12. The stopped times in seconds with statistics are shown in Table 13. The respective tables for the Greenline are Tables 14 and 15. The tables showing these data for the Orangeline are Table 16 and 17. The tables that show these statistics for the Redline are Table 18 and 19. Tables 20 and 21 show these data for the Yellowline.

In the tables below, it can be noted that some cells do not have a numeric value, but a dash. In these cases, either a mistake was made by the surveyor who missed a bus stop or there was a

mistake made when writing the observation, especially for the travel times tables. For the case of the stopped times tables, there were many occasions when the time could not be reported because the bus did not make a stop. The average is obtained for the ten observations when possible. The datum that is not applicable is represented by the symbol: #N/A. This happened when there were not enough points to compute the statistics.

Table 12: Travel times recorded from the manual collection in the Blueline

Stop 1	Stop 2	10/8/2018	10/10/2018	10/11/2018	10/12/2018	10/18/2018	1/14/2019	1/15/2019	1/16/2019	1/18/2019	1/17/2019	Average	Standard Deviation
1	2	81.02	71.89	91.23	81.49	85.9	97.71	85.4	85.92	81.13	93.62	85.531	7.369
2	3	49.73	60.06	49.02	54.28	56.45	49.58	54.19	53.3	53.23	54.88	53.472	3.407
3	4	58.93	98.57	128.4	202.96	47.7	75.25	74.86	60.03	57.05	63.21	86.696	47.260
4	5	45.03	47.76	45.53	16.33	41.5	41.34	45.19	45.98	41.73	47.54	41.793	9.254
5	6	44.1	45.1	36.45	44.73	34.15	43.14	41.31	48.03	37.4	36.52	41.093	4.655
6	7	119.04	127.85	122.14	137.47	131.37	172.772	162.84	82.52	139.94	148.23	134.417	25.039
7	8	35.16	18.23	35.8	137.47	43.1	32.18	31.46	43.49	34.82	33.22	44.493	33.404
8	9	65.6	-	40.63	93.12	80.54	58.48	97.18	145.65	108.08	102.3475	87.959	31.056
9	10	124.39	-	116.7	126.44	134.7	138.32	120.05	108.77	117.18	121.08	123.070	9.177
10	11	-	43.67	39.2	45.58	43.22	40.9	44.0375	39.46	44.83	50.96	43.540	3.600
11	12	-	31.3	40.8	30.55	42.28	39.15	28.59	34.2	35.52	31.03	34.824	4.935
12	13	119.57	82.33	98.6	111.57	84.47	85.64	129.97	102.29	145.55	94.27	105.426	21.069
13	14	114.35	148.82	152.7	160.77	255.15	86.64	179.85	161.08	90.72	88.95	143.903	52.015
14	15	36.52	34.25	37.13	39.17	54.9	35.18	55.3	53.25	62.69	43.39	45.178	10.367
15	16	73.98	74.2	91.41	86.7	76.97	124.06	84.19	111.96	134.41	70.23	92.811	22.728
16	17	73.27	196.34	147.34	148.87	129.33	121.75	209.27	188.13	88.55	130.18	143.303	44.666
17	18	72.23	65.12	91.99	169.17	91.51	70.28	118.93	68.33	95.42	86.25	92.923	31.380
18	19	99.69	78.67	81.6	92.9	88.14	90.65	75.63	87.94	91	72.16	85.838	8.576
19	20	32.9	34.61	39.01	36.25	32.72	42.62	30.25	32.78	34.6	33.1	34.884	3.599
20	21	49.96	55.7	63.62	54.5	50	60.77	49.84	55.5	56.19	59.09	55.517	4.727
21	22	-	-	172.44	178	266.94	204.27	181.54	175.64	178.92	249.85	200.950	37.031
22	23	-	-	55.33	41.58	40.91	48.02	37.34	42.13	42.64	37.23	43.148	5.972
23	24	-	-	31.26	27.93	25.19	33.23	34.4	38.4	34.8	34.46	32.459	4.208
24	25	49.8	54.03	66.66	54.92	59.8	62.7	51.53	53.45	57.45	56.9	56.724	5.171
25	26	201.84	205.64	107.07	170.25	71.07	168.46	165.41	84.54	121.52	171.91	146.771	47.512
26	27	20.48	26.73	21.43	23.98	17.91	22.32	21.73	23.8	22.7	21.12	22.220	2.354

Table 12 (continued)

Stop 1	Stop 2	10/8/2018	10/10/2018	10/11/2018	10/12/2018	10/18/2018	1/14/2019	1/15/2019	1/16/2019	1/18/2019	1/17/2019	Average	Standard deviation
27	28	69.9	65.34	74.49	77.7	68.93	59.86	52.75	56.7025	60.04	54.16	63.987	8.616
28	29	62.5	67.36	51.57	76.63	52.68	46.6	52.26	51.92	76.79	50.66	58.897	11.172
29	30	31.51	41.97	36.56	42.9	38.51	36.81	32.4	38.81	38.46	35.03	37.296	3.678
30	31	38.93	4.4	45.2	43.9	36.38	88.13	64.87	39.93	50.11	37.69	44.954	21.398
31	32	31.66	-	20.38	22.41	21.76	24.54	21.44	24.73	21.35	102.55	32.313	26.556
32	33	85.65	-	107.07	69.72	84.66	77.33	81.82	73.97	128.16	113.02	91.267	20.019
33	34	39.42	28.75	35.55	38.89	36.3	33.8	24.92	27.91	31.14	35.35	33.203	4.851
34	35	127.06	116.79	193.61	390.57	96.24	111.59	91.95	79.01	146.19	132.9	148.591	90.962
35	36	51.26	43.73	54.85	52.5	51.92	55.46	43.63	45.65	52.45	51.04	50.249	4.348
36	37	65.17	54.38	60.9	61.7	73.04	68.48	53.01	63.78	60.9	57.37	61.873	6.144
37	38	67.7	71	80.49	85.15	86.85	113.59	69.39	84.95	87.6	83.6	83.032	13.124

Table 13: Stopped times recorded from the manual collection in the Blueline

Stops	10/8/2018	10/10/2018	10/11/2018	10/12/2018	10/18/2018	1/14/2019	1/15/2019	1/16/2019	1/18/2019	1/19/2019	Average	Standard deviation
1	489.36	327.71	60.53	-	-	-	-	-	-	-	292.533	216.568
2	-	-	-	-	-	-	-	-	-	-	#N/A	#N/A
3	-	-	-	-	20.36	-	-	-	-	-	20.360	#N/A
4	9.87	-	10.56	16.33	-	-	-	-	-	-	12.253	3.547
5	10.07	-	-	12.54	-	15.65	-	-	18.46	23.8	16.104	5.341
6	14.73	18.07	8.8	10.4	-	-	-	29.4	19.63	-	16.838	7.451
7	-	-	-	-	-	-	-	-	-	-	#N/A	#N/A
8	56.07	23.39	6.46	23	55.1	33.81	25.58	97.14	31	54.27	40.582	25.752
9	9.72	-	10.38	-	-	15.43	19.8	-	-	17.615	14.589	4.428
10	-	-	-	14.62	18.3	18.13	-	36.27	20.05	16.21	20.597	7.903
11	-	13.34	-	12.83	-	27.1	25.86	-	24.37	26.1	21.599	6.655
12	37.53	-	-	-	-	38.06	308.55	-	-	88.75	118.223	129.139

Table 13 (continued)

Stops	10/8/2018	10/10/2018	10/11/2018	10/12/2018	10/18/2018	1/14/2019	1/15/2019	1/16/2019	1/18/2019	1/19/2019	Average	Standard deviation
13	61.35	228.97	123.76	47.63	127.03	50.21	33.6	80.76	46.67	54.91	85.489	59.809
14	-	-	-	-	11.27	-	-	-	23.03	27	20.433	8.180
15	10.76	-	-	-	-	-	119.28	18.74	-	-	49.593	60.482
16	143.75	131.12	54.87	31.73	160.29	49.36	56.13	11.17	70.69	22.91	73.202	52.931
17	-	-	-	14.65	-	-	18	-	-	-	16.325	2.369
18	-	-	-	-	-	-	-	-	-	-	#N/A	#N/A
19	-	11.1	9.53	12.58	-	20.33	-	15.81	27.33	-	16.113	6.703
20	11.08	8.33	37.4	16.9	-	23.31	25	-	-	30.16	21.740	10.377
21	-	57.49	94.18	12.67	-	-	119.84	37.9	22.36	30.93	53.624	39.710
22	176.03	-	-	-	-	-	-	-	-	-	176.030	#N/A
23	-	-	-	-	-	-	-	-	-	16.5	16.500	#N/A
24	-	-	-	-	-	-	-	23.57	-	-	23.570	#N/A
25	-	-	-	-	14.66	-	43.13	-	75.02	20.49	38.325	27.372
26	-	9.27	-	12.2	-	15.89	17.61	-	-	21.86	15.366	4.865
27	-	13.82	-	-	-	-	-	29.26	29.26	-	24.113	8.914
28	12.23	9.32	-	20.87	14.59	-	40.28	29.59	-	18.9	20.826	10.833
29	-	21.18	-	11.4	12.92	16.5	22.83	-	51.77	20.81	22.487	13.616
30	18.66	8.58	-	11.22	-	45.25	-	14.02	29.56	-	21.215	13.894
31	-	22.83	13.36	-	10.85	-	-	-	21.43	16.62	17.018	5.119
32	-	-	-	-	-	-	-	-	-	-	#N/A	#N/A
33	17.28	-	-	24.34	-	-	15.74	-	29.05	16.27	20.536	5.890
34	58.44	22.65	57.88	55.58	20.76	77.37	110.25	68.65	30.85	43.33	54.576	27.277
35	9.34	-	15.05	-	-	-	-	-	16.06	-	13.483	3.624
36	-	-	-	-	-	21.6	12.8	-	22.03	-	18.810	5.209
37	-	-	-	14.9	-	-	-	-	-	-	14.900	#N/A
38	489.36	327.71	60.53	-	-	-	-	-	-	-	292.533	216.568

Table 14: Travel times recorded from the manual collection in the Greenline

Stop 1	Stop 2	10/17/2018	10/18/2018	10/18/2018	10/19/2019	10/19/2019	2/4/2019	2/5/2019	2/6/2019	2/7/2019	2/8/2019	Average	Standard deviation
1	2	150.52	270.63	211.56	149.95	161.14	169.45	168.56	157.67	308.82	122.38	187.068	59.155
2	3	85.11	84.97	67.55	77.42	-	88.8	100.46	90.13	97.67	74.68	85.199	10.656
3	4	119.12	107.93	96.36	124.71	78.98	137.6	108.89	108.26	120.48	135.94	113.827	17.809
4	5	36.81	37.85	39.63	41.85	46.97	43.14	48.76	42.6	38.7	34.7	41.101	4.447
5	6	52.69	155.91	157.79	148.07	132.1	163.23	78.59	92.58	122.2	98.17	120.133	38.005
6	7	251.33	173.02	228.9	-	171.59	264.8	405.74	219.64	198.34	224.12	237.498	70.577
7	8	44.1	42.42	53.87	-	49.62	49.82	64.83	54.18	45.41	95.1	55.483	16.314
8	9	25.22	27.4	30.56	-	28.23	28.94	36.1	27.7	24.53	24.26	28.104	3.661
9	10	118.97	-	99.85	-	154.87	98.35	44.2	41.34	34.94	32.63	78.144	46.102
10	11	93.94	125.87	114.43	-	106.6	99.3	124.39	139.74	122.75	75.26	111.364	19.684
11	12	226.57	208.63	197.4	223.67	194.28	201.74	263.53	211.73	192.37	232.21	215.213	21.955
12	13	53.25	-	114.6	77.04	127.1	135	86.85	116.13	74.02	98.85	98.093	27.299
13	14	1587.6	745.06	781.45	-	805.67	618.47	834.98	668.87	626.17	577.63	805.100	307.106
14	15	982.05	849.41	940.47	785.79	739.6	746.25	862.29	591.7	610.49	649.01	775.706	134.219
15	16	164.39	109.84	99.33	167.65	147.81	213.08	225.95	165.57	137.33	84.35	151.530	46.117
16	17	209.03	201.74	258.33	201.65	228.05	243.16	307.32	198.62	217.92	242.03	230.785	33.779
17	18	120.04	154.22	112.52	93.08	96.89	102.99	129.61	95.79	89.79	117.93	111.286	20.031
18	19	26.65	25.35	30.7	31.13	31.89	31.58	161.9	33.43	32.16	28.01	43.280	41.760
19	20	24.51	22.95	27.18	26.01	26.34	27.83	31.68	26.16	25.39	25.65	26.370	2.304
20	21	46.4	56.38	57.09	56.69	59.2	53.09	57.33	63.1	49.51	38.98	53.777	7.049
21	22	240.19	251.42	241.66	206.12	342.99	170.62	363.27	291.36	267.85	293.51	266.899	58.631
22	23	219.03	142.27	164.34	-	187.19	213.98	224.8	242.75	262.81	197.81	206.109	37.721
23	24	134.21	45.58	49.71	269.87	50.11	48.93	137.88	51.71	49.16	89.04	92.620	71.911
24	25	28.26	28.52	26.83	23.57	33.73	28.82	35.81	25.42	24.78	27.25	28.299	3.837
25	26	197.11	-	185.05	226.03	209.16	172.94	217.11	181.8	194.15	165.64	194.332	20.234

Table 15: Stopped times from the manual collection in the Greenline

Stops	10/17/2018	10/18/2018	10/18/2018	10/19/2018	10/19/2018	2/4/2018	2/5/2019	2/6/2019	2/7/2019	2/8/2019	Average	Standard deviation
1	-	31.03	47.64	25.7	36.64	-	-	-	-	-	35.253	9.389
2	-	14.17	-	-	-	-	-	-	-	-	14.170	#N/A
3	-	-	-	33.06	-	18.26	18.83	18.98	-	-	22.283	7.192
4	-	17.85	-	-	-	22.72	21.66	-	24.66	-	21.723	2.865
5	-	-	-	-	-	-	-	-	-	-	#N/A	#N/A
6	41.15	25.7	19.94	25.34	48.84	34.74	27.98	41.86	23.46	57.93	34.694	12.429
7	-	-	16.56	-	-	14.24	42.7	11.11	24.22	6.46	19.215	12.937
8	-	-	13.05	-	-	-	-	-	-	-	13.050	#N/A
9	10.74	13	15.53	-	12.13	19.09	143.97	26.03	15.76	26.34	31.399	42.590
10	-	-	-	-	-	-	-	-	-	52.62	52.620	#N/A
11	-	14.1	-	-	-	-	-	-	15.73	-	14.915	1.153
12	-	53.25	43.1	21.68	24.93	51.63	42.9	22.97	-	15.83	34.536	14.769
13	-	-	-	-	-	-	-	-	10.39	28.76	19.575	12.990
14	42.52	44.17	74.63	60.12	30.13	80.85	41.27	38.27	62	15.98	48.994	20.140
15	-	-	-	-	-	22.72	-	-	-	-	22.720	#N/A
16	14.47	13.75	22.41	16.2	36.76	-	-	-	-	-	20.718	9.594
17	65.43	-	-	-	-	-	-	-	-	-	65.430	#N/A
18	-	-	-	-	-	-	-	-	-	-	#N/A	#N/A
19	-	-	-	-	-	-	-	25.21	-	-	25.210	#N/A
20	-	-	-	-	-	-	-	-	-	-	#N/A	#N/A
21	-	-	-	-	-	21.78	54.09	18.03	81.98	34.8	42.136	26.362
22	49.24	22.76	44.56	17.64	13.79	22.76	15.33	24.71	7.04	9.53	22.736	14.001
23	-	-	-	-	-	-	-	-	-	-	#N/A	#N/A
24	-	28.26	-	-	28.38	33.77	-	-	-	-	30.137	3.147
25	-	-	-	-	27.05	-	-	-	-	-	27.050	#N/A
26	-	31.03	47.64	25.7	36.64	-	-	-	-	-	35.253	9.389

Table 16: Travel times recorded from the manual collection in the Orangeline

Stop 1	Stop 2	10/8/2018	10/10/2018	10/10/2018	10/18/2018	10/18/2018	1/7/2019	1/8/2019	1/9/2019	1/10/2019	1/11/2019	Average	Standard deviation
1	2	141.1	109.88	110.9	120.97	105.86	97.9	64.88	116.35	114.2	100.45	108.249	19.454
2	3	68.24	84.18	92.89	64.22	71.61	117.7	100.72	65.57	79.08	76.42	82.063	17.228
3	4	45.07	47.42	49.14	48.37	45.65	49.35	36.53	42.6	43.15	35.93	44.321	4.857
4	5	64.73	40.36	46.09	64.3	64.91	72.93	38.54	43.1	46.48	60.76	54.220	12.502
5	6	-	73.78	119.13	153.53	148.54	164.28	160.33	142.98	91.2	148.61	133.598	31.967
6	7	166.74	135.6	136.75	149.8	160.26	203.18	142.84	123.58	121.96	163.82	150.453	24.315
7	8	58.94	135.6	73.85	34.05	38.33	46.22	102.66	45.6	91.2	36.73	66.318	33.968
8	9	75.57	71.31	74.94	56.15	63.57	85.78	78.13	48.62	49.19	51.99	65.525	13.415
9	10	131.7	87.23	101.07	90.65	90.12	86.59	85.47	95.58	96.4	116.97	98.178	15.027
10	11	-	62.19	55.47	54.63	58.28	72.57	61.74	53.96	53.43	45.91	57.576	7.445
11	12	-	122.97	95.57	54.23	81.5	71.68	79.3	70.43	66.21	70.79	79.187	19.955
12	13	-	-	-	-	-	34.81	27.43	56.16	30.15	25.52	34.815	12.433
13	14	94.68	80.4	104.49	111.14	90.76	108.02	83.22	100.16	87.2	71.51	93.158	12.890
14	15	52.84	39	54.74	40.03	38.2	44.3	30.36	38.48	37.63	38.83	41.441	7.351
15	16	44.4	50.12	48.6	41.9	49.25	62.93	37.2	43.48	45.48	37.49	46.085	7.429
16	17	51.29	-	53.44	45.52	46.08	48.46	36.27	48.74	41.13	36.72	45.294	6.095
17	18	20.3	-	14.25	25.33	23.596	25.55	23.9	22.28	24.22	24.11	22.615	3.515
18	19	123.37	74.1	102.31	101.6	109.03	87.67	93.21	74.94	88.72	56.15	91.110	19.414
19	20	89.42	94.82	85.53	80.12	78.66	91.32	85.1675	89.03	87.21	73.11	85.439	6.505

Table 17: Stopped times recorded from the manual collection in the Orangeline

Stops	10/18/2018	10/10/2018	10/10/2018	10/18/2018	10/18/2018	1/7/2019	1/8/2019	1/9/2019	1/10/2019	1/11/2019	Average	Standard deviation
1	11.58	315	28.25	94.18	488.38	-	-	-	-	-	187.478	207.245
2	5.29	-	-	15.05	8.3	-	-	32.86	-	-	15.375	12.350
3	-	-	-	-	-	-	-	-	-	-	#N/A	#N/A
4	-	11.34	-	15.95	-	-	-	-	-	-	13.645	3.260
5	25.95	27.98	-	10.17	9.26	16.06	-	25.8	-	-	19.203	8.443
6	1.29	8.38	11.43	36.98	16.16	29.71	34.91	23.49	44.63	21.36	22.834	13.818
7	10.51	-	-	-	-	-	11.96	-	18.47	-	13.647	4.240
8	-	13.06	13.54	28.4	15.1	17.45	58.78	75	24	17.11	29.160	22.330
9	-	-	-	-	10.61	-	-	38.16	-	-	24.385	19.481
10	-	25.96	18.45	30.2	13.89	16.12	11.48	23.75	-	25.96	20.726	6.679
11	-	20.91	-	-	21.25	-	-	24.75	-	-	22.303	2.126
12	25.23	-	23.72	18.51	16.21	-	38.12	-	-	122.82	40.768	40.915
13	-	-	-	-	-	-	15.1	24.46	-	-	19.780	6.619
14	12.22	-	9	-	-	-	-	-	-	16.33	12.517	3.674
15	16.42	-	12.32	17.31	-	17.24	-	-	-	14.73	15.604	2.110
16	-	-	-	-	-	-	-	-	-	-	#N/A	#N/A
17	-	-	-	-	-	-	-	-	-	-	#N/A	#N/A
18	-	13.07	-	9.26	-	-	12.25	-	16.41	21.43	14.484	4.643
19	-	-	-	-	-	-	-	-	-	-	#N/A	#N/A
20	11.58	315	28.25	94.18	488.38	-	-	-	-	-	187.478	207.245

Table 18: Travel times recorded from the manual collection in the Redline

Stop 1	Stop 2	10/16/2019	10/17/2018	10/17/2018	10/18/2018	10/19/2018	1/7/2019	1/8/2019	1/9/2019	1/10/2019	1/11/2019	Average	Standard deviation
1	2	95.99	83.12	96.93	105.05	106.58	78.75	77.93	93.45	74.59	120.4	93.279	14.798
2	3	93.34	98.18	81.9	118.09	86.18	79.76	77.47	70.7	70.73	91.19	86.754	14.334
3	4	82.85	86.99	69	97.9	76.3	85.74	61.32	73.81	89.01	94.33	81.725	11.494
4	5	49.97	47.9	45.23	45.41	49.43	41.75	41.56	52.69	46.81	46.18	46.693	3.504
5	6	35.05	31.89	26.88	32.73	45.05	38.7	34.37	40.36	31.18	42.4	35.861	5.642
6	7	79.71	34.11	89.9	42.55	62.11	60.54	86.37	37.91	34.63	43.54	57.137	21.802
7	8	45.48	60.4	41.23	65.73	46.27	42.2	50.75	47.16	49.5	40.37	48.909	8.283
8	9	63.68	27.36	68.22	93.4	61.94	89.01	56.88	54.2	101.67	91.16	70.752	22.857
9	10	75	90.68	65.04	83.34	66.67	58.27	75.13	132.79	86.81	135.82	86.955	26.901
10	11	30.83	33.29	31.29	37.12	34.53	29.73	32.09	32.01	42.58	30.49	33.396	3.888
11	12	74.1	166.55	143.05	77.61	79.3	41.01	79.67	58.44	54.34	94.88	86.895	39.295
12	13	40.91	34.79	35.66	37.13	45.86	33.96	37	36.23	31.51	33.16	36.621	4.133
13	14	35.91	21.52	80.72	21.87	31.93	45.78	36	21.8	18.6	20.13	33.426	18.890
14	15	-	80.54	106.47	80.35	92.11	113.07	85.12	98.35	65.35	152.71	97.119	25.417
15	16	-	41.6	41.17	38.55	57.23	32.63	34.71	35.76	63.2	75.15	46.667	14.904
16	17	173.27	77.5	112.29	128.87	82.65	177.07	178.53	75.36	113.45	160.84	127.983	42.066
17	18	122.24	108.35	92.66	112.01	139.11	112.38	106.37	119.67	113.72	92.07	111.858	13.815
18	19	143.73	145.45	111.863	147.34	114.34	137.22	120.34	177.26	112.07	140.28	134.989	20.708
19	20	77.58	69.49	48.23	67.13	50.33	102.34	49.16	51.34	57.48	54.64	62.772	17.080
20	21	-	18.74	18.87	23.1	20.6	17.01	19.01	18.65	20.53	18.46	19.441	1.747
21	22	-	74.2	67.08	79.81	76.36	69.62	183.79	77.62	70.3	77.37	86.239	36.828
22	23	47.08	72.81	39.66	43.53	120.67	37	42.78	56.74	40.14	39.52	53.993	25.790
23	24	28.96	34.7	31.62	37.69	34.92	32.85	28.6	30.26	28.66	28.36	31.662	3.269
24	25	22.96	29.69	28.18	30.33	38.47	24.9	23.42	24.94	22.86	21.06	26.681	5.169
25	26	33.26	41.45	42.7	44.35	51.21	36.73	39.17	37.58	35.6	35.4	39.745	5.339
26	27	64.93	62.7	42.4	61.87	54.72	47.21	59.39	77.53	43.91	103.71	61.837	18.196

Table 18 (continued)

Stop 1	Stop 2	10/16/2019	10/17/2018	10/17/2018	10/18/2018	10/19/2018	1/7/2019	1/8/2019	1/9/2019	1/10/2019	1/11/2019	Average	Standard deviation
27	28	60.68	70.13	94.32	61.6	69.47	48.86	60.37	54.98	59.22	54.44	63.407	12.641
28	29	-	131.09	180.6	105.84	129.8	83.21	75.23	56.95	97.28	154.72	112.747	39.801
29	30	-	49.41	42.12	39.4	56.24	46.65	47.96	47.39	38.35	42.48	45.556	5.595
30	31	38.55	38.33	37.5	39.93	51.46	37.53	41.3	37.48	43.43	35.03	40.054	4.631
31	32	98.01	105.46	99.12	108.87	102.64	103	132.24	98.39	98.87	112.9	105.950	10.461
32	33	128.49	108.95	132.69	137.02	86.28	116.35	112.72	85.94	108.26	93.46	111.016	18.389
33	34	22.67	43.12	29.25	36	38.32	37.48	35.6	38.52	29.33	39.12	34.941	6.065
34	35	190.44	220.62	93.9	126.42	164.53	99.55	132.94	115.42	163.49	156.59	146.390	40.365
35	36	39.79	40.77	33.81	42.4	36.68	39.35	37.85	36.5	32.81	19.85	35.981	6.405
36	37	69.44	98.07	76.15	84.2	73.15	64.06	68.31	66.46	76.44	88.23	76.451	10.780
37	38	79.47	80.62	73.78	-	90.13	80.03	82.89	77.31	70.84	93.84	80.990	7.282
38	39	81.53	96.54	-	-	86.42	66.29	75.87	70.8	83.25	68.52	78.653	10.244

Table 19: Stopped times recorded from the manual collection of the Redline

Stops	10/16/2019	10/17/2018	10/17/2018	10/18/2018	10/19/2018	1/7/2019	1/8/2019	1/9/2019	1/10/2019	1/11/2019	Average	Standard deviation
1	589.01	-	-	534.9	-	-	-	-	-	-	561.955	38.262
2	-	-	-	-	-	-	18.335	-	-	-	18.335	#N/A
3	16.44	-	-	10.23	45.26	13.13	20.54	-	-	13.46	19.843	12.933
4	17.01	9.67	-	-	-	-	-	-	-	-	13.340	5.190
5	-	-	-	-	-	-	-	27.53	-	22.65	25.090	3.451
6	13.8	-	-	-	15	10.5	-	-	-	-	13.100	2.330
7	15.4	16.41	28.68	14.73	16.54	-	16.93	14.4	33.92	26.16	20.352	7.250
8	-	-	-	10.03	-	-	-	12.33	26.71	45.93	23.750	16.526
9	16.23	-	12.5	13.83	25.25	30.52	14.16	33.35	14.05	46.68	22.952	11.874
10	-	-	-	-	-	-	-	20.9	-	-	20.900	#N/A
11	28.55	46.93	40.83	15.96	16.61	34.69	45.48	7.78	36.75	58.9	33.248	16.031

Table 19 (continued)

Stops	10/16/2019	10/17/2018	10/17/2018	10/18/2018	10/19/2018	1/7/2019	1/8/2019	1/9/2019	1/10/2019	1/11/2019	Average	Standard deviation
12	17.36	12.1	-	-	13.69	-	-	-	-	-	14.383	2.698
13	-	-	-	24.02	18.3	20.98	24.99	80.96	-	64.59	38.973	26.793
14	11.11	-	-	-	-	-	21.04	12.89	-	-	15.013	5.295
15	-	15.54	17.26	12.31	-	-	-	-	15.96	34.32	19.078	8.713
16	18.61	12.15	25.84	14.57	16.6	15.73	-	27.62	-	29.23	20.044	6.549
17	26.69	13.46	27.61	-	-	29.32	21.96	19.58	-	-	23.103	5.976
18	43.44	124.23	26.17	56.13	29.93	58.8	31.7	28.9	225.34	105.45	73.009	63.196
19	-	-	-	-	-	-	-	15.5	-	17.2	16.350	1.202
20	-	-	-	83.36	-	-	-	-	25.85	16.06	41.757	36.361
21	-	-	-	-	-	-	14.25	-	-	-	14.250	#N/A
22	-	15.95	-	-	-	22.77	14.42	13.3	24.88	-	18.264	5.217
23	-	-	21.53	15.83	-	22.56	-	-	23.5	15.1	19.704	3.940
24	-	-	-	-	14.06	-	-	12.85	-	-	13.455	0.856
25	-	13.51	9.76	12.4	23.16	16.31	-	-	-	-	15.028	5.117
26	-	-	17.58	14.51	28.27	36.35	12.10	-	30.05	-	23.143	9.755
27	-	11.04	21.06	29.93	16.24	21.82	-	18.78	-	-	19.812	6.305
28	4.83	-	-	-	-	26.03	23.96	40.73	18.73	46.64	26.820	15.127
29	-	-	18.51	-	12.66	-	12.81	11.21	-	87.31	28.500	32.994
30	-	-	-	-	-	-	11.86	12.1	-	-	11.980	0.170
31	-	-	9.04	-	10.76	-	-	-	-	-	9.900	1.216
32	16.11	-	-	-	-	15.12	-	48.1	27.31	17.98	24.924	13.828
33	-	20.46	-	-	20.93	-	13.32	-	29.23	20.19	20.826	5.646
34	32.5	-	-	27.92	-	-	-	16.78	-	14.94	23.035	8.526
35	12.41	-	-	13.59	-	17.33	-	-	-	35.14	19.618	10.559
36	-	10.56	-	-	-	-	23.36	-	-	-	16.960	9.051
37	-	12.19	-	-	-	-	-	12.89	21.78	16.46	15.830	4.385
38	-	10.86	-	-	-	-	-	-	-	-	10.860	#N/A
39	589.01			534.9							561.955	38.262

Table 20: Travel times recorded from the manual collection in the Yellowline

Stop 1	Stop 2	10/10/2018	10/10/2018	10/18/2018	10/18/2018	10/19/2018	1/14/2019	1/15/2019	1/16/2019	1/17/2019	1/18/2019	Average	Standard deviation
1	2	180.87	135.17	171.33	137.46	125.76	166.05	169.68	211.16	141.71	141.66	158.085	26.337
2	3	70.12	69.45	70.28	74.64	72.51	71.37	67.51	69.20	73.07	75.70	71.385	2.575
3	4	146.50	133.68	62.13	104.75	121.37	90.00	126.70	123.79	140.72	59.06	110.870	31.165
4	5	44.43	39.83	43.10	47.66	47.77	41.84	42.92	42.98	42.12	30.49	42.314	4.836
5	6	138.81	81.45	125.68	157.13	97.67	61.52	137.05	136.14	103.57	67.84	110.686	33.072
6	7	166.28	213.91	236.43	209.36	191.96	196.89	171.48	209.48	189.27	211.23	199.629	20.983
7	8	46.46	45.83	46.40	44.16	45.30	37.01	44.40	42.85	43.48	44.46	44.035	2.742
8	9	48.58	42.18	35.83	42.85	46.25	35.76	67.22	34.23	33.83	31.52	41.825	10.603
9	10	34.49	47.29	39.51	43.42	-	36.65	38.30	38.82	37.71	70.68	42.986	11.052
10	11	64.89	61.58	64.80	68.11	-	64.18	57.25	60.94	66.35	65.22	63.702	3.266
11	12	23.70	28.11	25.65	31.95	-	22.64	25.22	23.70	25.63	21.23	25.314	3.184
12	13	37.00	37.75	36.91	36.55	-	26.64	31.07	32.75	33.83	32.82	33.924	3.601
13	14	28.59	31.89	25.48	36.55	49.78	50.13	30.43	31.47	39.23	31.00	35.455	8.540
14	15	44.25	52.28	55.58	47.18	56.50	49.99	51.90	48.90	47.31	53.79	50.768	3.946
15	16	23.75	17.35	17.73	31.00	17.92	21.47	21.13	19.35	20.48	20.72	21.090	4.004
16	17	147.62	58.38	118.42	-	-	92.68	150.83	89.31	89.09	69.10	101.929	34.090
17	18	71.72	94.85	84.30	-	-	78.97	74.57	70.65	84.37	82.70	80.266	8.030
18	19	56.46	104.00	48.18	130.47	86.11	117.95	77.43	59.91	66.61	108.57	85.569	28.408
19	20	27.14	39.62	30.43	32.66	29.74	30.35	26.10	26.85	28.88	33.13	30.490	3.972
20	21	224.34	227.26	228.89	180.71	180.56	183.11	200.46	256.74	238.95	253.51	217.453	29.363

Table 21: Stopped times recorded from the manual collection in the Yellowline

Stops	10/10/2018	10/10/2018	10/18/2018	10/18/2018	10/19/2018	1/14/2019	1/15/2019	1/16/2019	1/17/2019	1/18/2019	Average	Standard deviation
1	57.29	60.13	40.20	71.21	67.99	-	-	-	-	-	59.364	12.113
2	16.16	-	23.43	-	23.76	-	32.26	25.85	-	15.45	22.818	6.296
3	-	11.06	11.03	-	15.14	30.97	20.06	33.17	-	-	20.238	9.772
4	22.39	12.78	17.48	-	12.06	21.42	30.26	31.75	18.23	29.54	21.768	7.404
5	-	19.27	-	8.85	-	-	-	-	-	-	14.060	7.368
6	29.35	19.46	28.10	15.85	60.80	58.73	53.89	77.03	27.63	29.69	40.053	20.729
7	-	-	21.36	-	1.07	-	24.60	36.14	20.38	15.76	19.885	11.489
8	10.14	-	14.71	7.42	13.80	36.92	30.33	24.06	16.85	17.10	19.037	9.627
9	6.51	-	11.78	12.65	-	28.04	20.18	16.64	15.36	-	15.880	6.859
10	-	-	-	-	27.02	-	33.93	34.83	18.03	19.53	26.668	7.827
11	15.95	-	8.70	9.62	-	14.88	18.78	20.73	22.67	-	15.904	5.319
12	7.70	-	-	9.23	8.07	-	18.80	20.94	-	-	12.948	6.389
13	-	-	-	-	16.75	-	29.82	-	20.35	46.27	28.298	13.189
14	20.47	9.20	10.10	-	-	17.71	22.27	22.34	18.23	23.60	17.990	5.536
15	7.50	-	-	-	-	32.83	-	-	21.58	17.87	19.945	10.454
16	8.92	-	-	-	-	16.35	21.26	-	38.53	19.03	20.818	10.939
17	-	-	-	14.00	-	-	27.01	-	-	-	20.503	9.202
18	31.94	-	-	22.97	-	20.73	-	-	19.22	21.51	23.274	5.030
19	13.83	-	8.50	8.23	-	17.36	21.59	-	19.60	23.85	16.137	6.176
20	-	-	-	-	-	18.31	-	-	-	17.06	17.685	0.884
21	57.29	60.13	40.20	71.21	67.99	-	-	-	-	-	59.364	12.113

3.5. Smart Station Data

Collecting data with the Smart Station is a key component of this research. These data were analyzed with statistical tools and algorithms to infer passenger ridership, OD matrices, wait times and travel times. The raw data collected by the SS consist of the creation of two datasets: the Wi-Fi data and the GPS data. The Wi-Fi data are Extensible Markup Language (XML) files, which is a technology that uses tags to mark and delineate pieces of datum (Huh, 2014). The GPS data are the National Marine Electronics Association (NMEA) files, which is a set of character strings that provides navigation system information (Varun, Singh, & Nagaraj, 2013).

These archives required a preprocessing step before they were fed into the algorithms. The Kismet Wireless software gathers a wealth of information and reports the output in XML format. The XML file consists of texts and descriptions of the data; therefore, it was converted into Excel files. In this process, more than 200 attributes were created, many of them being repetitive or unusable for analysis. The initial filter only selected fourteen attributes, these are: (1) Wireless Network ID, (2) First time of detection, (3) Last time of detection, (4) Type of wireless network, (5) MAC address, (6) Wireless network channel, (7) Average latitude, (8) Average longitude, (9) Manufacturer, (10) Maximum rate of data transfer, (11) Maximum speed, (12) Minimum speed, (13) Maximum signal strength, and (14) Minimum signal strength. These attributes were selected because they revealed useful information to estimate the parameters of interest of this research. Descriptions of these attributes are provided in Appendix C.

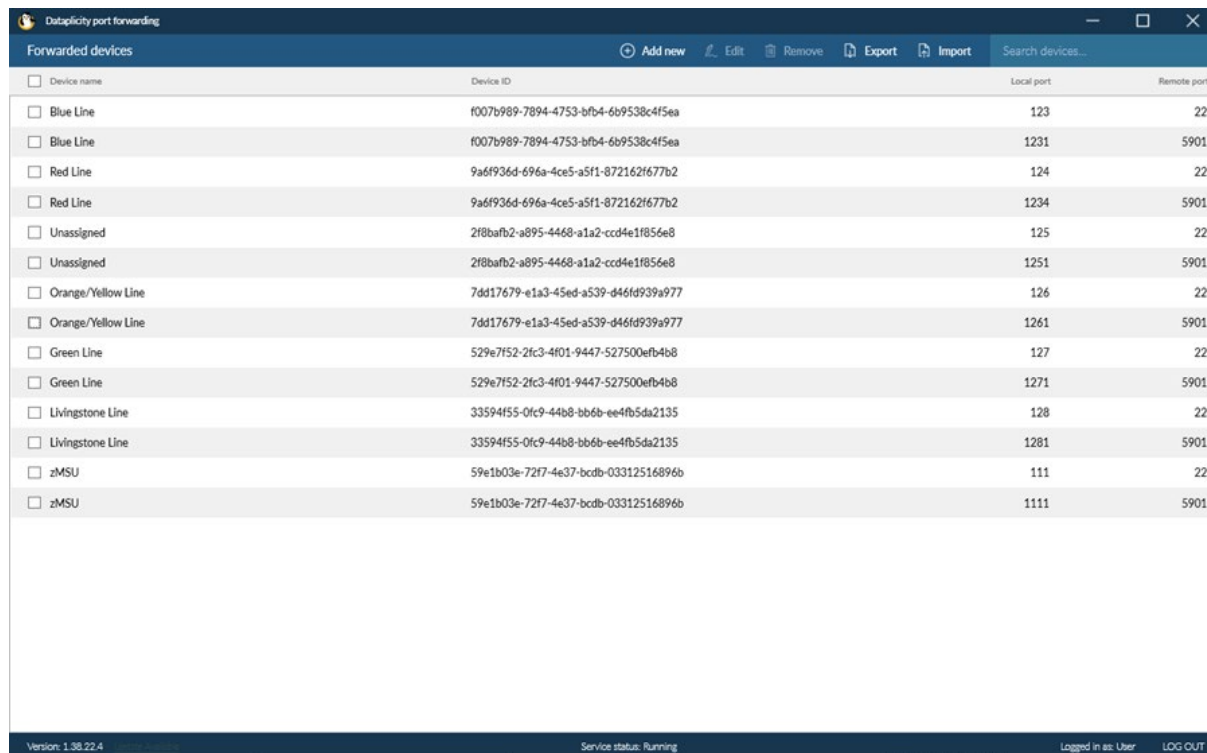
In the same way, the NMEA files were preprocessed and converted to Excel tables for analysis. The processed GPS data provides the following attributes: (1) Coordinated Universal Time (UTC), which references the local time at Greenwich, England, (2) Universal Transverse Mercator (UTM) Easting in Zone 12T, (3) UTM Northing in Zone 12T, (4) Distance traversed, and (5) Speed. The UTM Zone 12T was used because the city of Bozeman is located there. This section provides the summary statistics of the SS data collected during the pilot study in April 2018 and during the manual counts of October 2018 and January 2019.

3.5.1. Pilot Data

During the pilot data collection, Smart Stations were onboard the buses during the day. The batteries were replaced at the end of the day shift, recharged overnight and reinstalled back in the morning. In order to ensure that all the SS were functioning correctly, the Dataplicity Porthole program was used to monitor the devices. This software allows remote access to Raspberry Pi computers (Dataplicity, 2019). Figure 29 shows the interface of the remote control program. In addition, the SS were programmed to restart every 30 minutes to capture passengers who had alighted but later boarded the same bus on its returning run, because once a MAC address is detected by a SS, attributes like the sign in and sign out time are updated for that same MAC address.

From the pilot data, the researchers identified a number of issues. First, the complexity of preprocessing large amounts of information with Kismet Wireless prompted the team to convert the raw data to excel format for easier data management. Second, Wi-Fi data produced a time unit that is different from the one used by GPS data. To make time units consistent, the NMEA files were projected into the UTM coordinates. Third, the Raspberry Pi was prone to error upon reboots;

therefore, data were not collected until the time was synchronized. The time synchronizes within a period of one minute after reboot. Lastly, the team also made small but important modifications like the necessity of connecting Raspberry Pi to the Internet and extending the GPS antennas to provide accurate locations. These modifications helped reduce errors and enhanced data quality. The pilot data was not fed into the algorithms.



Device name	Device ID	Local port	Remote port
<input type="checkbox"/> Blue Line	f007b989-7894-4753-bfb4-6b9538c4f5ea	123	22
<input type="checkbox"/> Blue Line	f007b989-7894-4753-bfb4-6b9538c4f5ea	1231	5901
<input type="checkbox"/> Red Line	9a6f936d-696a-4ce5-a5f1-872162f77b2	124	22
<input type="checkbox"/> Red Line	9a6f936d-696a-4ce5-a5f1-872162f77b2	1234	5901
<input type="checkbox"/> Unassigned	2f8bafb2-a895-4468-a1a2-ccd4e1f856e8	125	22
<input type="checkbox"/> Unassigned	2f8bafb2-a895-4468-a1a2-ccd4e1f856e8	1251	5901
<input type="checkbox"/> Orange/Yellow Line	7dd17679-e1a3-45ed-a539-d46fd939a977	126	22
<input type="checkbox"/> Orange/Yellow Line	7dd17679-e1a3-45ed-a539-d46fd939a977	1261	5901
<input type="checkbox"/> Green Line	529e7f52-2fc3-4f01-9447-527500efb4b8	127	22
<input type="checkbox"/> Green Line	529e7f52-2fc3-4f01-9447-527500efb4b8	1271	5901
<input type="checkbox"/> Livingstone Line	33594f55-0fc9-44b8-bb6b-ee4fb5da2135	128	22
<input type="checkbox"/> Livingstone Line	33594f55-0fc9-44b8-bb6b-ee4fb5da2135	1281	5901
<input type="checkbox"/> zMSU	59e1b03e-72f7-4e37-bc8b-03312516896b	111	22
<input type="checkbox"/> zMSU	59e1b03e-72f7-4e37-bc8b-03312516896b	1111	5901

Figure 29: Interface of the remote-control program

3.5.2. Data Collection

This section refers to the Wi-Fi and GPS data collected with the Smart Stations during the months of October 2018 and January 2019. The manual counts of passengers and travel times were recorded as mentioned earlier in this chapter. The SS were turned on before the start of the loops. On this occasion, the SS were not monitored remotely, and the data were retrieved after the loops had ended and were transferred to a desktop computer.

3.5.2.1. Wi-Fi Data

The Wi-Fi data consist of the fourteen attributes that were obtained after preprocessing the raw data. Figure 30 shows the total number of networks scanned by the SS per loop. The following observations were not correct because there was a malfunction with the battery or the GPS receiver: Green 10/18/2018, Orange 10/18/2018 and Yellow 10/10/2018. These malfunctions were fully fixed when the data were collected in January 2019. As expected, the number of networks detected generally increased on the longer loops. It is important to note that, although the Greenline

is a longer trip, it travels through many areas that are less densely populated, which is why fewer networks are detected.

One outlier can be observed for the Orangeline, which is one of the loops obtained on 10/10/2018. This is because the SS was turned on for a longer duration of time due to the Yellowline and Orangeline sharing a bus, which extended the recording time on that specific day. Therefore, this observation will be useful in testing travel times but not for ridership counts.

Figure 31 shows the percentage distribution of the detected networks by type. Infrastructure networks represent 69% of all detected observations. Approximately one-third of networks are probes, which is characteristic of the phone probes. Therefore, all the passengers will be part of the probe category of this classification.

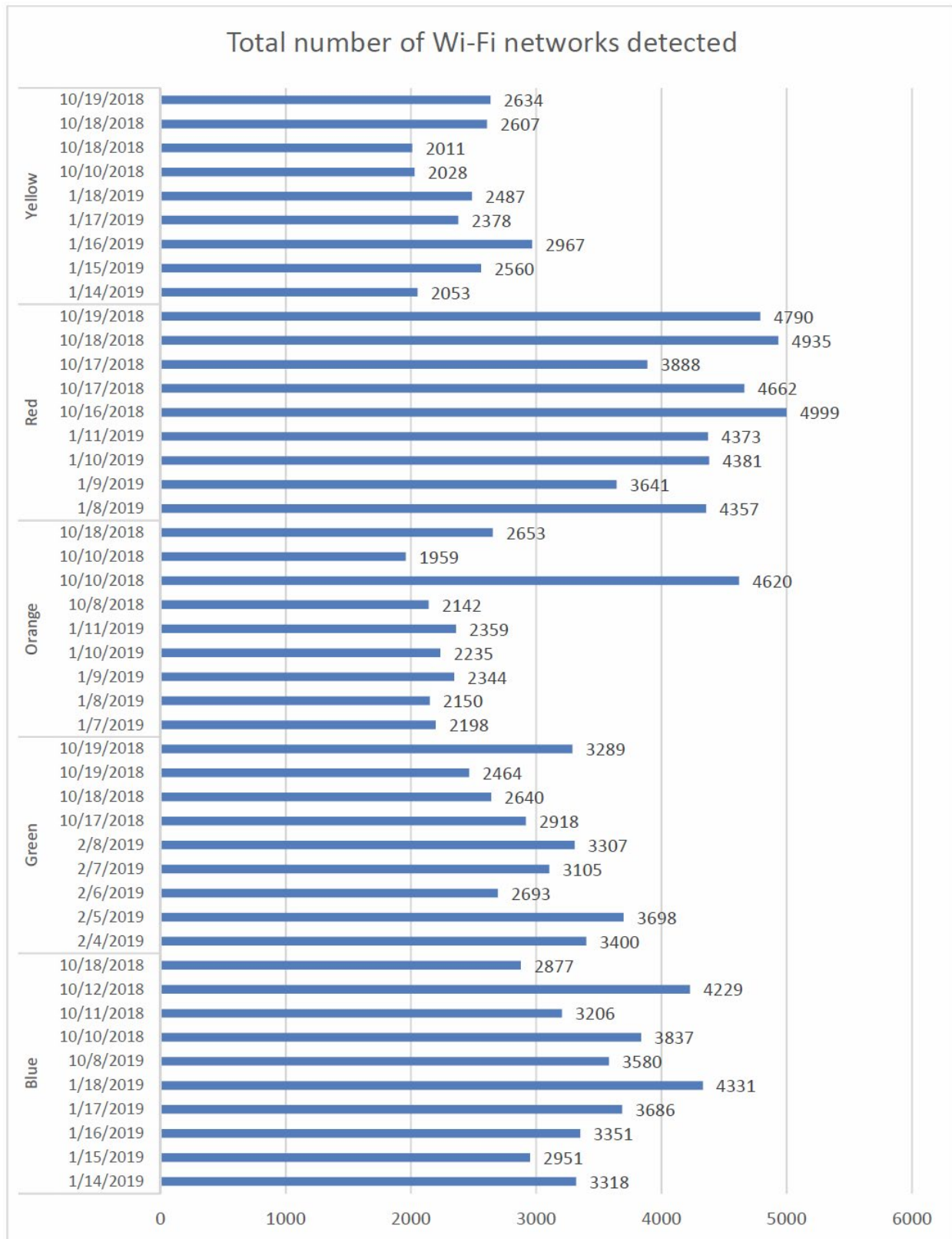


Figure 30: Total number of unique Wi-Fi networks detected

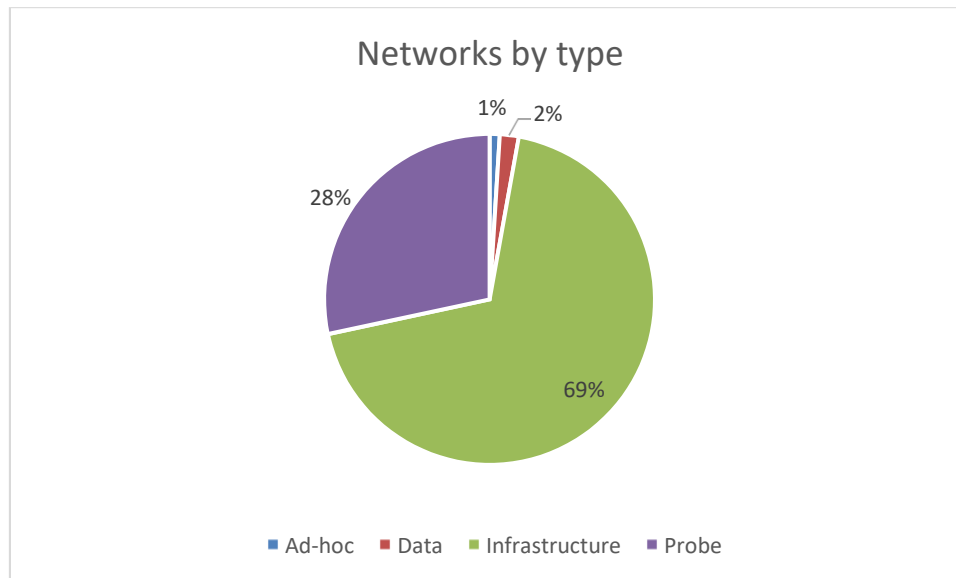


Figure 31: Percentages of networks by type

Figure 32 shows the percentage distribution of the networks by network channel, associated with the frequency at which the signals are being transmitted by the source devices. The phone probes use channel 0 (2407 MHz) to communicate. The most common channels are channel 0 with 33% of the networks; channels 1, 6 and 11 maintained 18% each.

Figure 33 shows the percentage distribution of the networks by the rate of data transfer. Out of the 151,539 networks detected, around 80% had a rate of data transfer of 1000 megabits per second (Mbps). The rate that mobile devices use is between 1000 and 6000 Mbps. Therefore, around 93% of the networks that were detected communicate in that rate range. The number of devices that were detected to transfer data at 0 and 9 Mbps was so low that it rounds to zero percent compared to the total.

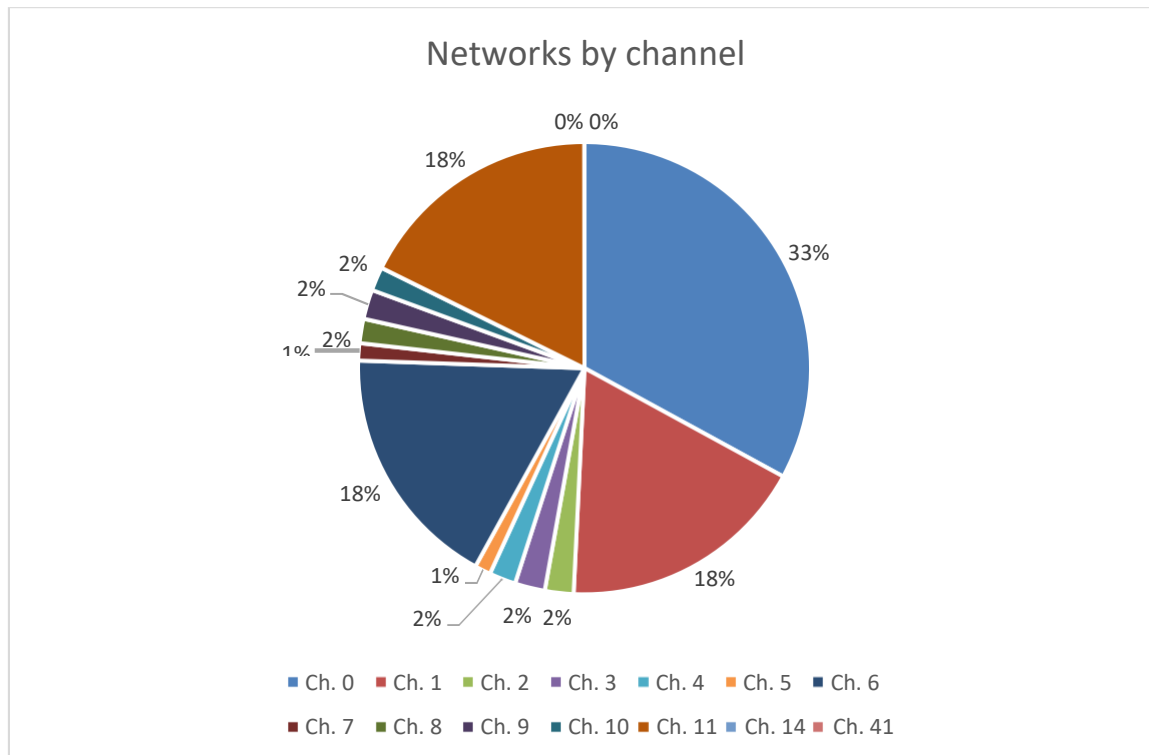


Figure 32: Percentages of networks by channel

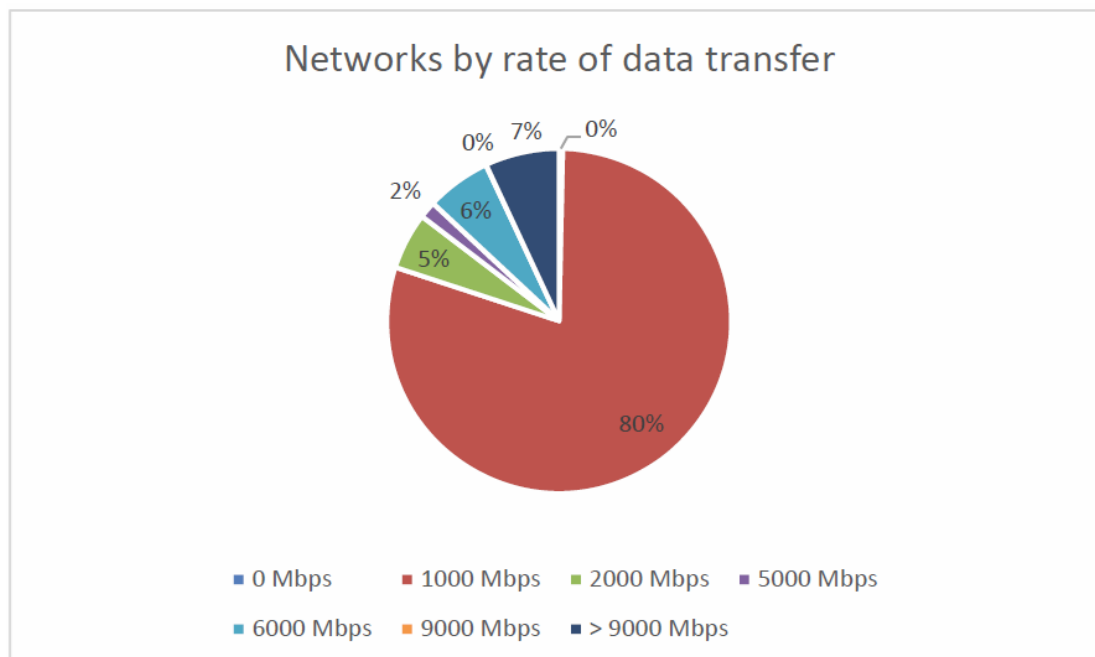


Figure 33: Percentages of networks by the rate of data transfer

Table 22 shows the summary statistics of the numeric variables: average latitude, average longitude, maximum speed, minimum speed, maximum signal strength, minimum signal strength and detection time. In some cases, the detection time would yield a negative value. This was because the Raspberry Pi can collect Wi-Fi signals before the time is synchronized, generating inconsistencies. The negative values of detection times were discarded, negating a total of 81 observations. The detection time is the difference between the last and first times of detection.

Table 22: Summary Statistics of the Wi-Fi data (no. of observations = 147,141)

Attributes	Average	Maximum	Minimum	Standard deviation
Average latitude	45.680	45.776	45.637	0.018
Average longitude	-111.064	-111.009	-111.186	0.038
Maximum speed (m/s)	9.235	75.845	0	5.562
Minimum speed (m/s)	6.710	34.383	0	5.903
Maximum signal strength (dBm)	-77.831	-17	-95	11.113
Minimum signal strength (dBm)	-84.210	-17	-97	10.082
Detection time (seconds)	368.048	8251	0	937.725

The table shows a total sample size of 147,141 networks, which is inferior to the value of 151,539 total networks detected. This was due to the GPS devices' failure to detect geographical locations because of an obstruction to the antenna. Figure 34 shows the histograms of the first six attributes shown in Table 22. The values deviate minimally and are within the normal range of movement for the buses. Figure 35 shows the histogram of the detection time.

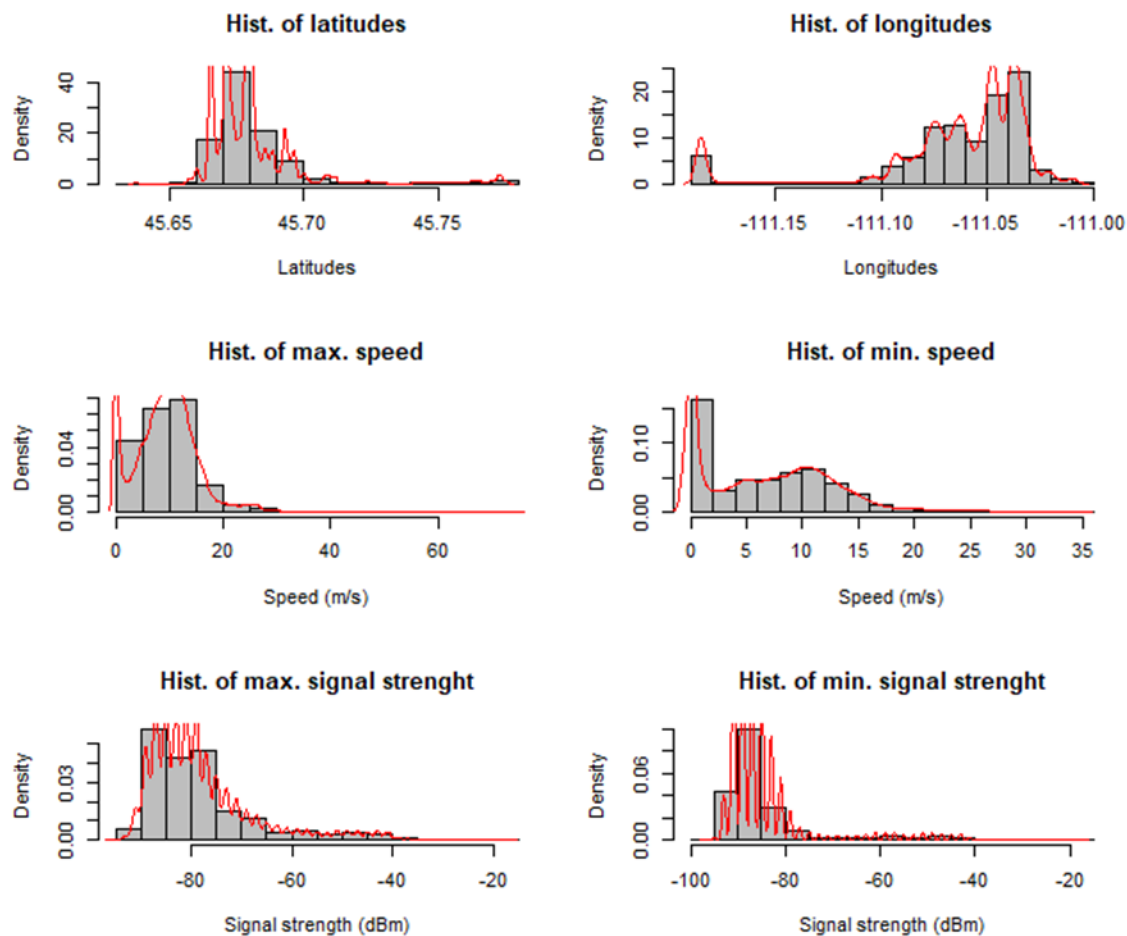


Figure 34: Histograms of the numeric variables of the Wi-Fi data

The detection time has more variability than the other variables, as shown in the summary statistics. Some signals have over 8,000 seconds of total detection, which is more than two hours. This happens because the Raspberry Pi takes some time to synchronize the hour after rebooting. However, when signals are immediately obtained, they are most likely permanent signals because the Smart Stations were started before boarding the buses.

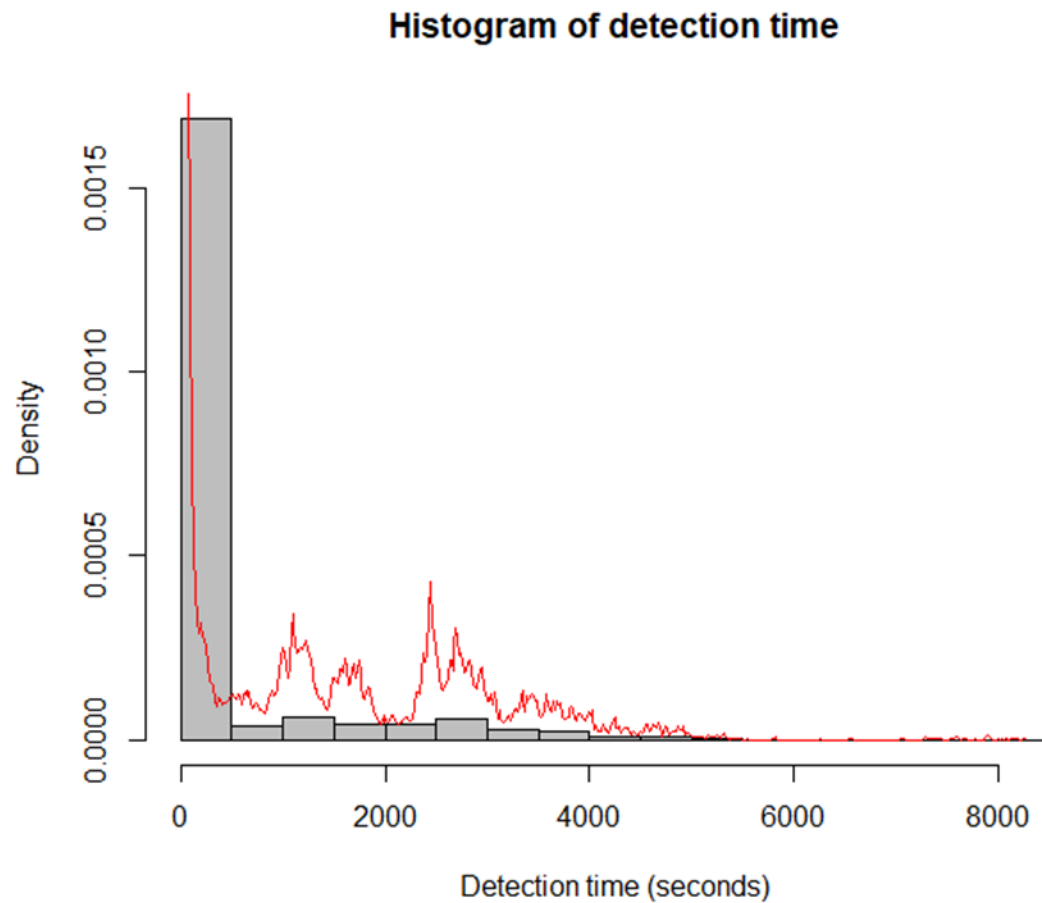


Figure 35: Histogram of the detection time of the Wi-Fi data

3.5.2.2. GPS Data

The GPS data logs the location of the Smart Stations every second in most cases. The same issues of logging GPS data as explained in the previous section may arise. Table 23 shows the summary statistics of the data obtained. The sample size differs from the Wi-Fi data because this is not the number of networks but locations of every time that was recorded. Figure 36 displays the histograms of the latitudes and longitudes, which show the same pattern as the average coordinates of the devices previously exposed. Figure 37 shows the histogram of the speed.

Table 23: Summary Statistics of the GPS data (no. of observations = 145,620)

Attributes	Average	Maximum	Minimum	Standard deviation
Latitude	-111.073	-111.009	-111.186	--
Longitude	45.68277	45.77604	45.63691	--
Speed (m/s)	7.152467	1724.13	0	8.048492

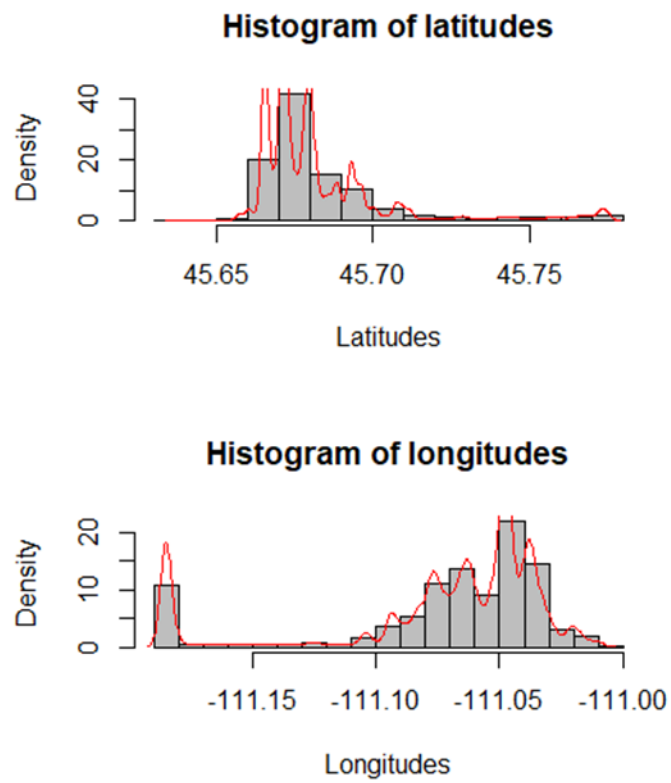


Figure 36: Histograms of the coordinates of the GPS data

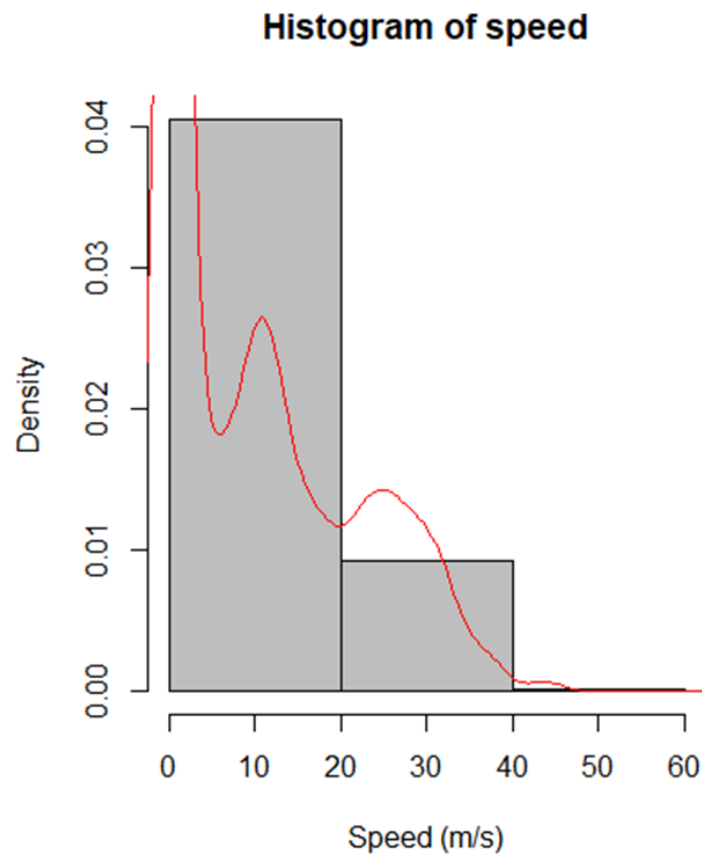


Figure 37: Histogram of the speed of the GPS data

4. METHODOLOGIES

This chapter provides a description of the procedure developed to analyze the data and the statistical tools. This study investigates the use of IoT technologies to estimate ridership, OD flow characteristics, wait time and arrival time, and therefore this chapter is divided in the same manner. Figure 38 displays the general approach of this study. Explanations of the algorithms implemented on the datasets and of the statistical tools will be provided for each characteristic that is being estimated.

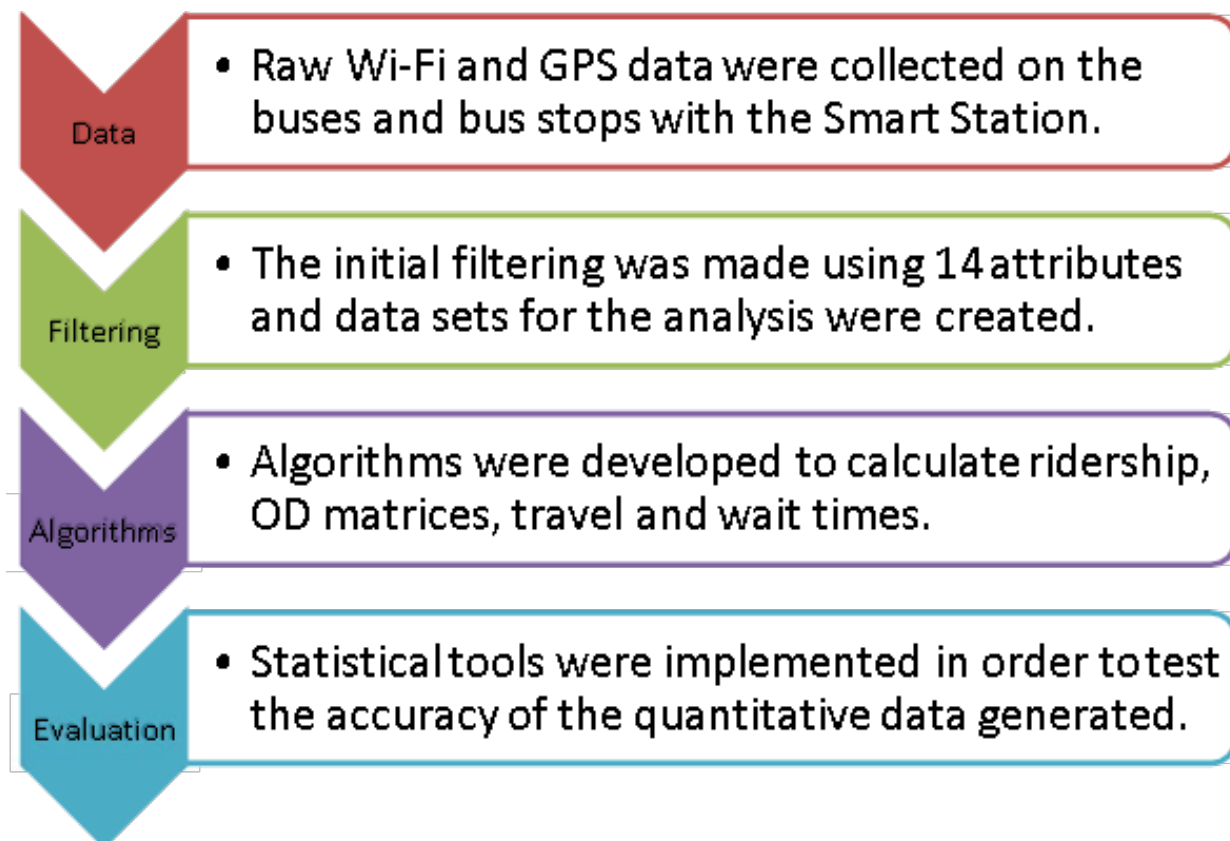


Figure 38: Methodological approach of the research

4.1. Ridership

Passenger counts of a system can be estimated by scanning the Wi-Fi networks of their mobile devices that are present in the transportation system on the premise that passengers' smartphones can be recognized. In an urban area, there can be thousands of devices sharing information through the Wi-Fi technology; therefore, there can be an overestimation of the passengers if other devices like smartphones used by pedestrians, other vehicles or devices in the buildings are not filtered out.

A novel filtering algorithm is designed to count discoverable devices that are within the detection range of the Smart Station. The algorithm is intended to filter out detections that would not belong

to passengers using the attributes provided by Kismet Wireless software. The algorithm will be referred to as the rule-based method. An enhancement of the rule-based method is also tested, and it provides the option of accounting for those devices that randomize their MAC addresses. Finally, a clustering approach is also tested in order to see if there are specific patterns in the data that could be used to detect passengers from the dataset using unsupervised machine learning. The details of the algorithms implemented, and the tools used to test the accuracy are presented below.

4.1.1. Rule-based Method

This method uses the attributes contained in the Wi-Fi datasets in order to filter out those devices that do not belong to passengers based on the characteristics of network type, network channel, the maximum rate of data transfer, maximum speed, minimum speed, and first and last times of detection. The difference between the last two characteristics is the total time of detection for a specific device. Table 24 shows the methodology used to filter out the networks that are not considered passengers.

Table 24: Rule-based Method

Rule	Description	Attribute(s) used
1	Passengers should be detected for a longer period as the bus is running. Therefore, a minimum detection time of 60 seconds is used to identify those networks that are detected sporadically. Additionally, those networks that are detected for the entire trip are either riding the entire time (like the bus driver's smartphone) or signals that are constantly being detected but represent networks that are on fixed points. Hence, networks that are detected for nearly the total period of the loop are discarded as well. Likewise, networks that were detected for a period longer than half a loop plus four minutes were discarded because they would be considered permanent signals.	First time of detection, Last time of detection
2	Network type needs to be a probe. This rule filtered out Wi-Fi routers.	Network type
3	Network channel needs to be zero. Mobile devices that passengers would be carrying primarily use channel 0 (2.407 GHz).	Network channel
4	Mobile devices can only transfer a maximum of 6 MB of data. Therefore, devices exceeding this limit were ruled out.	Maximum rate of data transfer
5	Since passengers would be on buses most of the time, the minimum speed detected should be zero. However, estimations do not always yield a zero value; therefore, it is proposed that the minimum speed should be less than five meters per second. If the minimum speed is too high, the device is not moving alongside the Smart Station and it is not a passenger.	Minimum speed

6	The maximum speed that would be registered by passengers is when they board and alight the buses because of the relative motion. However, buses would not develop high speed near bus stops because they prepare to stop when passengers board or alight. Therefore, the travel speed of fifteen meters per second in the urban setting is considered.	Maximum speed
---	--	---------------

Table 25: Enhanced Rule-based method

Rule	Description	Attribute(s) used
1	Network type needs to be a probe. This rule filtered out Wi-Fi routers.	Network type
2	Network channel needs to be zero. Mobile devices that passengers would be carrying were channel 0 (2.407 GHz).	Network channel
3	Wi-Fi routers contain an SSID; however, mobile devices do not. If an SSID was found, the device was ruled out because it would not represent a bus rider.	Service set identifier (SSID)
4	Only a specific range of device manufacturers would correspond to possible mobile devices. Only these manufacturers were considered: Apple, Google, LG Electronics, Motorola, Samsung Electronics, TCT Mobile and Unknown. The unknown devices were classified using the random forest algorithm.	Manufacturer
5	Mobile devices can only transfer a maximum of 6 MB of data. Therefore, devices exceeding this limit were ruled out.	Maximum rate of data transfer
6	Mobile devices must be first and last detected nearby a bus stop. These times had to be within two minutes of a bus stop. Otherwise, they would be discarded.	First time of detection, Last time of detection, SS GPS
7	Since the bus would have to stop in order to pick up a passenger, high minimum speed would indicate that the detected device was a passerby and not a bus rider. A minimum speed of 55 meters per second was used.	Minimum speed
8	A valid bus rider would certainly ride the bus for at least the time to the next bus stop. If the detection time was less than 5 seconds, the device was filtered out.	First time of detection, Last time of detection

4.1.2. Enhanced Rule-based Method

The rule-based method does not use any statistical tool to account for mobile randomization and the GPS data containing the location of the Smart Station. Aided by undergraduate research performed by Jeremy Tate, a senior majoring in Statistics at Montana State University, the initial rule-based approach was enhanced, and more rules were considered in this analysis. Table 25 shows the methodology used in the enhanced rule-based method.

The enhanced method consisted of five absolute rules (1 through 5) and three tunable rules (6 through 8). The tuned values were obtained by minimizing a cost function. The cost function consisted of the sum of squared errors (SSE) of the predicted and the estimated number of people at each bus stop plus the SSE of the predicted and estimated number of passengers per loop. The cost function is described by the following equation:

$$f = \sum_{i=1}^n (x_i - \hat{x}_i)^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (1)$$

where n is the number of observations. x_i is the value of the number of passengers boarding and alighting at each bus stop for the i^{th} observation, and \hat{x}_i denotes the estimated number of passengers boarding and alighting at each bus stop. y_i denotes the true ridership of the i^{th} loop, and \hat{y}_i denotes the estimated ridership. The optimization method has the following structure (Bihorel, 2018),

$$\min f(x) \quad (2)$$

$$l_i \leq x_i \leq h_i, i = 1, n \quad (3)$$

$$g_j(x) \geq 0, j = 0, nbineq \quad (4)$$

where f is the cost function, x is the vector of parameter estimates, l and h are vectors of lower and upper bounds for the parameters estimates, n is the number of parameters, and $nbineq$ the number of inequality constraints $g(x)$. The cost function was applied initially to the data collected in October for the Redline for training the model. These consist of five loops. Later, the algorithm was also implemented on all the Wi-Fi data, including the data that was used for training due to the small number of datasets.

Because in general, not all bus riders carry detectable devices, there are probably more passengers on any given route than those that are truly detected. Therefore, a Poisson regression model was implemented to correct the ridership estimate based on the ground truth values. The Poisson distribution models the probability of an event (Cameron & Trivedi, 1999), and its mathematical form is:

$$Pr(Y = y | \mu) = \frac{e^{-\mu} \mu^y}{y!} \quad (y = 0, 1, 2, \dots) \quad (5)$$

where μ is the mean incidence rate of an event per unit of exposure. The exposure could be time, space, distance, area, volume, or population size. e is the number 2.71828..., commonly known as Euler's number, the base of the natural logarithm.

The Poisson regression model is obtained from the Poisson distribution by parameterizing the existing relation between the mean incidence rate μ and covariates (regressors) x . The mathematical form of the Poisson regression model is shown in Equation (6).

$$\mu_i = \exp(x_i' \beta), \quad i = 1, \dots, n \quad (6)$$

where by assumption, there are k linear covariates that are independent, and the model usually includes a constant. The regression coefficients are denoted by β .

Sometimes, when a device's manufacturer was unknown, it could indicate that the MAC address was randomized. However, it is not certain that all unknown manufacturers belonged to passengers' devices. Therefore, these manufacturers needed to be classified as riders or non-riders.

After implementing the first three rules, the devices were separated by manufacturers. The known manufacturers continued the entire process through the eighth rule. These results were fed into a random forest model to train a classifier of riders and non-riders. The variables of maximum signal strength, minimum signal strength, and detection time were used to train the model. The random forest classifier was applied to the unknown manufacturer devices that had been previously segregated.

The random forest was built from a random sample of the data. The data are the Wi-Fi datasets for the Redline during the month of October. The bootstrap sample has a size of $0.632n$. The observations that are not used to build any tree are denoted as out-of-bag observations (Janitza & Hornung, 2018). In contrast to cross-validation and other data splitting approaches, only one random forest needs to be constructed (Bylander, 2002; G. Zhang, Zhang, & Zhang, 2010).

The training data for the model was the Redline datasets of the month of October. These datasets included the known manufacturers. The classifier developed was implemented on the unknown manufacturers once the known manufacturers had been classified with the enhanced rule-based method. The classification model had an error of 35.76%. This error may introduce uncertainty into the data. The ridership estimation is the sum of the number of known signals classified as riders by the implemented rules plus the number of unknown signals classified as riders by the random forest classifier.

Random forests consist of an amalgamation of tree predictors such that each tree is dependent on the values of a random vector sampled independently and has the same distribution for all the trees in the forest (Breiman, 2001). The mathematical representation of the random forest classification model was defined by Breiman as:

$$\{h(x, \theta_k), \quad k = 1, \dots\} \quad (7)$$

where the $\{\theta_k\}$ are identical and independently distributed random vectors for each tree that represents a unit vote for the most popular class for the input X . The random forest method combines Bagging (Breiman, 1994), with a random variable selection at each node (Amit, Geman, & Wilder, 1997). Thanks to these two strategies, random forest algorithms represent one of the

most effective machine learning tools that enjoy wide applications across fields. Indeed, decision trees are the best candidates for ensemble methods because they tend to have low bias and high variance, which makes them very prone to benefit from the averaging process. The only assumption of the method is that the sample is representative (Louppe, 2014). Since the Wi-Fi dataset used in this research has the same structure and values, it is expected that random forest works well in the classification process for other datasets after being trained.

Some of the characteristics of the random forest are accuracy at least as good as contemporary classifiers, relative robustness to outliers and noise, quicker computing times compared to Bagging, and provision of internal error (Hedemalm, 2017).

4.1.3. Unsupervised Machine Learning

Unsupervised machine learning applies to data of which the response variable is not labeled. While the Smart Stations record the MAC addresses emitted by devices within their radar, they do not show whether each MAC address is from a bus rider. This research explored the use of a clustering algorithm in the classification of riders using the Wi-Fi data obtained with the Smart Stations. Clustering is a method of grouping objects to classify or categorize them into subsets called clusters. There are many methods to calculate similarity in the individual data points; however, there is not a best approach for all the datasets (Miao, 2015). The Euclidean distance between two points, denoted as x and y , is the length of the straight-line segment that connects them (Danielsson, 1980), and this distance is one of the most common approaches for cluster classification. In a Cartesian plane, for two points $x = (x_1, x_2, \dots, x_p)$ and $y = (y_1, y_2, \dots, y_p)$, their Euclidian distance d is provided by Equation (8):

$$d(x, y) = \sqrt{\sum_{i=1}^p |x_i - y_i|^2} \quad (8)$$

where p is the dimension of a data point. Clusters belong to unsupervised learning problems because the data are unlabeled and the method is considered as a type of exploratory statistics because it is a heuristic process (Bock, 1996).

This research utilizes the K-means algorithm to analyze hidden patterns in the data typically discoverable through clustering methods. K-means is a centroid-based clustering method, where K represents the number of clusters, which is an input parameter. Each data point is categorized into the clusters based on the smallest distance to each cluster. There are two main steps to implementing the K-means algorithm: 1) find the centroids and 2) categorize each data point based on the distance. Figure 39 illustrates the K-means algorithm. Mathematically, the k centroids are denoted as m_1, m_2, \dots, m_k . The initial centroids are selected randomly. The data points x_i are assigned to the clusters S_j based on the form:

$$S_j = \{x_i: ||x_i - m_j||^2 \leq ||x_i - m_t||^2, 1 \leq t \leq k\} \quad (9)$$

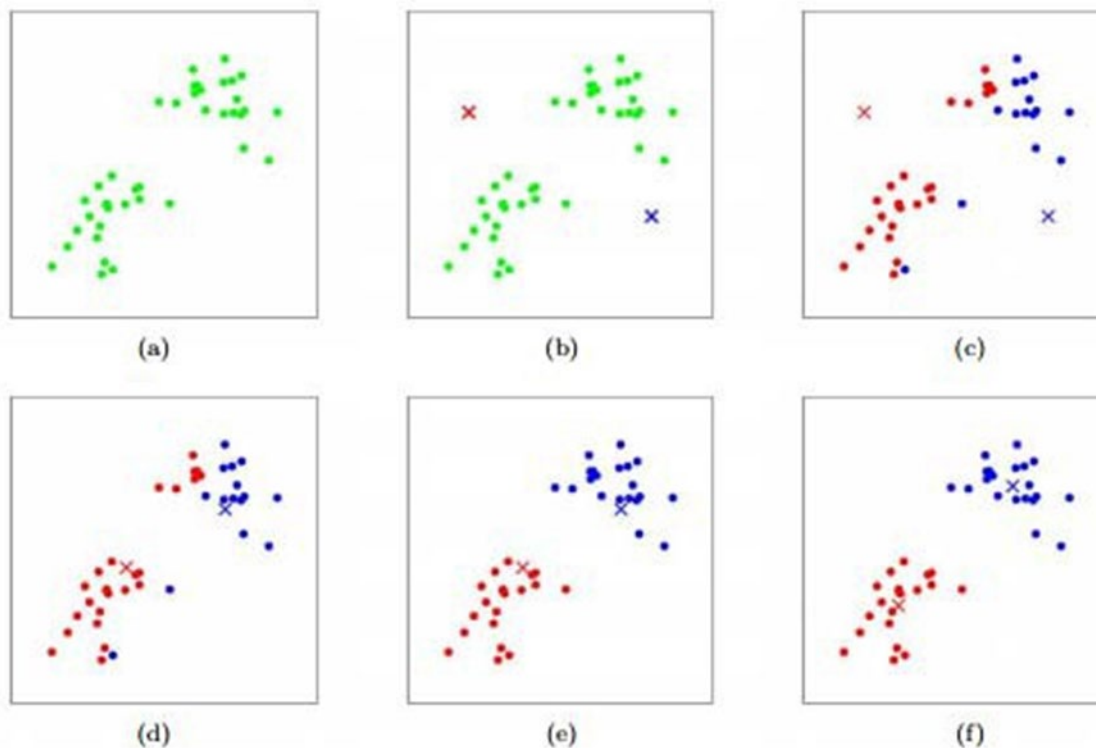


Figure 39: K-means algorithm (Piech, 2013)

In Figure 39, every element is denoted as dots, cluster centroids are denoted as crosses. (a) is the original data set. (b) is the random initialization of the clusters. (c - f) illustrate the iterations of the algorithm until the points converge to the same cluster based on Equation (9). K-means is the most common clustering algorithm and it is easy to implement (Oyelade, Oladipupo, & Obagbuwa, 2010); therefore, this study explores the use of this algorithm for classification of riders. The quantitative variables of the Wi-Fi datasets were used in the clustering analysis; these variables are maximum speed, minimum speed, maximum signal strength, minimum signal strength and detection time. Before implementing the K-means algorithm, unrealistic values were discarded. The unrealistic values are those that are discarded by the first three rules of the rule-based algorithm (see Table 24).

This study explored the use of principal component analysis (PCA) to reduce the dimensions of the quantitative variables before the implementation of the K-means algorithm. PCA is widely used in statistical analysis and it makes K-means implementation easier (Chris Ding & He, 2004). In addition, PCA allows easier representation of the data. However, PCA can be sensitive to heavy-tailed noise (Shahid, Perraudin, Kalofolias, Puy, & Vanderghenst, 2016). This procedure of implementing PCA and then applying K-means is a recognized methodology (Zha, Ding, Gu, He, & Simon, 2001).

The main goal of PCA is to compress the size of a dataset by keeping only the most important information. PCA helps better understand the data structure (Abdi & Williams, 2010). To achieve this goal, PCA calculates new variables that are called principal components. These components

are obtained as linear combinations of the original variables. The first component is required to have the largest variance, also called inertia, which explains more of the original data. The second component is calculated under the condition that it is orthogonal to the first component. The other components are computed likewise. This process uses the definition of the covariance matrix, mathematically represented as:

$$C^{n \times n} = (c_{ij}, c_{ij} = \text{cov}(\text{Dim}_i, \text{Dim}_j)) \quad (10)$$

in this case, $C^{n \times n}$ is a matrix with n rows and n columns, and Dim_x is the x^{th} dimension. This formula states that the matrix is square and each entry in the matrix is the result of computing the covariance among separated dimensions.

Additionally, the concept of eigenvectors is applied to the reduction of dimensions. The definition of eigenvectors is shown as follows:

$$A\vec{v} = \lambda\vec{v} \quad (11)$$

where v (with an arrow over it) is the eigenvector that when multiplied by the square matrix A , results in the product of the same vector by a factor λ . Even when a multiple of the eigenvector is used, the same factor multiplied by the scaled vector is obtained. The factor on the right side of the equation is called eigenvalue.

Before the data analysis, the values are standardized. The process of data standardization used was the Z-score scaling formula shown below:

$$Z_i = \frac{x_i - \bar{X}}{\sigma} \quad (12)$$

where Z_i is the new value of the individual data point. X_i is the original observation. \bar{X} represents the mean value for all the values in a particular dimension. σ corresponds to the standard deviation of the values in that specific dimension of the data. The results change the data into different values whose mean is zero and they have a standard deviation of one. The advantages of data standardization, as previous researchers have noted, is an easier comparison of different dimensions whose values can have very dissimilar ranges (Tanioka & Yadohisa, 2012), hence multivariate analysis becomes easier.

After the obtainment of the eigenvectors and the eigenvalues, the eigenvector with the largest eigenvalues corresponds to the first principal component, also known as the new dimension. These eigenvectors are unit vectors, i.e. their lengths are one. The eigenvalues are then ordered from highest to lowest to sort the components by order of significance. The components with lesser significance are ignored, and some information may be lost. However, if the eigenvalues are small, the lost information is not significant. When the least significant components are discarded, the dimensionality is reduced.

Lastly, the final data set was derived from the eigenvectors and the initial scaled data set. This procedure is shown in the following formula:

$$FinalData = RowFeatureVector \times RowDataAdjust \quad (13)$$

In this case, the *RowFeatureVector* is the matrix with the eigenvectors that become part of the row; the columns are transposed. The *RowDataAdjust* is the transposed data, each data value is in each column, and the rows hold separate dimensions.

The principal component analysis is a tool that helps analyze patterns in high-dimensional data which is not suitable for graphical representation. Additionally, computational time is reduced (Baranski, Wytyczak-Partyka, & Walkowiak, 2008). The data was not scaled because it was tested on the dataset of January 14, 2019 of the BlueLine, that when the data was scaled, the formed clusters contained too many values which was unrealistic for ridership classification. This dataset was kept in results which may introduce bias.

4.1.4. Evaluation of Ridership Estimation Methods

The estimates of passenger counts were compared to the ground truth counts performed in the field. The mean squared error (MSE) is a well-established statistical method to compare the closeness of measurements to a defined target value (Holst & Thyregod, 1999). The mathematical form of the MSE is shown as follows:

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \tilde{y}_i)^2 \quad (14)$$

where n is the number of data points, y_i represents the observed values, and \tilde{y}_i represents the predicted value. In addition, the absolute percentage error (APE) is calculated using the following formula:

$$APE = \frac{|y_i - \tilde{y}_i|}{y_i} * 100\% \quad (15)$$

In addition, the Pearson correlation (r) was used to test the linear dependence of the actual and estimated ridership. The plot of the counts versus the estimated number of passengers was generated in order to see a visual trend. The correlation was tested using the Pearson correlation formula:

$$r = \frac{\sum (x - m_x)(y - m_y)}{\sqrt{\sum (x - m_x)^2 \sum (y - m_y)^2}} \quad (16)$$

where m_x and m_y are means of x and y variables. The p-value of the correlation could be determined by using degrees of freedom equal to $n - 2$, where n is the number of observations in x and y variables.

4.2. Wait Time

In this research, the wait time is equivalent to the total time a passenger spends waiting at a bus stop, measured from the time he or she arrives at the bus stop to the time when the bus leaves the bus stop. Wait time at a bus stop strongly influences the travel attitude of passengers toward public transportation. However, what passengers perceive as their wait time can be completely different from the actual time they wait (Psarros, Kepaptsoglou, & Karlaftis, 2015).

Wait time is an important element in passengers' decision making to choose a mode of transportation, thus transit agencies can provide a more attractive service by understanding passengers' travel behavior (Salek & Machemehl, 1999).

In practice, wait times are estimated by surveying the passengers at bus stops or by field observations. Notably, reported wait times obtained from surveys can be misleading because of the subjective perception of the wait time by passengers. The same is true for field observations, due to the complexity of observing passengers in large crowds (McCord, Mishalani, & Wirtz, 2006).

This study aims to estimate the average wait time of passengers through the innovative approach of deploying Smart Stations. Two Smart Stations were deployed: one at a bus stop, and the other on the bus. An algorithm was developed to calculate the wait time based on the MAC addresses that were seen at a chosen bus stop and at a bus stopping there. Figure 40 shows the implemented algorithm. The detection time is calculated as explained in the Rule-based Method section.

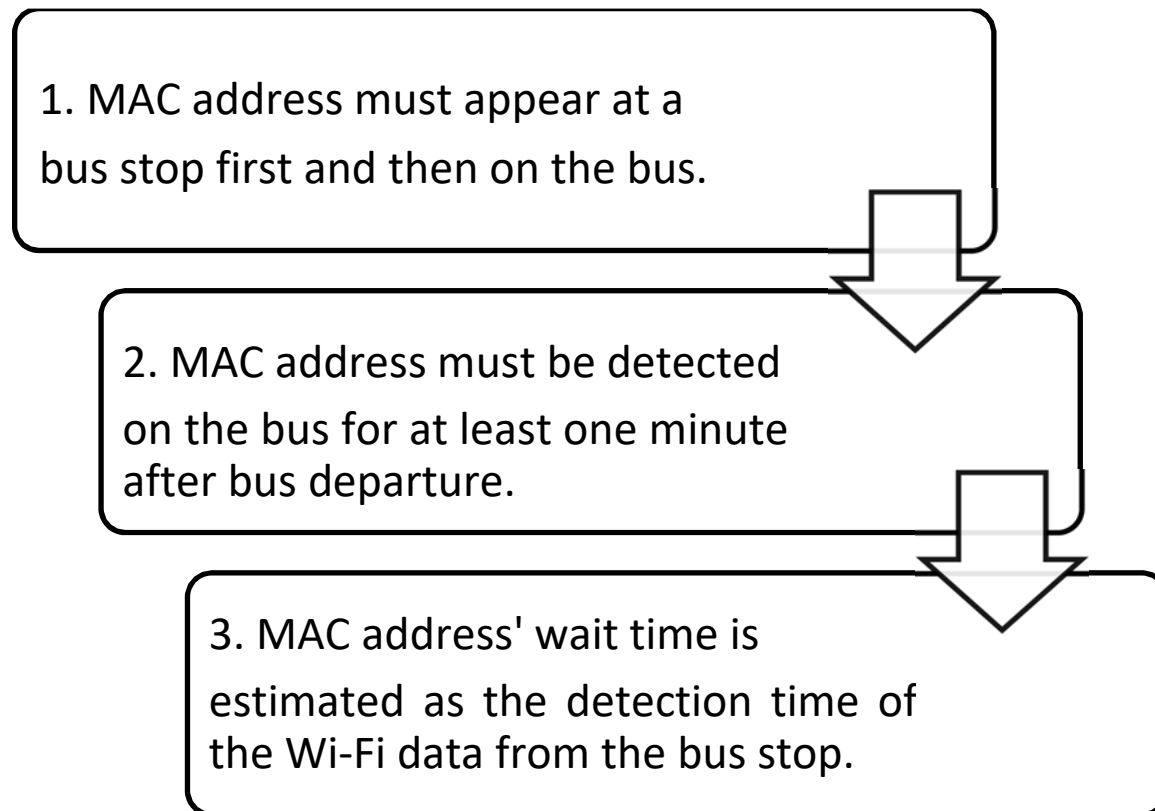


Figure 40: Algorithm to estimate the wait time

4.2.1. Experiment to Validate Detection Time

In order to know if the detection time was an unbiased estimate, an experiment was developed to test if the estimation was statistically similar to ground truth observations. The experiment consisted of having a Smart Station scan data at the Montana State University Transfer Station bus stop. Meanwhile, the Wi-Fi on mobile devices was turned on and off at predetermined intervals of time which were used as the ground truth values. In addition, to avoid repeating the counts, the Smart Stations were rebooted after every measurement. Table 26 shows the variables and their different categories which were tested in this experiment.

Table 26: Variables and categories per variable used in wait time experiment

Variable	Category 1	Category 2	Category 3
Device type	LG	Samsung	Apple
Distance (meters)	1	5	10
Time* (seconds)	60	180	300

*Total time the device's Wi-Fi was turned on

The main purpose of this experiment was to account for the different variables of device type, distance, and the detectable time to discover if they influence the accuracy of the results provided by the Smart Station's measurements. An initial three-way interaction linear model was proposed to see if the variables influenced the error. The error is the difference between the Smart Station obtained time and the actual time the device's Wi-Fi was turned on. A p-value above 0.05 was used to determine the variables that were not significant in this model. The error is represented as *Time.Error*, and the model is shown in the following equations:

$$\text{Time. Error} = a_0 + \beta_1 * \text{Distance} + \beta_2 * \text{Device.Type} + \beta_3 * \text{Actual.Time} + \mathcal{E} \quad (17)$$

Equation (16) says that the error \mathcal{E} has a normal distribution centered at zero and has a standard deviation of σ . When the experiments were performed, the Apple devices randomized their MAC addresses; therefore, their detection time could not be obtained. The only occasion in which Apple devices do not randomize their MAC addresses is when they are connected to a Wi-Fi network, which is not expected when they are on the Streamline buses. Thus, the Apple device was left out of this model.

Regression is a procedure for studying relations between variables, where such relations are approximated using functions (Schneider, Hommel, & Blettner, 2010). When the relation between the dependent variable (Y) and only one independent variable (X) is established, the model is called simple linear regression (SLR).

The multiple linear regression (MLR) model associates the dependent variable to several independent variables, also called predictor variables. MLR is mathematically expressed as follows:

$$Y = \alpha_0 + \beta_1 * X_1 + \dots + \beta_p * X_p + \mathcal{E} \quad (p = 1, \dots, n) \quad (18)$$

where the a and β symbols are called regression parameters or coefficients. Usually, the first coefficient is interchangeably denoted with the alpha or beta Greek letters. The idea behind the regression models is to obtain the coefficients such that the minimum SSE will be obtained. Considering the matrix notation of Equation (18):

$$Y = \alpha + \beta X + \mathcal{E} \quad (19)$$

the mathematical formula to obtain the minimum least squares is:

$$(\hat{\alpha}, \hat{\beta}) = \arg \min_{\alpha, \beta} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (20)$$

where $\hat{\alpha}$ and $\hat{\beta}$ are called the least squares estimates, y_i is the actual observation, and \hat{y}_i is the estimated value.

The assumptions of the SLR model are: (1) errors are normally distributed, (2) the means of the responses follow a straight-line function of the explanatory variable, (3) the errors for different values of the explanatory variable have equal variance, and (4) the observations are independent. The MLR model has the extra assumption that variables should not be correlated with each other (Cheng, 2006).

4.2.2. Experiment to Estimate Wait Time

The BlueLine's Wal-Mart bus stop was chosen as a location for the experiment of measuring the wait time using the Smart Stations. This bus stop was chosen because only one line serves the bus stop, which would reduce the noise and difficulty of observing passengers for multiple lines. It is also a common place for passengers, which would allow the surveyor to gather a sample size big enough for analysis.

One SS was located on the bus and one at the bus stop. The procedure of Figure 38 was implemented. The data were collected for a period of seven weekdays. For each day, a total of three hours was surveyed. The SS were rebooted every hour to avoid overestimation of the detection time. The two-sample t-test was applied to test whether the sample from the observations and the SS have statistically similar mean values.

The t-test uses the central limit theorem (CLT) as its most fundamental theory for application. The CLT states that the means of a random sample size, n , from a population with mean μ , and a variance of σ^2 , make up a normal distribution with mean μ and a variance of the fraction σ^2/n (Kwak & Kim, 2017). Mathematically, the CLT is written as:

$$\lim_{n \rightarrow \infty} \text{distribution} \left(\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \right) = N(0,1) \quad (21)$$

CLT holds even when the data X is not normal but is dependent on the sample size. For severely skewed data, an $n \geq 30$, is considered to behave normally. The test statistic of the t-test is mathematically shown as Equation (21):

$$T = \frac{X_1 - \bar{X}_2 - \Delta_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \quad (22)$$

Where Δ_0 is a specific value, usually zero, S_1 and S_2 are the estimated standard deviation of the first and second sample, respectively. The difference of the means of the two groups, $X_1 - X_2$, follow an approximately normal distribution as shown in the following equation:

$$X_1 - \bar{X}_2 \sim N(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}) \quad (23)$$

where σ_1 and σ_2 are the parametrical standard deviation of the first and second population, respectively. The confidence intervals CI for the difference of the means of the two groups is obtained with the following:

$$CI = X_1 - \bar{X}_2 \pm t_{1-\frac{\alpha}{2}, df} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \quad (24)$$

where df denotes the degrees of freedom, and α the significance level. The latter is 0.05 throughout this research. The degrees of freedom are calculated by subtracting the sample size and the number of estimates.

4.3. OD Flow Characteristics

An OD matrix describes the commuting patterns over time at selected geographical locations (Bahoken & Raimond, 2013). This research uses the SS data to infer these patterns of the Streamline riders based on geographical locations and time stamps of the passengers' smartphones detected by their Wi-Fi signals. An OD matrix is a description of the spatiotemporal trajectory T_k which is composed of consecutive points, defined as:

$$T_k = \{p_1, p_2, \dots, p_{n-1}, p_n\} \quad (25)$$

where $p_i = (x_i, y_i, t_i)$ is a record point which has a spatial location (x, y) at moment t , $t_1 < t_2 < \dots < t_i < t_{n-1} < t_n$, and $i = 1, 2, \dots, n$ representing the number of points composing T_k .

An OD matrix definition states that for each trajectory T_k , there is a flow between the points (i, j) , noted as $F(i, j)$ if the next conditions are satisfied:

$$p_1 \subset i \text{ OR } (p_2 \subset i \text{ AND } p_1 \in \text{neighbors list of } p_2) \quad (26)$$

$$p_n \subset j \text{ OR } (p_{n-1} \subset j \text{ AND } p_n \in \text{neighbors list of } p_{n-1}) \quad (27)$$

The concept of an OD matrix is shown in Figure 41, which exhibits movements, also known as OD pairs, between five locations. From the left side of the graph, an OD matrix was built, located on the right side. Each OD pair becomes a cell and if no flow is recorded, the cell acquires a value of zero. This research utilizes this concept to show OD flows obtained using Wi-Fi signals.

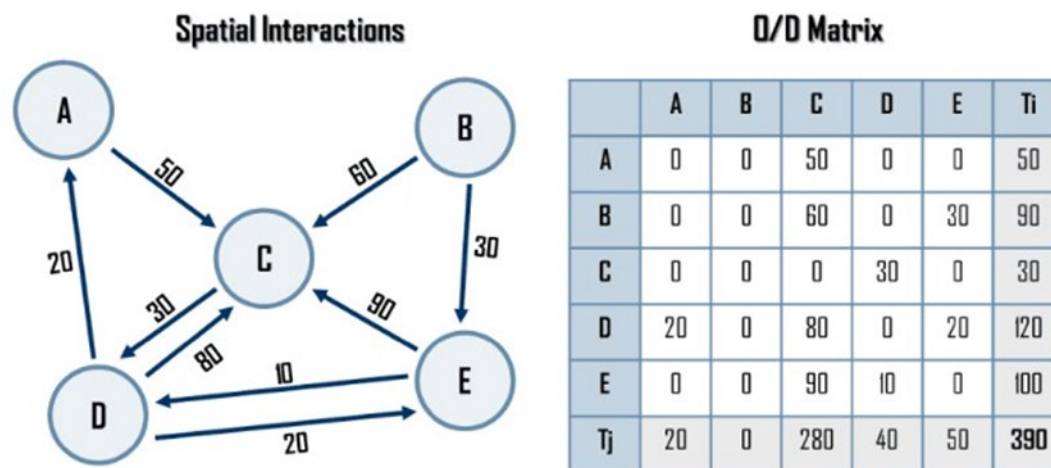


Figure 41: OD flows and matrix representation (Rodrigue et al., 2017)

4.4. Estimated Time of Arrival

Previous research suggests that passengers perceive waiting at bus stops as a burdensome task (Ben-Akiva & Lerman, 1987). This is associated with passengers' high level of uncertainty experienced when they have to wait for a transit vehicle when they do not know when it will arrive (Jaffe, 2015). Therefore, knowing the wait time of a transit system improves the user experience of the passengers.

To test if the Smart Station estimates provided accurate measurements of arrival time, ground truths were recorded in the field. Later, the ground truths were compared to the estimated values of travel times.

The times were obtained manually as mentioned in the Datasets chapter. The Smart Station was also collecting GPS data. These data contain the coordinates with time stamps. An algorithm was developed to calculate the travel time between every bus stop. The bus stop coordinates and the timestamps were used to estimate the travel times.

The algorithm obtains the arrival time and the departure time from all the bus stop based on the distance threshold. When the bus is located approximately fifteen meters from the bus stop, both the arrival and departure time are recorded. The arrival time corresponds to the time the bus arrives at a bus stop and is the first time recorded under the threshold condition. On the other hand, the departure time corresponds to the time the bus leaves a bus stop and is the last time under the threshold condition. The following equations show this algorithm in mathematical form:

$$\text{Travel time} = \text{Arrival time } (i + 1) - \text{Departure time } (i) \quad (28)$$

$$\text{Stopped time} = \text{Departure time } (i) - \text{Arrival time } (i) \quad (29)$$

where $i = (1, \dots, n)$ and represents the i^{th} bus stop. Although the algorithm is simple, many inconveniences occur because fifteen meters is not always the adequate threshold since the bus does not always park within this distance. This threshold was the first tried and then it was adjusted for every bus stop depending on the distance the bus stopped from the bus stop. In addition, the buses returned to the same location, generating more complexity to the calculation of the arrival and departure times.

The travel times were evaluated using a paired t-test because the observations were highly correlated to each other. The difference between the observed and estimated travel time was obtained. The t-test assessed if the difference was equal to zero on average or not. The null and alternative hypotheses take the form of:

$$H_0: (y_i - \tilde{y}_i) = 0 \quad (30)$$

$$H_a: (y_i - \tilde{y}_i) \neq 0 \quad (31)$$

where y_i is the observed value, measured with the chronometer, and \tilde{y}_i is the estimated value obtained with the implementation of the algorithm on the GPS datasets.

Moreover, the mean difference or error was evaluated using a multi-factor analysis of variance (ANOVA). This test evaluates if there is a difference mean among groups for many variables or factors (H. Kim, 2014). The factors utilized in this research are bus line, peak or off-peak time, and day of the week. The assumptions of ANOVA are the following:

(1) random sampling, (2) independence of the measurements, (3) homogeneity of variance of the residuals, and (4) normal distribution of the residuals.

The hypotheses that are being evaluated by ANOVA are:

$$H_0: (y_i - \tilde{y}_i)_{i,j,k} = 0 \quad (32)$$

$$H_a: (y_i - \tilde{y}_i) \neq 0, \text{ for some } i, j, k, l \quad (33)$$

where i represents the bus lines, j represents the days of the week, k signifies the morning or afternoon period of the day, l denotes the peak or off-peak hours. The TukeyHSD test was performed to obtain the confidence intervals and for graphical representation of the results.

In order to analyze whether outliers play a dominant role in the model, a combination of leverage and Cook's distance was implemented. The leverage is a distance between an explanatory variable value and the average of the explanatory variable values in the entire dataset (Ranganai, Van Vuuren, & De Wet, 2014). The i^{th} leverage h_i is denoted with the following equation:

$$h_i = \left[\frac{SE(\hat{\mu}\{Y|X_1, X_2, \dots\})}{\hat{\sigma}} \right]^2 \quad (34)$$

where SE is the standard error of the developed model and $\hat{\sigma}$ is the estimated standard deviation of the residuals. The Cook's distance quantifies the influence of the i^{th} datapoint by determining the effect of omitting such a datapoint on the fitted values of the model (Q. Gao, Ahn, & Zhu, 2014). The mathematical form of the Cook's distance D_i is the following:

$$D_i = \sum_{j=1}^n \frac{Y_{j(i)} - Y_j}{\tilde{p}\hat{\sigma}^2} \quad (35)$$

where Y_j is the j^{th} fitted value using all observations, $Y_{j(i)}$ is the j^{th} fitted value excluding observation i , \tilde{p} represents the number of regression coefficients $\beta_0, \beta_1, \dots, \beta_{p-1}$, and $\hat{\sigma}^2$ is the estimate of σ^2 .

It can be noted that the sample size of the data collected was not a parameter that was prioritized. This research was made in the period of a year, and data collection, especially data processing was time-consuming. However, this research focused on evaluating if the estimated values of the transit system were similar to the ground truths obtained.

5. RESULTS AND ANALYSIS

This chapter summarizes the results obtained in this research. In addition, a discussion based on the results is provided. The Results and Analysis chapter is separated into the following: (1) Ridership (2) OD Flow Characteristics, (3) Wait Time, and (4) Travel Time.

5.1. Ridership

The ridership estimation methods focus on the use of Wi-Fi and GPS data collected by Smart Stations. There are three approaches explored in this research: (1) Rule-based method, (2) Enhanced rule-based method, and (3) unsupervised machine learning.

5.1.1. Rule-based Method

This method employed six rules which intended to exclude those signals that did not belong to passengers. The rules are explained in Table 24 in Chapter 3. After the implementation of each rule, the number of signals from the raw data decreased. Table 27, 28, 29, 30, and 31 show the number of signals after the implementation of the rules for the Blueline, Greenline, Orangeline, Redline, and Yellowline respectively. The final ridership estimate of each route is the number when the not applicable values, shown as NA in the datasets, were removed.

Table 27: Number of signals after implementation of Rule-based method for the Blueline

Date	Signals	R 1	R.2	R.3	R.4	R.5	R.6	Estimates
1/14/2019	3318	435	35	35	34	28	26	26
1/15/2019	2951	340	37	36	35	32	28	28
1/16/2019	3351	461	36	36	36	28	23	23
1/17/2019	3686	504	46	45	45	42	40	40
1/18/2019	4331	595	49	49	49	40	34	34
10/8/2018	3580	534	36	36	36	33	25	23
10/10/2018	3837	468	38	37	37	33	30	28
10/11/2018	3206	550	28	27	27	19	18	18
10/12/2018	4229	561	65	65	65	62	60	60
10/18/2018	2877	303	20	20	19	17	16	16

Table 28: Number of signals after implementation of Rule-based method for the Greenline

Date	Signals	R 1	R.2	R.3	R.4	R.5	R.6	Estimates
2/4/2019	3400	531	50	49	49	36	31	27
2/5/2019	3698	543	47	47	47	38	35	32
2/6/2019	2693	472	25	23	23	15	8	8
2/7/2019	3105	606	30	29	29	21	15	13
2/8/2019	3307	672	37	35	35	17	7	7
10/17/2018	2918	271	34	34	34	27	17	17
10/18/2018	2640	345	42	42	42	35	23	23
10/19/2019	2464	423	26	25	24	11	4	4

10/19/2019	3289	405	45	45	45	33	27	27
------------	------	-----	----	----	----	----	----	----

Table 29: Number of signals after implementation of Rule-based method for the Orangeline

Date	Signals	R 1	R.2	R.3	R.4	R.5	R.6	Estimates
1/7/2019	2198	173	13	13	13	11	10	10
1/8/2019	2150	185	22	22	22	22	18	18
1/9/2019	2344	190	14	14	14	14	12	12
1/10/2019	2235	194	28	28	28	27	24	24
1/11/2019	2359	147	17	17	17	17	17	6
10/8/2018	2142	200	11	11	11	10	9	9
10/10/2018	1959	157	3	3	3	3	2	2
10/18/2018	2653	297	44	44	44	43	41	41

Table 30: Number of signals after implementation of Rule-based method for the Redline

Date	Signals	R 1	R.2	R.3	R.4	R.5	R.6	Estimates
1/7/2019	4248	252	30	30	30	28	25	24
1/8/2019	4357	316	35	35	35	28	27	27
1/9/2019	3641	205	36	36	36	33	32	32
1/10/2019	4381	259	32	32	32	29	25	25
1/11/2019	4373	276	29	29	29	28	27	27
10/16/2018	4999	355	90	89	86	86	84	81
10/17/2018	4662	333	30	30	29	27	25	25
10/17/2018	3888	296	21	21	21	19	18	18
10/18/2018	4935	353	43	43	43	42	40	39
10/19/2018	4790	258	34	34	34	33	33	33

Table 31: Number of signals after implementation of Rule-based method for the Yellowline

Date	Signals	R 1	R.2	R.3	R.4	R.5	R.6	Estimates
1/14/2019	2053	115	17	17	17	16	15	15
1/15/2019	2560	162	41	41	41	39	37	37
1/16/2019	2967	178	51	51	50	49	47	47
1/17/2019	2378	133	29	29	29	28	25	25
1/18/2019	2487	139	21	21	21	19	19	19
10/10/2018	2028	186	26	26	26	24	24	24
10/18/2018	2011	175	10	10	10	8	6	6
10/18/2018	2607	133	23	23	23	22	22	22
10/19/2018	2634	161	52	52	52	51	49	37

Expectedly, the reduction of signals is considerable. On average, only around 0.8% of the detected signals are retained. This means that most networks detected are not passengers. As analyzed in

the Datasets chapter, a large percentage of networks were infrastructure-based. Therefore, all these needed to be removed. The first rule discards around 90 percent of all signals. This means that most of the networks that are detected are sporadic or permanent networks. These had to be thrown out because they would not belong to signals whose source are passengers' devices. The second rule only leaves those signals that belong to probes and it retained around one percent of all signals. The other rules discard a minor percentage of passengers; however, they are also believed to separate riders from non-riders.

Figure 42 shows the dispersion of the estimated values and the counted passengers in the field for the Blue line, Green line, Orange line, Red line, and Yellow line. In addition, a line of perfect prediction is shown in order to show the dispersion of the estimates with respect to the desired value. This line of perfect prediction has an intercept of zero, and a slope of one and has a red color.

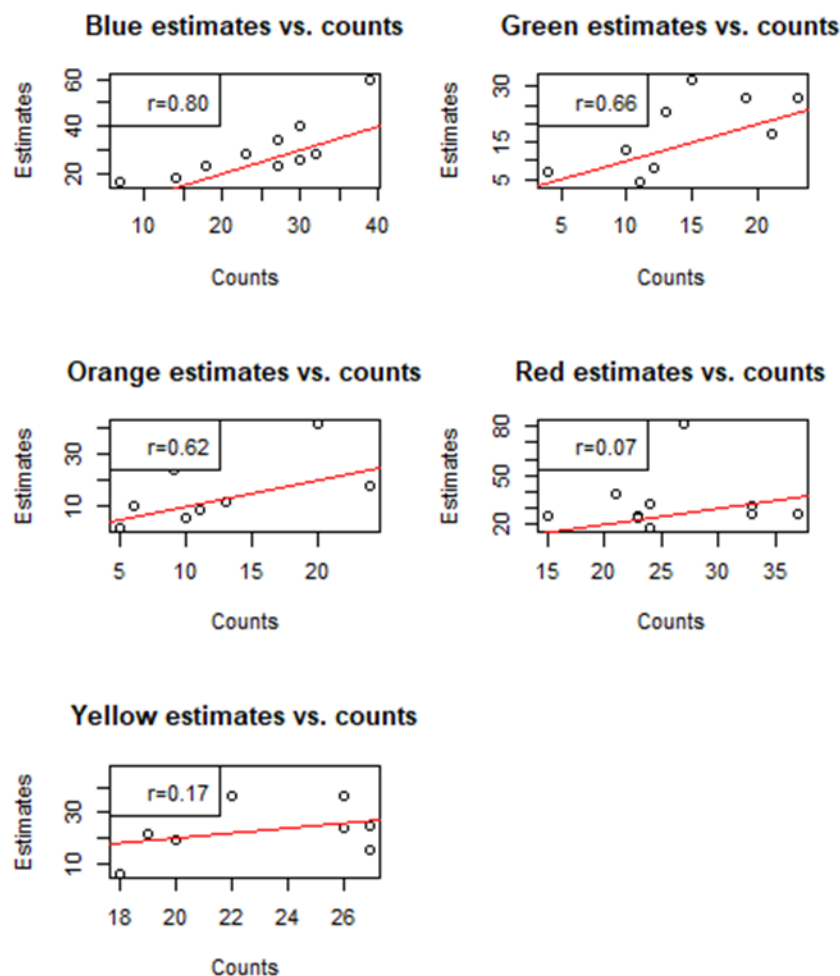


Figure 42: Plots of rule-based estimated values vs. counts by bus lines

It can be seen, from Figure 42, that there is a perceivable correlation between the number of estimated passengers and the manual counts performed. This correlation is positive, which provides an indication that the number of passengers can be estimated from the Smart Station's Wi-Fi scanning. For all the lines there are values that over and underestimate the true ridership. Therefore, a single graph of all the data points was computed. Figure 43 shows the dispersion of the estimated values and counted passengers for all the lines.

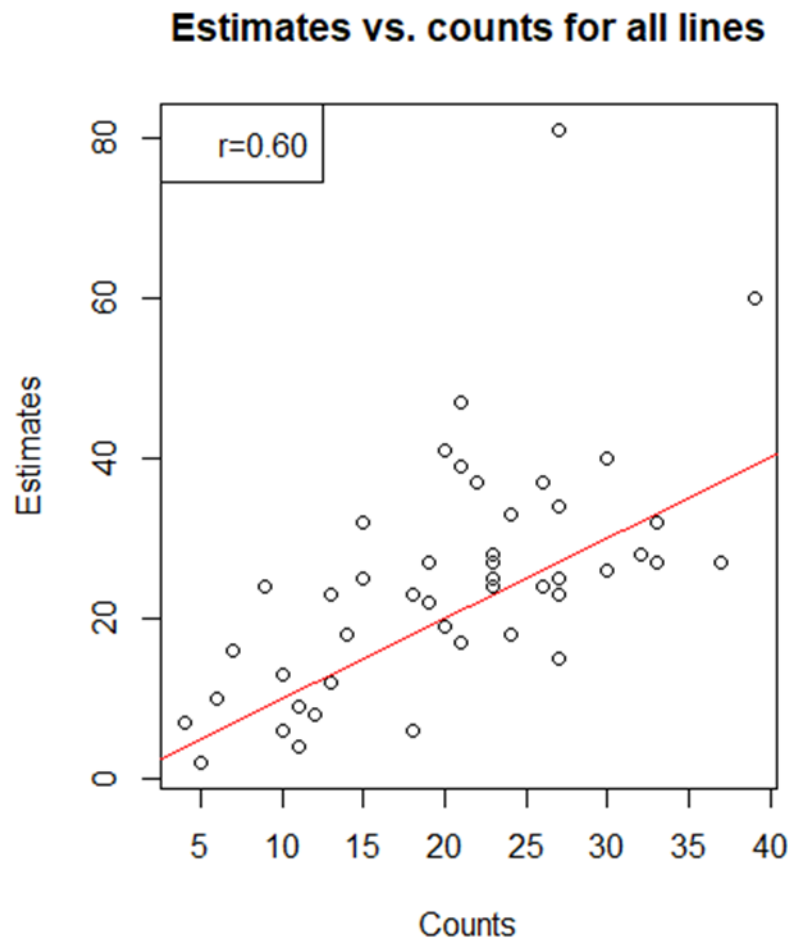


Figure 43: Plot of rule-based estimated values vs. counts for all lines

In Figure 43, there is a general overestimation of the number of passengers with the application of the rule-based method. Nevertheless, the estimated number of passengers increase as the number of counts increase. Therefore, a relationship can be established. Using the Poisson regression, the relationship was deduced. The resulting coefficients of this regression are shown in Table 32.

Table 32: Coefficients of the Poisson regression for the rule-based method

	Estimate	Standard Error	z value	Pr(> z)
Intercept	2.626784	0.063492	41.372	<2e-16***
Estimated ridership	0.014709	0.001919	7.665	1.79e-14***

Significance codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’

The coefficients show that the regression is significant at higher than a 0.001 level. Therefore, the results indicate a strong relationship. Even with the occurrence of outliers, the model is strong and can be used for ridership estimation. The estimated model is defined by the following equation:

$$\ln(\hat{\mu}) = 2.626784 + 0.014709W_i \quad (36)$$

where $\hat{\mu}$ represents the corrected estimated number of passengers and W_i represents the estimated number of passengers with the rule-based algorithm.

The number of passengers was estimated using Equation (36) in order to approximate them closer to the ground truth data. These corrected estimated values were plotted in the y axis and the manual counts were presented in the x axis. This dispersion is shown in Figure 44.

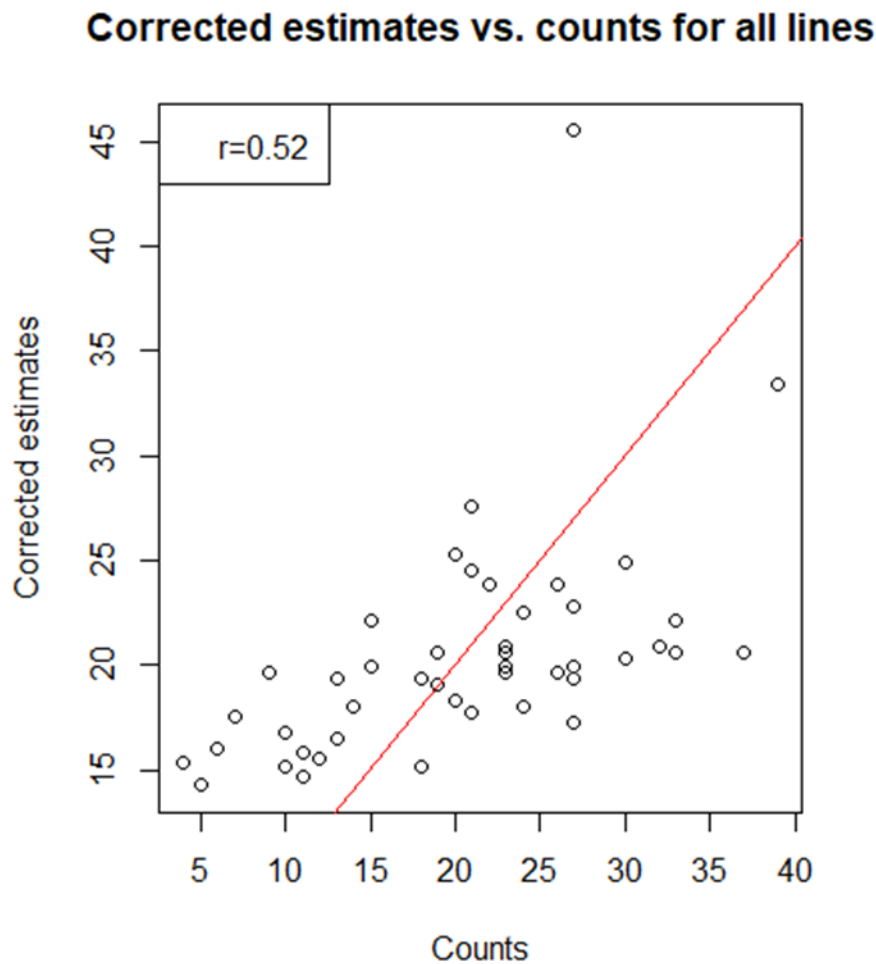


Figure 44: Plot of corrected ridership estimates vs. counts of the rule-based

Although the errors are reduced because that is the goal of the data correction, the data points seem to be less associated with the line of perfect prediction. This could be an indication of overfitting the model to make the estimates closer to the ground truth values. However, the corrected estimates do not vary greatly, and the accuracy is improved overall. More research using the simple rule-based model will provide a better understanding for making informed decisions on whether to correct the estimates or not.

Tables 33, 34, 35, 36, and 37 show the mean squared errors (MSE) and the absolute percentage errors (APE) of Blueline, Greenline, Orangeline, Redline, and Yellowline, respectively. The corrected estimates were rounded down to the closer integer.

Table 33: MSE and APE of the Blueline after implementation of rule-based method.

Date	Passenger	Estimate	MSE	APE	Estimate*	MSE*	APE*
1/14/2019	30	26	16	-13%	20	100	-50%

1/15/2019	32	28	16	-13%	20	144	-60%
1/16/2019	27	23	16	-15%	19	64	-42%
1/17/2019	30	40	100	33%	24	36	-25%
1/18/2019	27	34	49	26%	22	25	-23%
10/8/2018	18	23	25	28%	19	1	5%
10/10/2018	23	28	25	22%	20	9	-15%
10/11/2018	14	18	16	29%	18	16	22%
10/12/2018	39	60	441	54%	33	36	-18%
10/18/2018	7	16	81	129%	17	100	59%

* Indicates values after Poisson correction

Table 34: MSE and APE of the Greenline after implementation of rule-based method.

Date	Passenger	Estimate	MSE	APE	Estimate*	MSE*	APE*
2/4/2019	23	27	16	17%	20	9	-15%
2/5/2019	15	32	289	113%	22	49	32%
2/6/2019	12	8	16	-33%	15	9	20%
2/7/2019	10	13	9	30%	16	36	38%
2/8/2019	4	7	9	75%	15	121	73%
10/17/2018	21	17	16	-19%	17	16	-24%
10/18/2018	13	23	100	77%	19	36	32%
10/19/2019	11	4	49	-64%	14	9	21%
10/19/2019	19	27	64	42%	20	1	5%

* Indicates values after Poisson correction

Table 35: MSE and APE of the Orangeline after implementation of rule-based method.

Date	Passenger	Estimate	MSE	APE	Estimate*	MSE*	APE*
1/7/2019	6	10	16	67%	16	100	63%
1/8/2019	24	18	36	-25%	18	36	-33%
1/9/2019	13	12	1	-8%	16	9	19%
1/10/2019	9	24	225	167%	19	100	53%
1/11/2019	10	6	16	-40%	15	25	33%
10/8/2018	11	9	4	-18%	15	16	27%
10/10/2018	5	2	9	-60%	14	81	64%
10/18/2018	20	41	441	105%	25	25	20%

* Indicates values after Poisson correction

Table 36: MSE and APE of the Redline after implementation of rule-based method

Date	Passenger	Estimate	MSE	APE	Estimate*	MSE*	APE*
1/7/2019	23	24	1	4%	19	16	-21%

1/8/2019	37	27	100	-27%	20	289	-85%
1/9/2019	33	32	1	-3%	22	121	-50%
1/10/2019	23	25	4	9%	19	16	-21%
1/11/2019	33	27	36	-18%	20	169	-65%
10/16/2018	27	81	2916	200%	45	324	40%
10/17/2018	15	25	100	67%	19	16	21%
10/17/2018	24	18	36	-25%	18	36	-33%
10/18/2018	21	39	324	86%	24	9	13%
10/19/2018	24	33	81	38%	22	4	-9%

* Indicates values after Poisson correction

Table 37: MSE and APE of the Yellowline before and after Poisson correction

Date	Passenger	Estimate	MSE	APE	Estimate*	MSE*	APE*
1/14/2019	27	15	144	-44%	17	100	-59%
1/15/2019	26	37	121	42%	23	9	-13%
1/16/2019	21	47	676	124%	27	36	22%
1/17/2019	27	25	4	-7%	19	64	-42%
1/18/2019	20	19	1	-5%	18	4	-11%
10/10/2018	26	24	4	-8%	19	49	-37%
10/18/2018	18	6	144	-67%	15	9	-20%
10/18/2018	19	22	9	16%	19	0	0%
10/19/2018	22	37	225	68%	23	1	4%

* Indicates values after Poisson correction

Figures 45, 46, 47, 48, and 49 display the number of passengers counted, estimated and corrected for each of the five lines respectively.

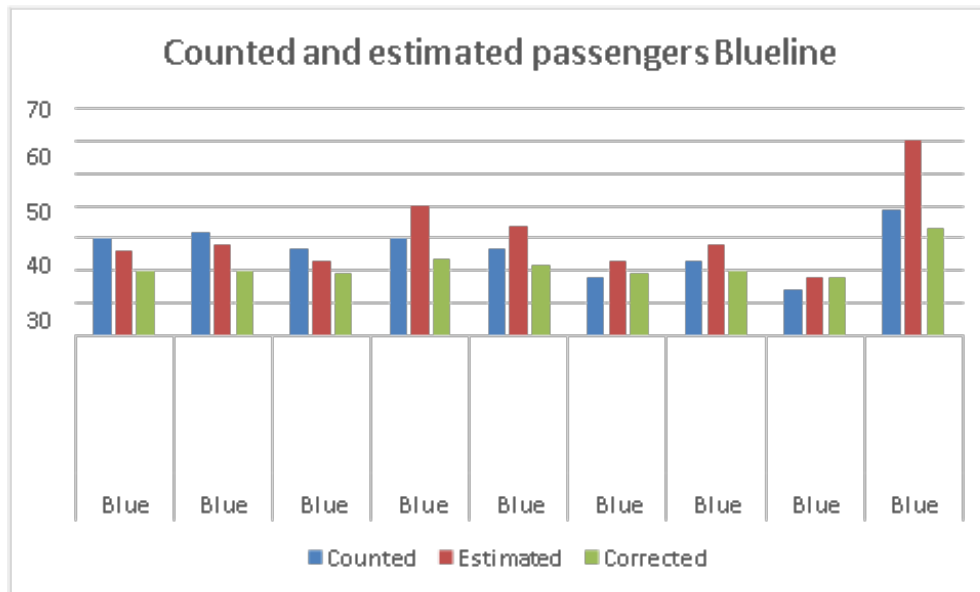


Figure 45: Counted, estimated, and corrected passengers with the rule-based method of the Blueline

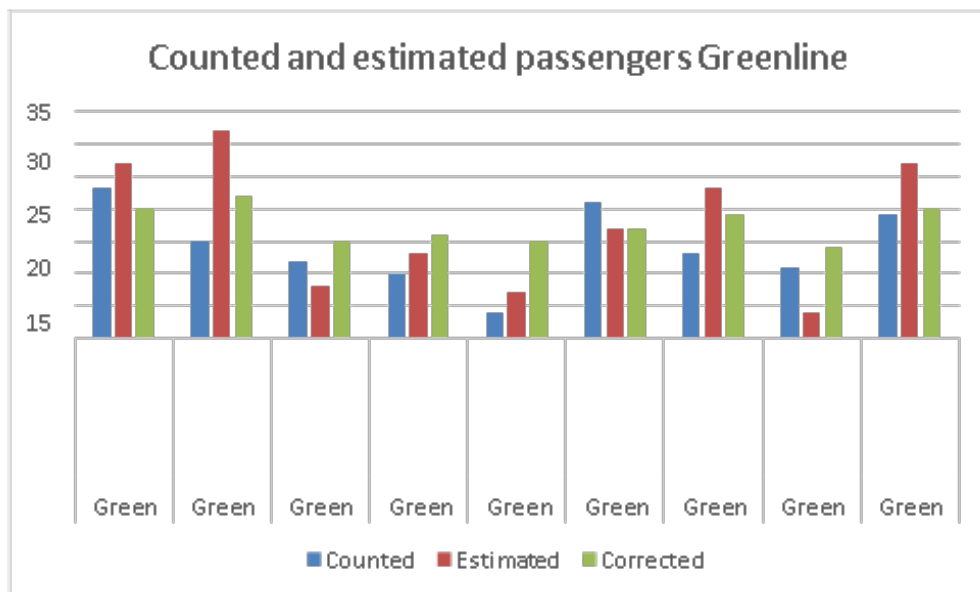


Figure 46: Counted, estimated, and corrected passengers with the rule-based method of the Greenline

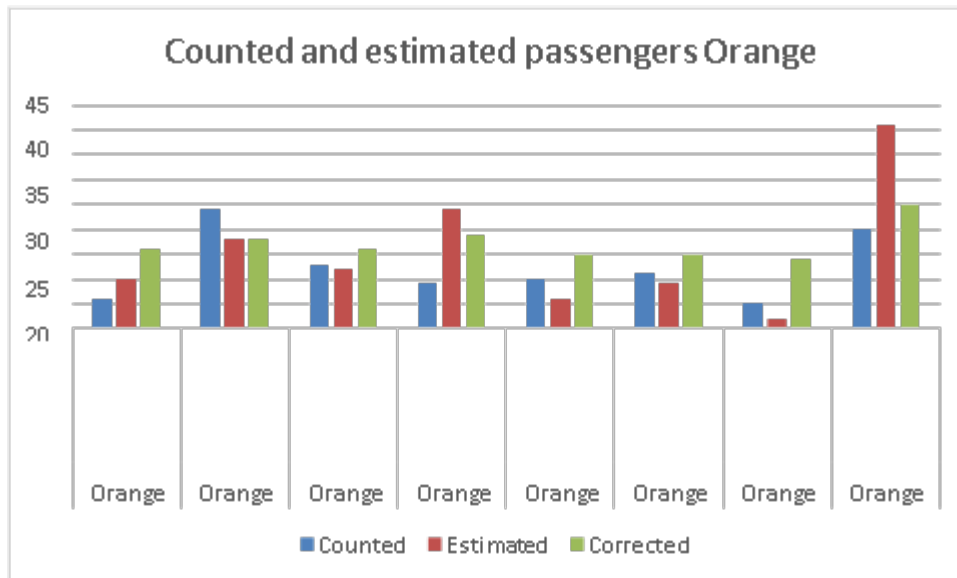


Figure 47: Counted, estimated, and corrected passengers with the rule-based method of the Orangeline

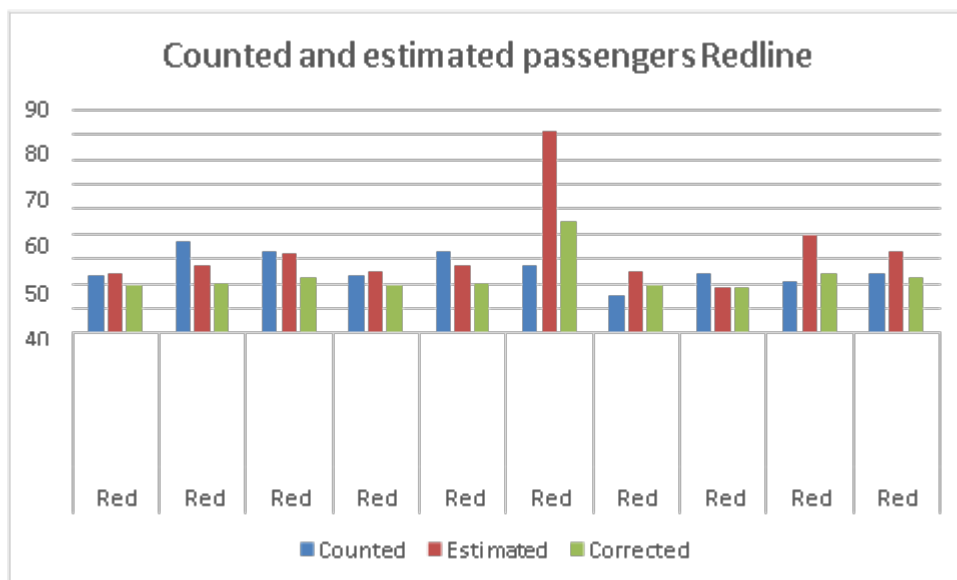


Figure 48: Counted, estimated, and corrected passengers with the rule-based method of the Redline

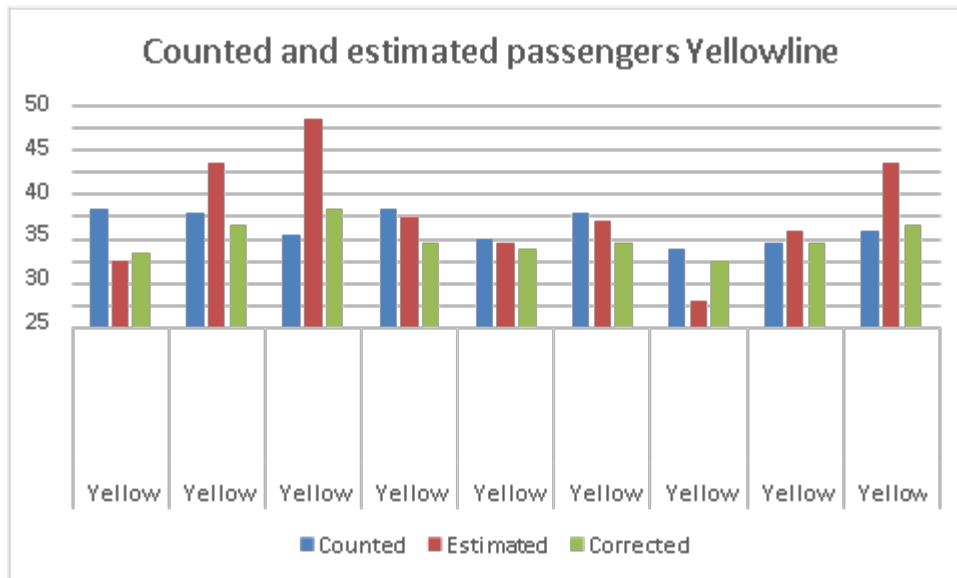


Figure 49: Counted, estimated, and corrected passengers with the rule-based method of the Yellowline

For the whole data, the MSE is 152.8 on average before correction. After Poisson correction, the average MSE is reduced to 53.9. Therefore, the Poisson correction decreases the error of the estimated values with respect to the ground truths. Considering the APE, the value is 25% on average before correction. After correction, the APE is -2%. This indicates that the estimated ridership before correction overestimates the actual ridership by 25%. On the other hand, the Poisson regression model provides an accuracy of 98% for the case study.

These results are the best found in the literature. This could be associated with the use of a powerful Wi-Fi scanning software. This software provides estimated speed, location, signal strength, type of network and more information for each device detected. These attributes were used in a relatively simple rule-based model for the first time in ridership estimation.

5.1.2. Enhanced Rule-based Method

Utilizing the enhanced rule-based method, the estimated values for ridership were obtained. The code for the enhanced rule-based algorithm was provided externally as explained in Chapter 4. Figure 50 displays the dispersion of the estimated values and the counted passengers for the Blue line, Green line, Orange line, Red line, and Yellow line. Similarly to the rule-based plot, a red line was drawn in order to show the dispersion with respect to the desired values.

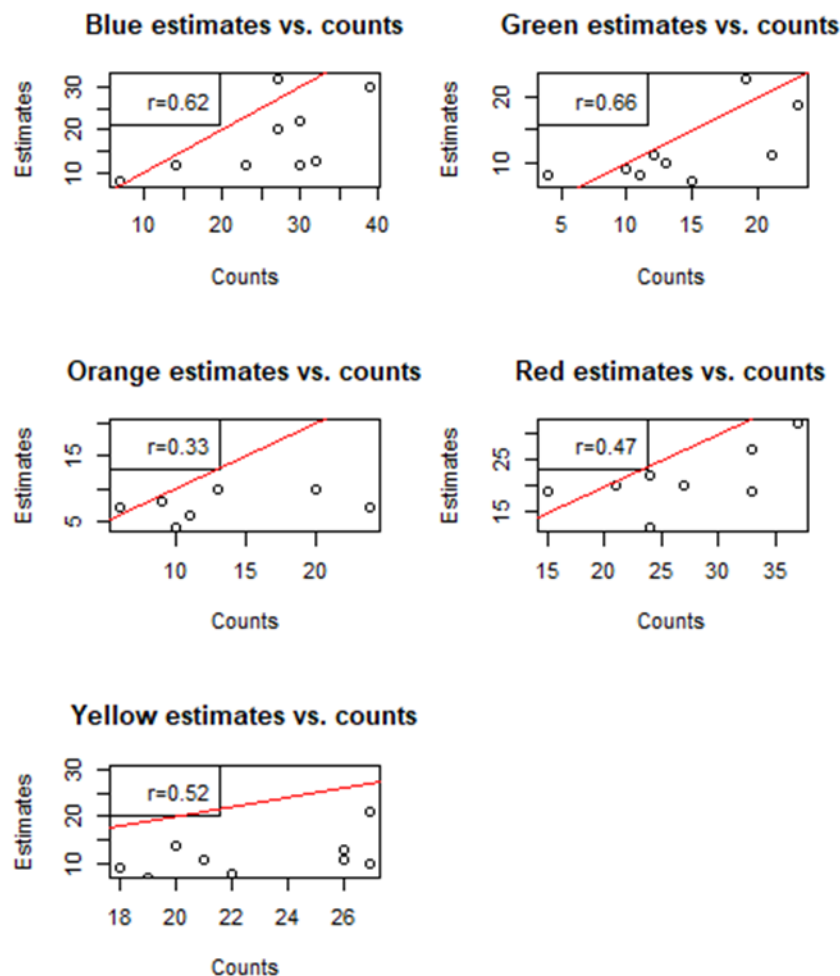


Figure 50: Plots of enhanced rule-based estimated values vs. counts by bus lines

Figure 50 shows that for all lines there is an overall trend to underestimate the number of passengers. The estimated values for the Redline seem to be closer to the line of perfect prediction. This could be since the model was trained using the Redline values for the month of October as. In order to have a better picture of the entire data, the scatterplot was made for all the lines as shown in Figure 51.

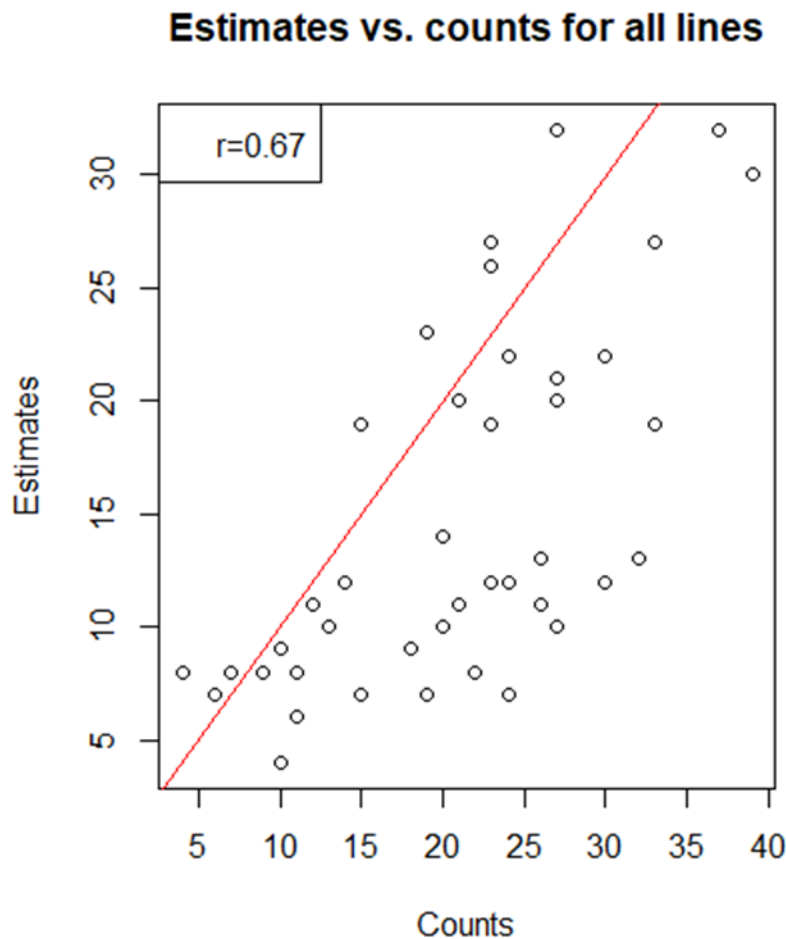


Figure 51: Plot of enhanced rule-based estimated values vs. counts for all lines

In Figure 51, the estimated values also show that there is an underestimation of ridership. However, the estimated values seem to increase as the counted values do. This is an indication that a correlation can be established. The Poisson regression model was implemented in order to correct the estimated values. The coefficients of this regression are shown in Table 38.

Table 38: Coefficients of the Poisson regression for the enhanced rule-based method

	Estimate	Standard Error	Z value	Pr(> z)
Intercept	2.522527	0.075284	33.507	<2e-16***
Estimated ridership	0.032702	0.004041	8.093	5.81e-16***

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.'

The coefficients show that the regression is significant at a higher than 0.001 level. This is indicative of a relationship. Even with the occurrence of the outliers, the model is strong and can be used for ridership estimate. In addition, even if the residuals of the model have fluctuating variances, the estimated values are unbiased. The estimated model is defined by the following equation:

$$\ln(\hat{v}) = 2.2552527 + 0.032702V_i \quad (37)$$

where \hat{v} represents the corrected estimated number of passengers and V_i represents the estimated number of passengers with the enhanced rule-based algorithm.

Utilizing Equation (37), the estimated number of passengers was corrected in order to have more accurate values. Figure 52 shows the plotted values of the corrected estimates in the y axis and the manual counts in the x axis.

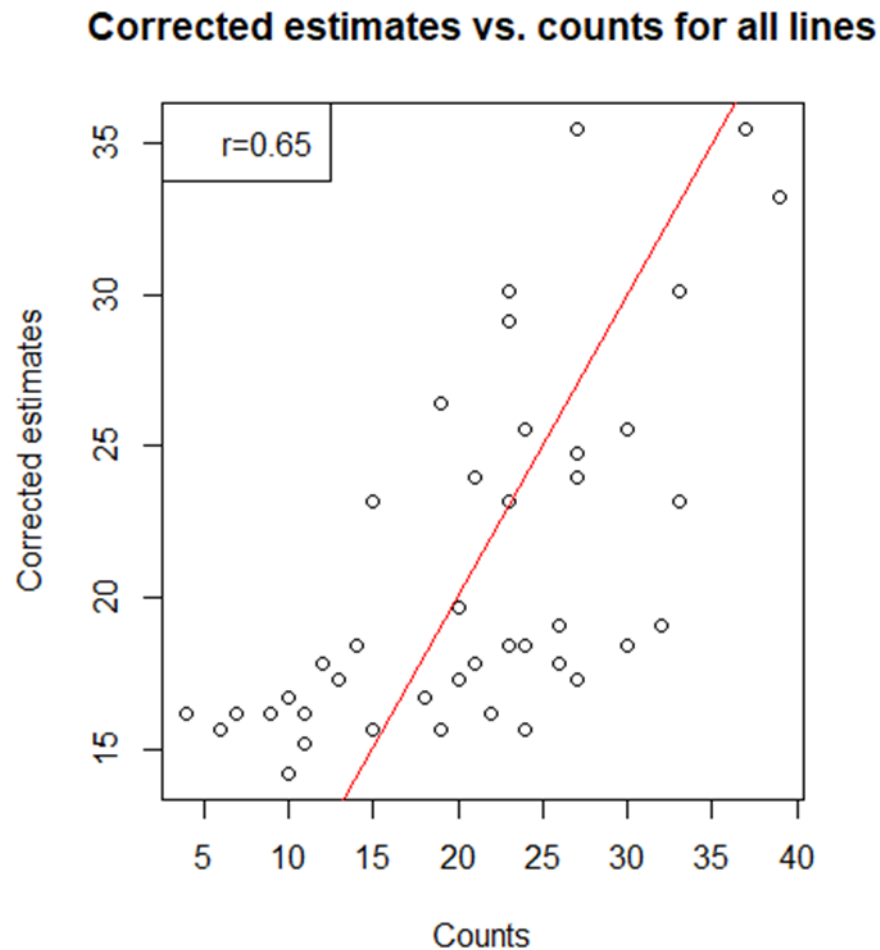


Figure 52: Plot of corrected ridership estimates vs. counts of the enhanced rule-based method.

After correcting the enhanced rule-based method estimates, the plot shows that the corrected values describe and visualize the line of perfect prediction more accurately. Tables 39, 40, 41, 42, and 43 show the mean squared errors (MSE) and the absolute percentage errors (APE) of the Blueline, Greenline, Orangeline, Redline, and Yellowline respectively. The corrected estimates were rounded down to the closest integer to minimize the MSE of the estimates.

Table 39: MSE and APE of the Blueline after implementation of enhanced rule-based method

Date	Passenger	Estimate	MSE	APE	Estimate*	MSE*	APE*
1/14/2019	30	12	324	-60%	18	144	-40%
1/15/2019	32	13	361	-59%	19	169	-41%
1/16/2019	27	20	49	-26%	23	16	-15%
1/17/2019	30	22	64	-27%	25	25	-17%
1/18/2019	27	32	25	19%	35	64	30%
10/10/2018	23	12	121	-48%	18	25	-22%
10/11/2018	14	12	4	-14%	18	16	29%
10/12/2018	39	30	81	-23%	33	36	-15%
10/18/2018	7	8	1	14%	16	81	129%

* Indicates values after Poisson correction

Table 40: MSE and APE of the Greenline after implementation of enhanced rule-based method

Date	Passenger	Estimate	MSE	APE	Estimate*	MSE*	APE*
2/4/2019	23	19	16	-17%	23	0	0%
2/5/2019	15	7	64	-53%	15	0	0%
2/6/2019	12	11	1	-8%	17	25	42%
2/7/2019	10	9	1	-10%	16	36	60%
2/8/2019	4	8	16	100%	16	144	300%
10/17/2018	21	11	100	-48%	17	16	-19%
10/18/2018	13	10	9	-23%	17	16	31%
10/19/2019	11	8	9	-27%	16	25	45%
10/19/2019	19	23	16	21%	26	49	37%

* Indicates values after Poisson correction

Table 41: MSE and APE of the Orangeline after implementation of enhanced rule-based method

Date	Passenger	Estimate	MSE	APE	Estimate*	MSE*	APE*
1/7/2019	6	7	1	17%	15	81	150%
1/8/2019	24	7	289	-71%	15	81	-38%
1/9/2019	13	10	9	-23%	17	16	31%
1/10/2019	9	8	1	-11%	16	49	78%
1/11/2019	10	4	36	-60%	14	16	40%
10/10/2018	11	6	25	-45%	15	16	36%

10/18/2018	20	10	100	-50%	17	9	-15%
------------	----	----	-----	------	----	---	------

* Indicates values after Poisson correction

Table 42: MSE and APE of the Redline after implementation of enhanced rule-based method

Date	Passenger	Estimate	MSE	APE	Estimate*	MSE*	APE*
1/7/2019	23	26	9	13%	29	36	26%
1/8/2019	37	32	25	-14%	35	4	-5%
1/9/2019	33	27	36	-18%	30	9	-9%
1/10/2019	23	27	16	17%	30	49	30%
1/11/2019	33	19	196	-42%	23	100	-30%
10/16/2018	27	20	49	-26%	23	16	-15%
10/17/2018	15	19	16	27%	23	64	53%
10/17/2018	24	22	4	-8%	25	1	4%
10/18/2018	21	20	1	-5%	23	4	10%
10/19/2018	24	12	144	-50%	18	36	-25%

* Indicates values after Poisson correction

Table 43: MSE and APE of the Yellowline after implementation of enhanced rule-based method

Date	Passenger	Estimate	MSE	APE	Estimate*	MSE*	APE*
1/14/2019	27	10	289	-63%	17	100	-37%
1/15/2019	26	11	225	-58%	17	81	-35%
1/16/2019	21	11	100	-48%	17	16	-19%
1/17/2019	27	21	36	-22%	24	9	-11%
1/18/2019	20	14	36	-30%	19	1	-5%
10/10/2018	26	13	169	-50%	19	49	-27%
10/18/2018	18	9	81	-50%	16	4	-11%
10/18/2018	19	7	144	-63%	15	16	-21%
10/19/2018	22	8	196	-64%	16	36	-27%

* Indicates values after Poisson correction

On average, the MSE is 79.4 before correction. After correction, the MSE is reduced to 40.6. This is expected because the Poisson regression minimizes the errors of the estimated and ground truth values. The APE calculated are -25% and 15% before and after correction, respectively. This means that before correction there is an underestimation, and after correction, there is an overestimation. The overestimation is smaller in magnitude for the values after correction, which is also implicitly stated by the MSE. A graphical representation of the results for the number of passengers counted, estimated, and corrected are shown in Figures 53, 54, 55, 56, and 57 for each of the five lines, respectively.

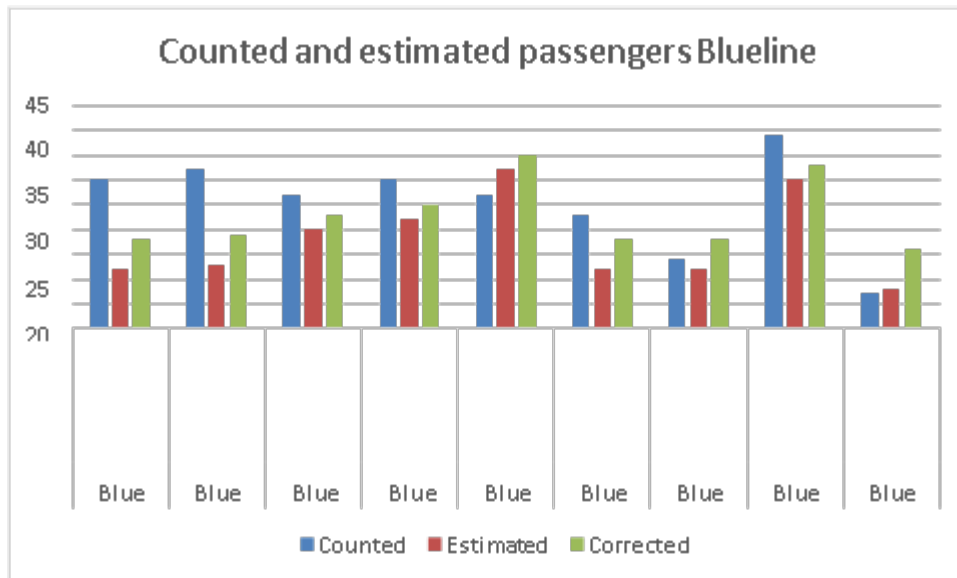


Figure 53: Counted, estimated, and corrected passengers with the enhanced rule-based method of the Blueline.

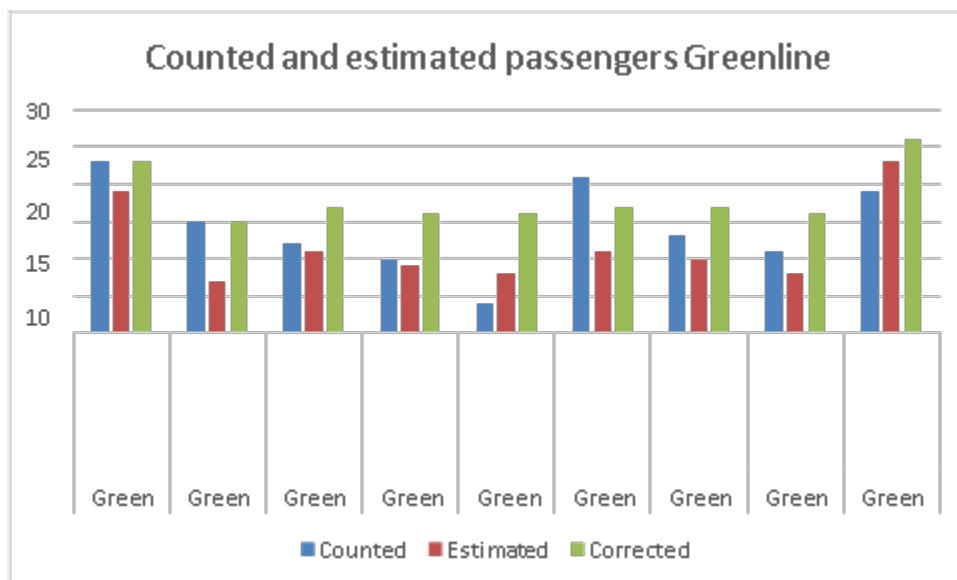


Figure 54: Counted, estimated, and corrected passengers with the enhanced rule-based method of the Greenline.

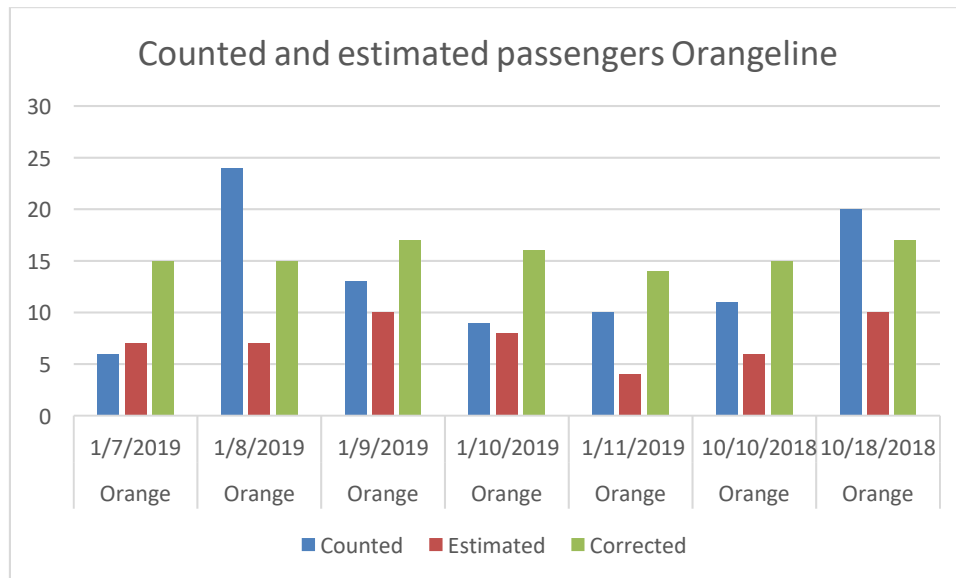


Figure 55: Counted, estimated, and corrected passengers with the enhanced rule-based method of the Orangeline.

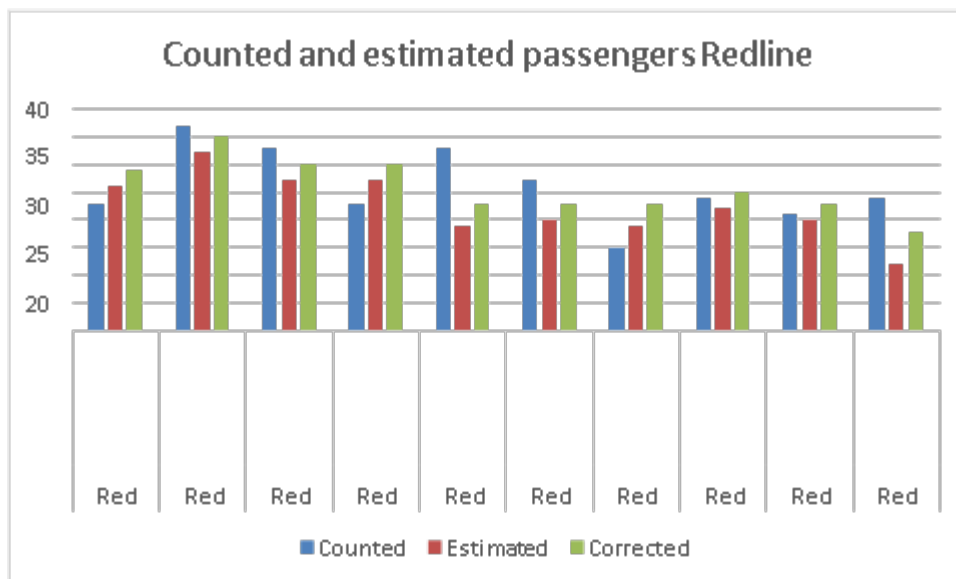


Figure 56: Counted, estimated, and corrected passengers with the enhanced rule-based method of the Redline.

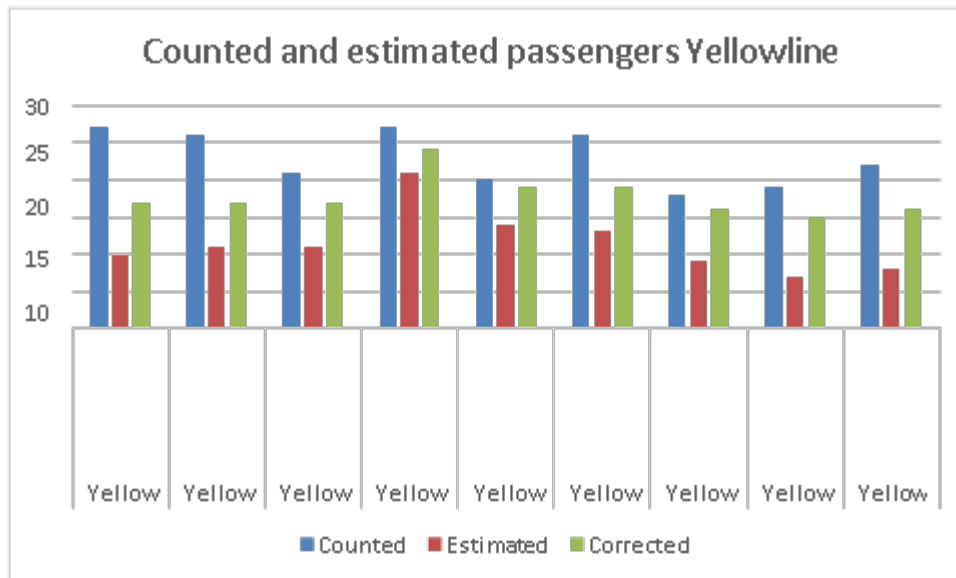


Figure 57: Counted, estimated, and corrected passengers with the enhanced rule-based method of the Yellowline.

It can be noted that for the Blueline there is one day (10/08/2018) not considered and for the Orangeline there are two days not considered (10/08/2018 and 10/10/2018). This was because the GPS and Wi-Fi data could not be joined together by the code. This may be due to some glitches that would have to be fixed for future implementation of such code. For the case of the rule-based method, the two datasets were not joined together. Therefore, those days are included in the model.

Furthermore, it can be noted that the Orangeline data for 10/10/2018 was included with this algorithm. This was because the enhanced rule-based method uses the location of the networks detected and the bus stops in order to estimate the number of passengers. Therefore, even when the Smart Station collected data for a longer period, those devices that were detected and did not belong to passengers using the Orangeline route were discarded.

5.1.3. Unsupervised Machine Learning

Although it is standard, the data were not scaled because the best results of data clustering were obtained without scaling the data. This could be due to some data loss when the data is scaled or the homogenization of all the networks that would not permit classification of riders and non-riders. Nevertheless, clustering methods are exploratory and data visualization helps researchers make better decisions. For all the datasets, a k value of two was used. This was confirmed by the elbow, silhouette, and gap statistic methods when tested on all the datasets. They suggested two centroids for most of the datasets as shown in Appendix D.

Figure 58 shows the results of the K-means clustering method on the Wi-Fi data for the Blueline on January 14, 2019. It can be observed that the vast majority of the data points belong to one cluster and only a small percentage belongs to the other cluster. For this case, 817 observations

(98%) belong to the first cluster, and 17 observations (2%) belong to the second cluster. The fact that one cluster yields a small value, and this is the trend for all the datasets, may indicate a classification that might be due to some specific networks. If these numbers are correlated with the total number of people counted, there might be an indication that the clustering method classifies the riders.

For this dataset, the percentage of variance in the data explained by the clusters is 86.5%. This is a strong explanation because it is closer to 100% than 50% even with only two clusters. Adding more clusters would increase this value, but the increase will not be significant enough to be necessary. The first two components explain 76.7% of the variance of the original data. This means that more than three-fourths of the information of the five variables can be described by two dimensions only.



Figure 58: Cluster plot of the Blueline Wi-Fi data on January 14, 2019

Figure 59 shows the elbow method performed on the dataset of the Blueline on January 14, 2019. For this same data set, the silhouette method (Figure 60) and the gap statistic method (Figure 61) were also implemented. All these methods provided a k value of two, as in most of the datasets. In order to have a standardized process for all the datasets, the same number of clusters was applied.

The variance explained by clusters and the first two principal components was similar for all the datasets.

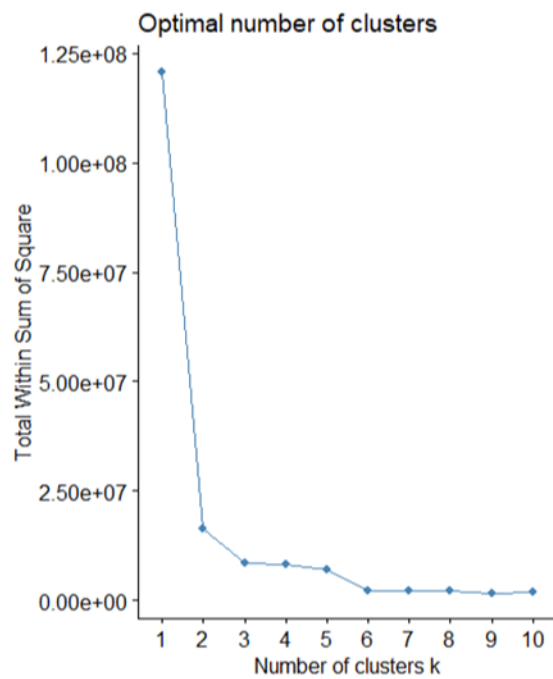


Figure 59: Elbow method for determination of number of clusters

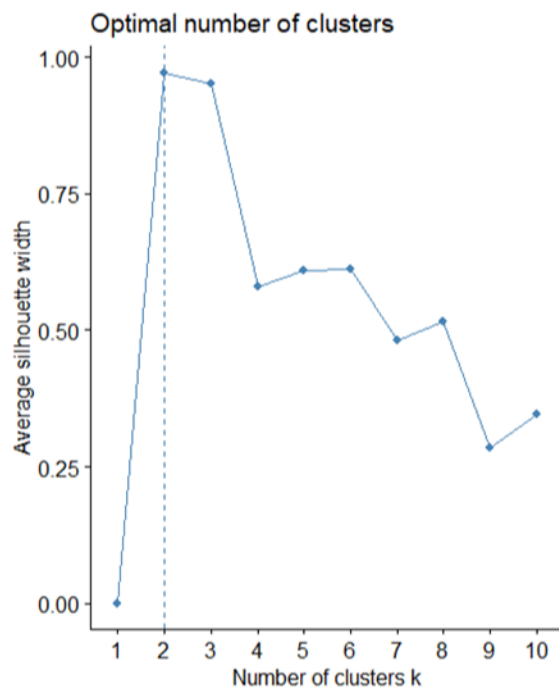
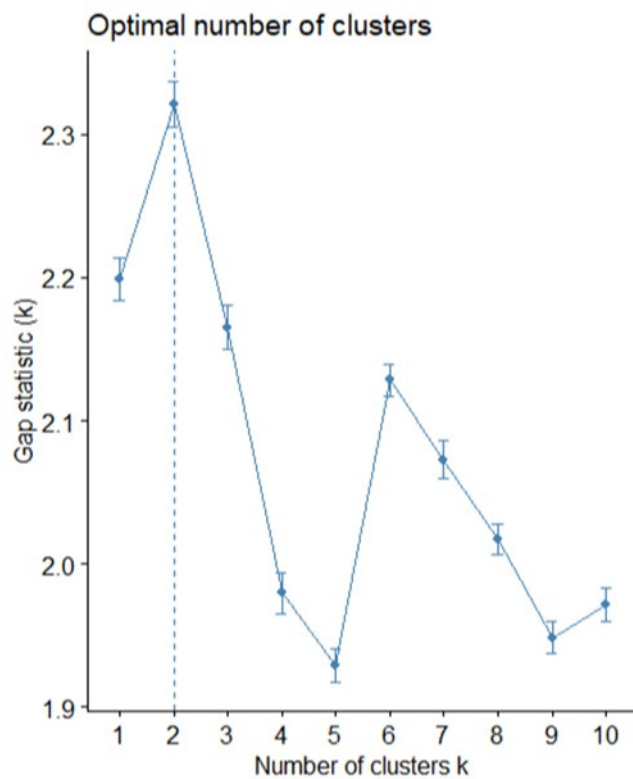


Figure 60: Silhouette method for determination of number of clusters**Figure 61: Gap statistic method for determination of number of clusters**

Tables 44, 45, 46, 47, and 48 show the results of implementing the K-means algorithm with two centroids on the BlueLine, Greenline, Orangeline, Redline, and Yellowline respectively. In addition, the explanation of the variance, known as R-squared, by the clustering algorithm is provided. On average, the first cluster condenses 98.2% of all the data points. The second cluster agglomerates a mean of around 1.8% of all the data points. The average R-squared of the models is 85%, which means that around the same percentage of total variance is explained by the clusters when only a k of two is used. On average, the first two components describe 77% of the original data, which means that more than three-fourths of the variance can be explained by two dimensions alone.

Table 44: Results of cluster analysis on the BlueLine datasets

Date	Cluster 1	Cluster 2	R-squared	PCA explanation
1/14/2019	817	17	86.5%	76.7%
1/15/2019	685	21	87.1%	76.1%
1/16/2019	789	39	85.7%	76.7%
1/17/2019	1019	27	88.5%	77.0%
1/18/2019	1236	47	91.2%	77.0%

10/8/2018	952	8	72.2%	77.4%
10/10/2018	903	11	79.9%	76.8%
10/11/2018	619	5	82.5%	76.8%
10/12/2018	1335	9	96.8%	76.0%
10/18/2018	707	13	93.0%	78.6%

Table 45: Results of cluster analysis on the Greenline datasets

Date	Cluster 1	Cluster 2	R-squared	PCA explanation
2/4/2019	1034	34	89.0%	77.8%
2/5/2019	1233	30	88.6%	79.0%
2/6/2019	779	18	88.6%	77.5%
2/7/2019	803	21	91.3%	76.5%
2/8/2019	843	35	90.5%	77.7%
10/17/2018	1126	22	84.4%	77.0%
10/18/2018	1122	12	86.3%	77.4%
10/19/2019	496	5	97.3%	76.5%
10/19/2019	1325	16	80.7%	78.2%

Table 46: Results of cluster analysis on the Orangeline datasets

Date	Cluster 1	Cluster 2	R-squared	PCA explanation
1/7/2019	474	13	96.6%	77.6%
1/8/2019	534	4	84.7%	76.3%
1/9/2019	472	8	98.7%	74.1%
1/10/2019	560	6	85.2%	76.5%
1/11/2019	391	4	94.7%	77.6%
10/8/2018	390	3	71.3%	77.3%
10/10/2018	284	3	53.1%	78.5%
10/18/2018	704	8	72.5%	76.6%

Table 47: Results of cluster analysis on the Redline datasets

Date	Cluster 1	Cluster 2	R-squared	PCA explanation
1/7/2019	741	7	82.5%	76.0%
1/8/2019	843	12	82.2%	75.8%
1/9/2019	732	10	75.5%	76.5%
1/10/2019	968	21	80.2%	76.3%
1/11/2019	1415	22	83.3%	77.5%
10/16/2018	1371	14	89.3%	75.5%
10/17/2018	1033	11	89.8%	76.1%
10/17/2018	898	6	79.3%	77.5%
10/18/2018	1034	12	91.1%	76.3%

10/19/2018	1107	8	83.9%	77.3%
------------	------	---	-------	-------

Table 48: Results of cluster analysis on the Yellowline datasets

Date	Cluster 1	Cluster 2	R-squared	PCA explanation
1/14/2019	549	15	67.7%	75.9%
1/15/2019	849	21	80.9%	76.2%
1/16/2019	1229	29	76.0%	76.6%
1/17/2019	745	29	82.9%	76.7%
1/18/2019	883	14	89.6%	77.2%
10/10/2018	490	16	78.7%	76.2%
10/18/2018	563	7	84.9%	77.3%
10/18/2018	939	11	95.6%	77.6%
10/19/2018	803	16	84.9%	76.2%

The fact that one cluster dominates the classification over the other is beneficial for this research because the cluster with fewer data points yields numbers that are not unrealistic for ridership classification. These results are surprising because a more even distribution of points was expected. Nonetheless, the number of points in the second cluster will be considered as estimated riders. Figure 62 shows the dispersion of the estimated ridership versus the counted passengers for all the lines.

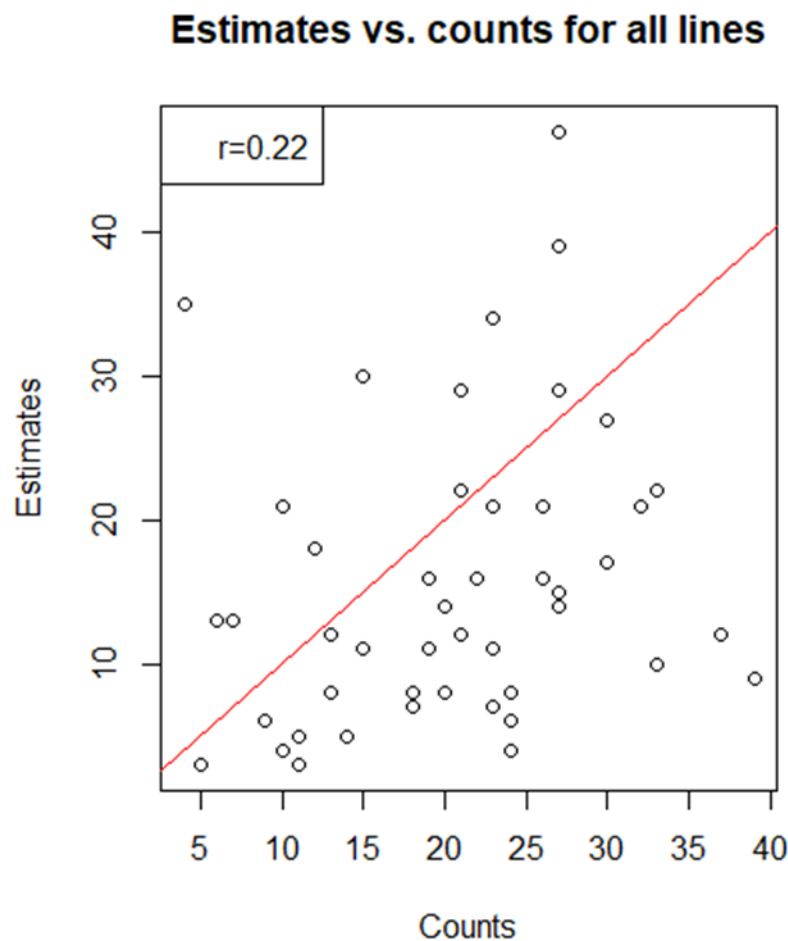


Figure 62: Plot of unsupervised machine learning estimated values vs. counts for all lines

The unsupervised machine learning estimates show a general trend of increasing when the real counts do. This is not so evident for counts that are higher than 35. Overall, there is an underestimation of the estimated values, as can be observed from the plot. A correlation was assumed in order to correct these values and make them closer to the line of perfect prediction. The Poisson regression model was implemented. The coefficients of this regression are shown in Table 49.

Table 49: Coefficients of the Poisson regression for the K-means clustering method

	Estimate	Standard Error	Z value	Pr(> z)
Intercept	2.875643	0.060541	47.499	<2e-16***

Estimated ridership	0.008721	0.003073	2.837	0.00455**
------------------------	----------	----------	-------	-----------

Significance codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’

The intercept is significant at a higher than 0.001 level, and the estimated ridership coefficient is significant at the 0.001 level. Although this value is smaller for this method, it is still statistically significant. Therefore, a Poisson relationship can be established. The estimated model is defined by the following equation:

$$\ln(\hat{p}) = 2.875643 + 0.008721U_i \quad (38)$$

where \hat{p} represents the corrected estimated number of passengers and U_i represents the estimated number of passengers with the unsupervised machine learning algorithm.

The estimated number of passengers were corrected using Equation (38). Figure 63 shows the scatterplot of the corrected estimates and the ground truths. It can be observed that the points follow along with the line of perfect prediction. However, a large error is observed as well. Nevertheless, overdispersion is not present in this model. Therefore, the Poisson model is a proposed model and its coefficients are significant at a very low level. More research could focus on the use of other models in order to explain more of the variability of the data and make them closer to the manual counts.

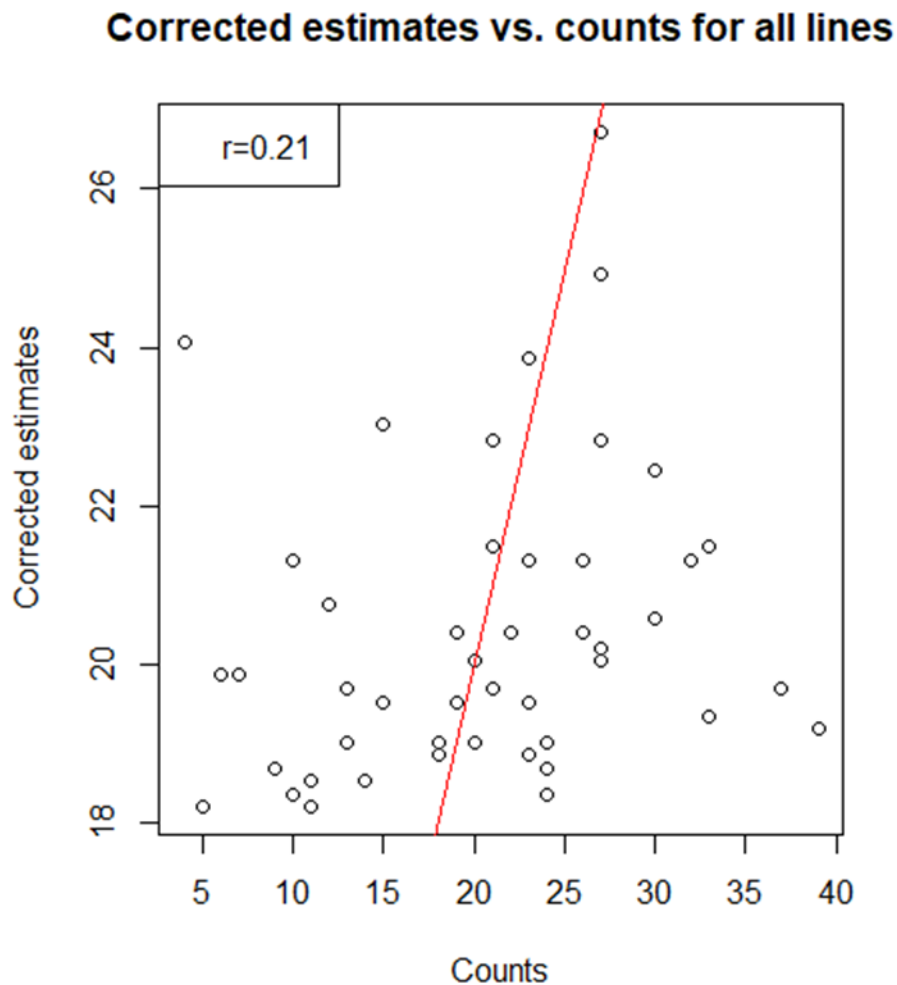


Figure 63: Plot of corrected ridership estimates vs. counts of unsupervised machine learning

The corrected estimated values are shown in Tables 50, 51, 52, 53, and 54 for the Blueline, Greenline, Orangeline, Redline, and Yellowline respectively. Additionally, the mean squared errors (MSE) and the absolute percentage errors (APE) are provided. It is important to note that the clustering algorithms are easy to implement with statistical software; therefore, clusters can be an important tool for data analysis.

Table 50: MSE and APE of the Blueline after implementation of machine learning method

Date	Passenger	Estimate	MSE	APE	Estimate*	MSE*	APE*
1/14/2019	30	17	169	-43%	20	100	-33%
1/15/2019	32	21	121	-34%	21	121	-34%
1/16/2019	27	39	144	44%	24	9	-11%
1/17/2019	30	27	9	-10%	22	64	-27%

1/18/2019	27	47	400	74%	26	1	-4%
10/8/2018	18	8	100	-56%	19	1	6%
10/10/2018	23	11	144	-52%	19	16	-17%
10/11/2018	14	5	81	-64%	18	16	29%
10/12/2018	39	9	900	-77%	19	400	-51%
10/18/2018	7	13	36	86%	19	144	171%

* Indicates values after Poisson correction

Table 51: MSE and APE of the Greenline after implementation of machine learning method

Date	Passenger	Estimate	MSE	APE	Estimate*	MSE*	APE*
2/4/2019	23	34	121	48%	23	0	0%
2/5/2019	15	30	225	100%	23	64	53%
2/6/2019	12	18	36	50%	20	64	67%
2/7/2019	10	21	121	110%	21	121	110%
2/8/2019	4	35	961	775%	24	400	500%
10/17/2018	21	22	1	5%	21	0	0%
10/18/2018	13	12	1	-8%	19	36	46%
10/19/2019	11	5	36	-55%	18	49	64%
10/19/2019	19	16	9	-16%	20	1	5%

* Indicates values after Poisson correction

Table 52: MSE and APE of the Orangeline after implementation of machine learning method

Date	Passenger	Estimate	MSE	APE	Estimate*	MSE*	APE*
1/7/2019	6	13	49	117%	19	169	217%
1/8/2019	24	4	400	-83%	18	36	-25%
1/9/2019	13	8	25	-38%	19	36	46%
1/10/2019	9	6	9	-33%	18	81	100%
1/11/2019	10	4	36	-60%	18	64	80%
10/8/2018	11	3	64	-73%	18	49	64%
10/10/2018	5	3	4	-40%	18	169	260%
10/18/2018	20	8	144	-60%	19	1	-5%

* Indicates values after Poisson correction

Table 53: MSE and APE of the Redline after implementation of machine learning method

Date	Passenger	Estimate	MSE	APE	Estimate*	MSE*	APE*
1/7/2019	23	7	256	-70%	18	25	-22%
1/8/2019	37	12	625	-68%	19	324	-49%
1/9/2019	33	10	529	-70%	19	196	-42%
1/10/2019	23	21	4	-9%	21	4	-9%

1/11/2019	33	22	121	-33%	21	144	-36%
10/16/2018	27	14	169	-48%	20	49	-26%
10/17/2018	15	11	16	-27%	19	16	27%
10/17/2018	24	6	324	-75%	18	36	-25%
10/18/2018	21	12	81	-43%	19	4	-10%
10/19/2018	24	8	256	-67%	19	25	-21%

* Indicates values after Poisson correction

Table 54: MSE and APE of the Yellowline after implementation of machine learning method

Date	Passenger	Estimate	MSE	APE	Estimate*	MSE*	APE*
1/14/2019	27	15	144	-44%	20	49	-26%
1/15/2019	26	21	25	-19%	21	25	-19%
1/16/2019	21	29	64	38%	22	1	5%
1/17/2019	27	29	4	7%	22	25	-19%
1/18/2019	20	14	36	-30%	20	0	0%
10/10/2018	26	16	100	-38%	20	36	-23%
10/18/2018	18	7	121	-61%	18	0	0%
10/18/2018	19	11	64	-42%	19	0	0%
10/19/2018	22	16	36	-27%	20	4	-9%

* Indicates values after Poisson correction

The MSE obtained for the estimates with the machine learning algorithm was 159.2 on average for all the datasets. After Poisson correction, the resulted MSE was 69.0 on average for all the datasets. This provides a reduction of the average distance of the estimates and the line of perfect prediction. In other words, the estimated values are closer (considering the Euclidean distance) to the ground truths, on average. Nevertheless, the average APE values before and after correction are -3% and 28%, respectively, which signifies a reduction in accuracy. The bar charts of the number of passengers counted and estimated for the five lines are shown in Figures 64, 65, 66, 67, and 68.

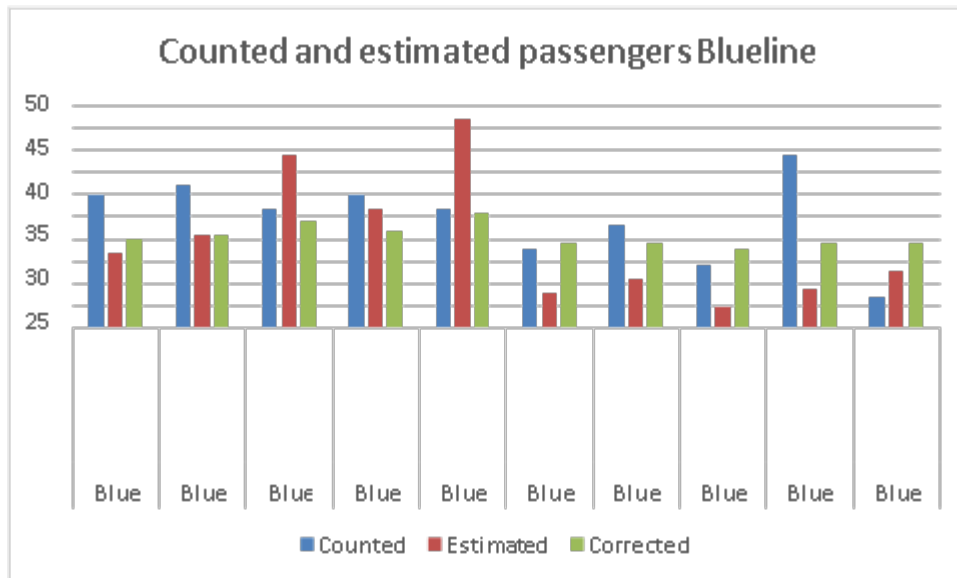


Figure 64: Counted, estimated, and corrected passengers with the clustering method of the Blueline

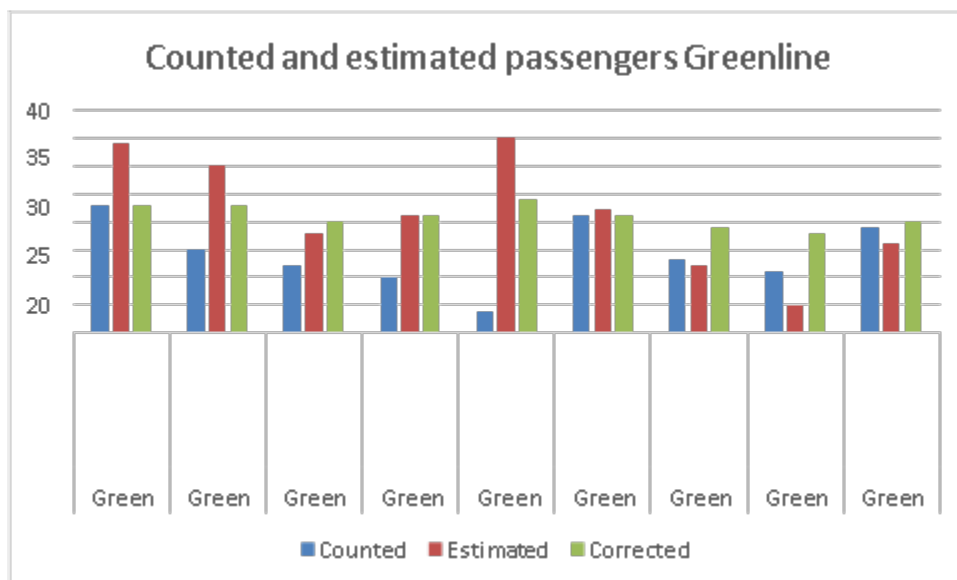


Figure 65: Counted, estimated, and corrected passengers with the clustering method of the Greenline

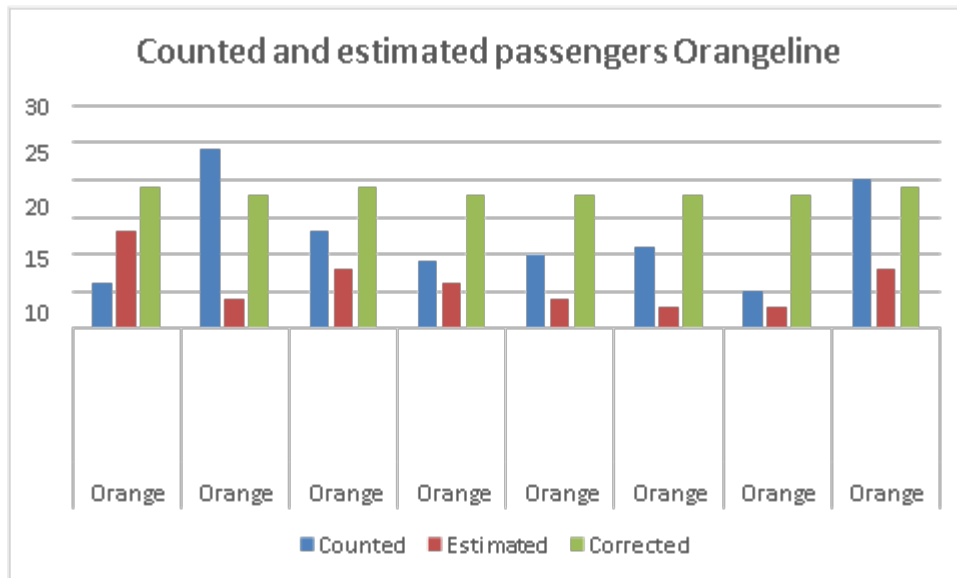


Figure 66: Counted, estimated, and corrected passengers with the clustering method of the Orangeline

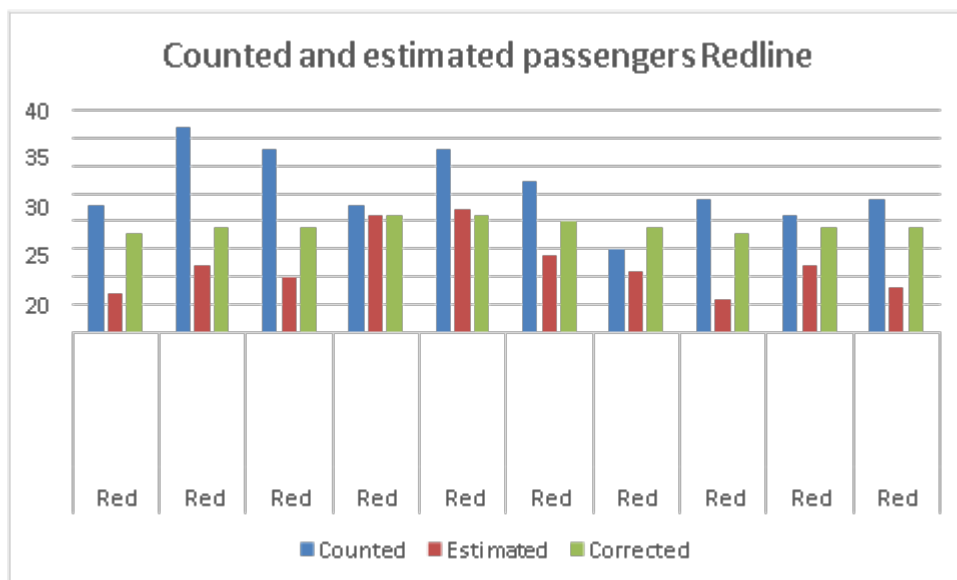


Figure 67: Counted, estimated, and corrected passengers with the clustering method of the Orangeline

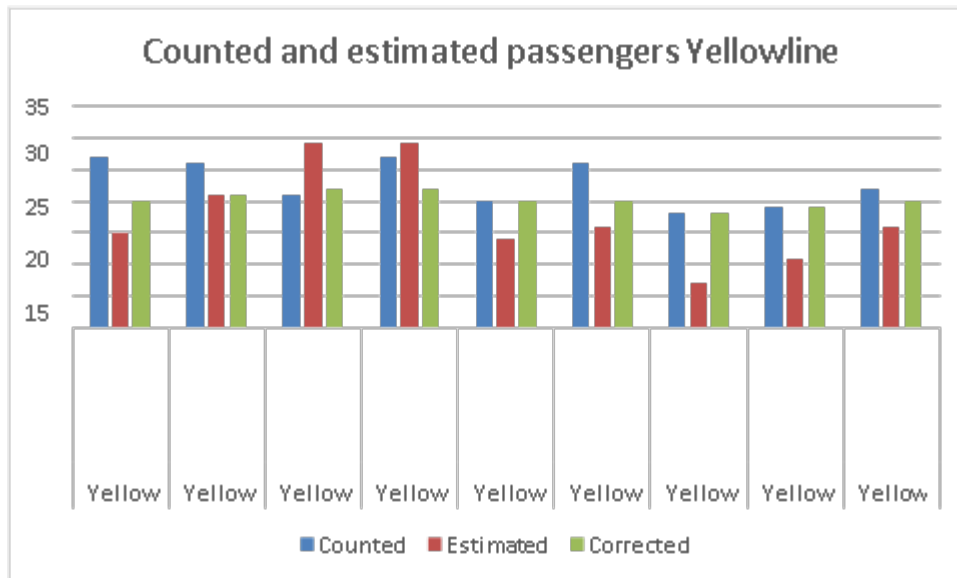


Figure 68: Counted, estimated, and corrected passengers with the clustering method of the Yellowline

The fact that the two indicators, MSE and APE, provide contradicting values is not good for the unsupervised machine learning approach. Clustering methods arrange data in a way that tries to find patterns in the nature of the data itself. The points are grouped without labels. They could be labeled, but this was not the approach of this study.

Future research could focus on labeling passengers and tracking them in the dataset by their MAC addresses in order to know whether they fall into the same cluster and if they are grouped into the cluster with fewer points. If this happens, it would indicate that the methodology used in this research is adequate to establish a correlation between the estimated number of passengers with the actual number of people using a transit system.

The clustering method proved to be less accurate after correction than the rule-based methods. Nonetheless, this research indicates that there is a significant correlation between the estimated number of passengers and the ground truths.

5.1.4. Evaluation of Ridership Estimation Levels

Overall, the three methods performed better than expected. Even when there are instances of under and overestimation, there is a correlation of the estimated ridership and the ground truths. This provides optimism for the use of Wi-Fi signals for ridership estimation. One advantage of using Wi-Fi signals over other methods (e.g., Bluetooth) is that the sample size increases considerably. In this study case, the Smartphone penetration rate was around 95%, which is advantageous for this study.

Luckily, the Smartphone penetration rate keeps increasing in the United States and worldwide. This is also beneficial for the implementation of the Smart Station technology or similar methods that use Wi-Fi signals for ridership estimation. A comparison of the three ridership estimation methods used in this research is presented in Table 55. The general trend observed for all the data

was that when there was an overestimation of riders, the Poisson regression transformed it into underestimation, and vice versa.

The correction decreased the MSE in all cases, which is expected when there is a correlation between the data. The clustering algorithm provided an accuracy higher without correction, which may be due to chance since the MSE is also the largest. The rule-based method seems to be the simplest for implementation and after correction, and it provides more accurate values than the other methods. The Pearson correlation is also shown.

Table 55: Comparison of the accuracy of the ridership estimation methods.

Method	MSE	MSE*	$ \Delta\text{MSE} $	APE	APE*	$ \Delta\text{APE} $	Pearson r
Rule-based	152.8	53.9	98.9	25%	-2%	27%	0.60
Enhanced rule-based	79.4	40.6	38.8	-25%	15%	40%	0.67
Clustering	159.2	69.0	90.2	-3%	28%	31%	0.22

* Indicates values after Poisson correction

5.2. OD Flow Characteristics

The origin-destination matrices were generated by estimating the bus stops that were the closest for the first and last detection times of a specific signal network. Only the networks belonging to passengers that were classified by the enhanced rule-based method were considered. The bus stops were obtained in accordance with the sixth rule. This is a novel methodology because previous research has not used time stamps to infer OD flows.

Every network's predicted boarding and alighting bus stops were introduced in the final OD matrix as one count. Later, all the estimated passengers were added into their corresponding cell. Figures 69, 70, 71, 72, and 73 display the OD matrices for the Blueline, Greenline, Orangeline, Redline, and Yellowline respectively.

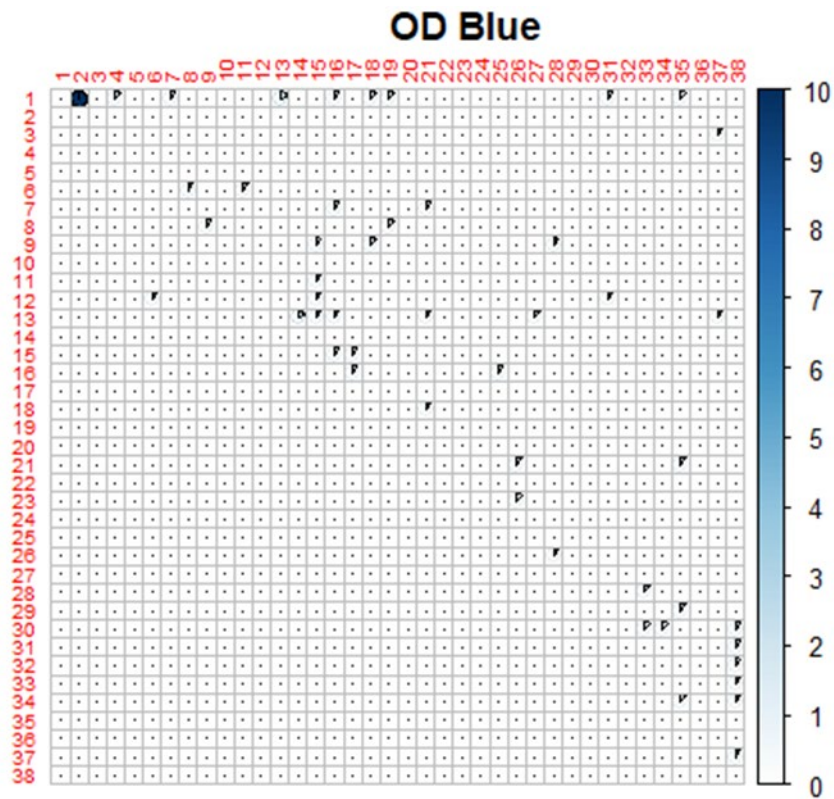


Figure 69: Estimated OD matrix of the Blueline

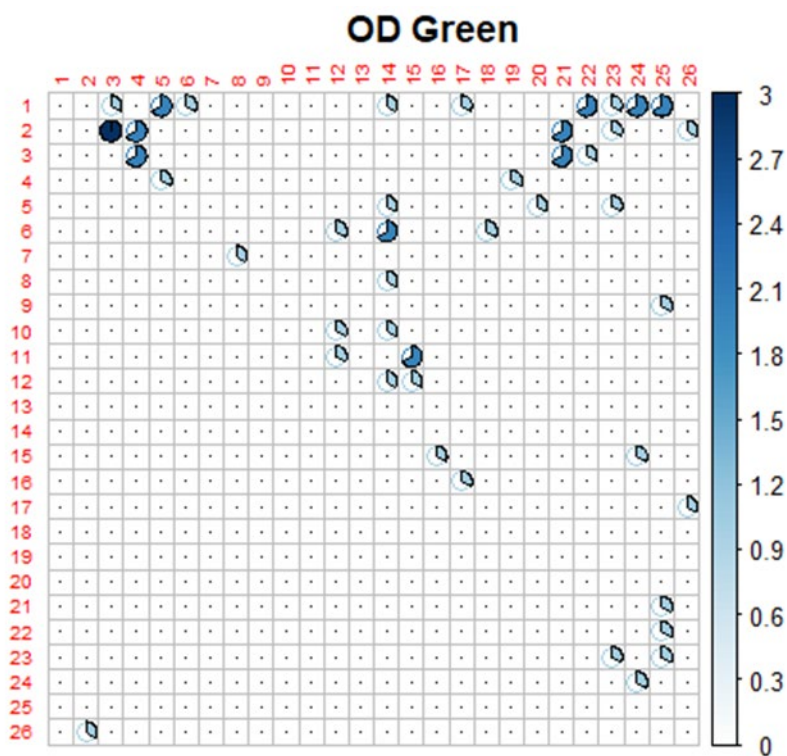


Figure 70: Estimated OD matrix of the Greenline



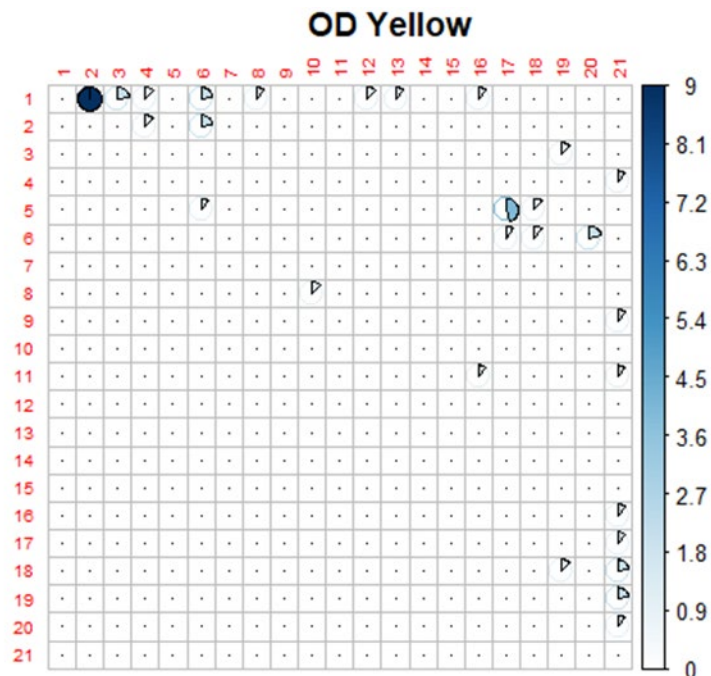


Figure 73: Estimated OD matrix of the Yellowline

For the Blueline, the bus stops that generated more passengers were MSU and Wal- Mart. The bus stops that received more passengers were MSU and 6th & Garfield. Additionally, Oak & 15th, Bridger Peaks Town Center, and Wilson & Curtis were shown to be destinations with a higher number of passengers alighting. Overall, the Blueline OD matrix was successfully identified the MSU and Wal-Mart bus stops as popular destinations. The 6th & Garfield stop is identified as a false positive because, in the surveys and observed data, not many passengers use that bus stop. The algorithm failed to detect the Downtown bus stop as a popular destination. However, passengers were reported to use this bus stop.

For the Greenline, the bus stops that generated more passengers are MSU and the surrounding stops which serve apartment areas. Likewise, the Gallatin Valley Mall stop was a trip generator. Regarding destinations, the bus stop that showed the most passengers was Smith & Missoula. In addition, the bus stops near the MSU bus stop also showed high alighting patterns. The fact that MSU and Smith and Missoula show a higher number of passengers encourages the use of Wi-Fi technology for estimation of OD flows because it is consistent with both the surveys and observed data.

The three most popular bus stops for generating passengers for the Orangeline were MSU, Downtown, and Highland & Main. The bus stops that were estimated to receive more passengers are Grant & Wilson, Bozeman Deaconess, South 7th @ MSU Police, and MSU. This general trend matches the characteristics shown in the surveys and observations on board. It should be noted that the stop at the police station did not register any stop during the manual surveys. It is believed that since the algorithm uses an approximation of time to infer the boarding and alighting bus stops, there may be some errors for bus stops that are close to each other. The police station is part of the campus and is separated by a travel time of around 30 seconds from the main MSU stop. Therefore, those alighting passengers are thought to belong to the MSU bus stop.

In the same way as the manual counts, the MSU stop was the most representative generator of passengers for the Redline. The other frequent generators of passengers were the Gallatin Valley Mall and Downtown bus stops. The 6th & Garfield stop was also another generator of passengers. The three main stop destinations were estimated to be MSU, 6th & Garfield, and Main & Babcock. With less frequency, 8th & Koch, Fowler & Loreda, and Tschache & 27th were estimated to be common destinations. These characteristics are consistent with the surveyed and observed data.

For the Yellowline, the main bus stops that generated passengers were MSU and its surroundings, and the Gallatin Valley Mall. The main destinations are MSU, its surrounding areas, Gallatin Valley Mall, and Huffine & Harmon Stream. The last stop is believed to be an overestimation made because the bus passes by this stop twice in a single loop and it is close to other bus stops.

In general, the OD matrices described patterns that are consistent with the observed movement of passengers. This is surprisingly good for the application of the proposed methodology in this research. There are, nonetheless, several signals that are sporadic, which indicates that they do not belong to passengers and constitute noise. Nevertheless, when only realistic values are considered, a characterization of the real OD patterns of passengers has been established by this methodology.

5.3. Wait Time

In general, the wait time of passengers is a value that is hard to observe in the field. It is particularly hard to estimate when several people stand by or walk near a bus stop. Surveys can be implemented to ask people their wait time; however, they provide a subjective, perceived time. This study proposes a new methodology to infer wait time using rule-based algorithms on Wi-Fi data obtained from Smart Stations located at the bus stop and on transit vehicles. The results of the experiments to determine if the Raspberry Pi provides an unbiased estimation are presented first. Secondly, the results of the evaluation of the estimate wait time are shown for a study case as explained in the Methodology chapter.

5.3.1. Experiment to Validate Detection Time

As expected, Apple devices randomized their MAC addresses which made the estimation of detection time unfeasible. Therefore, only the LG and Samsung devices were considered in the model. Figure 74 shows the estimated times of detection by device type. The line of perfect prediction is the desired value that the Smart Station should yield in order to have accurate estimates. It is shown that all the estimates follow the line of perfect prediction, which is indicative, at least visually, that the estimates are unbiased. The three clusters that are seen represent the three different times that were tested in this experiment. The times were one, three, and five minutes.

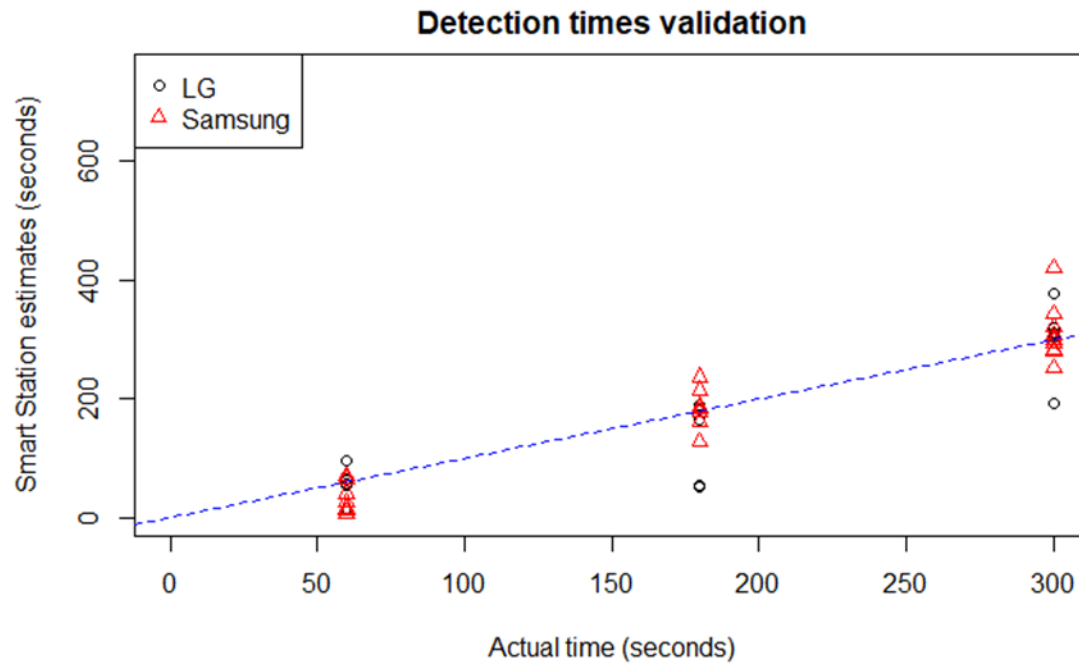


Figure 74: Smart station detection time versus actual time devices were detectable

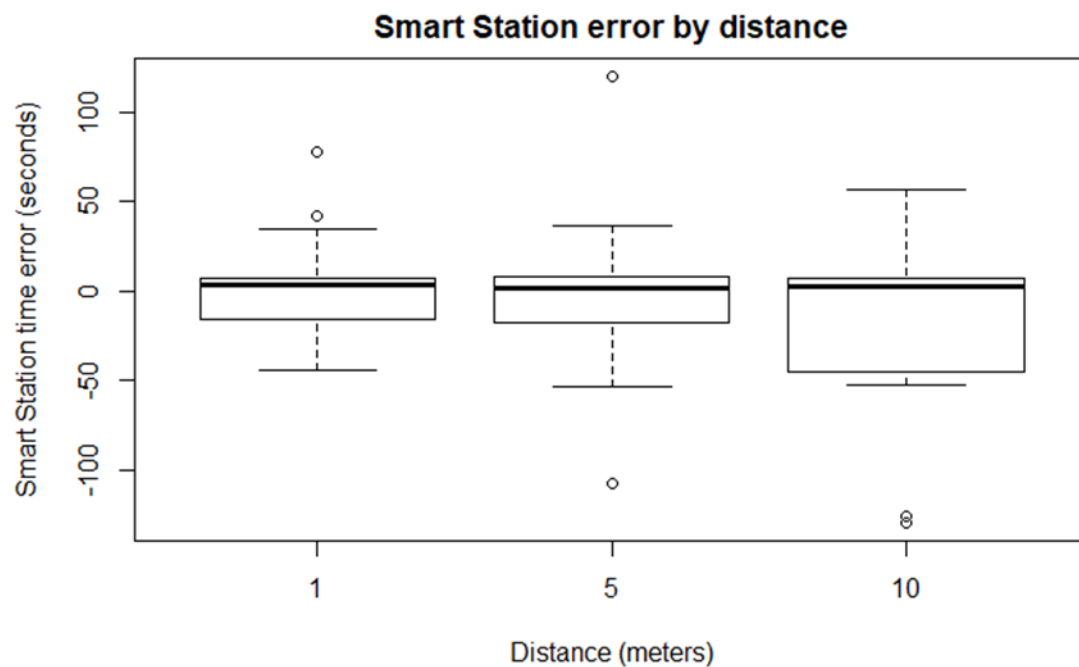


Figure 75: Boxplots of Smart Station time error by distance

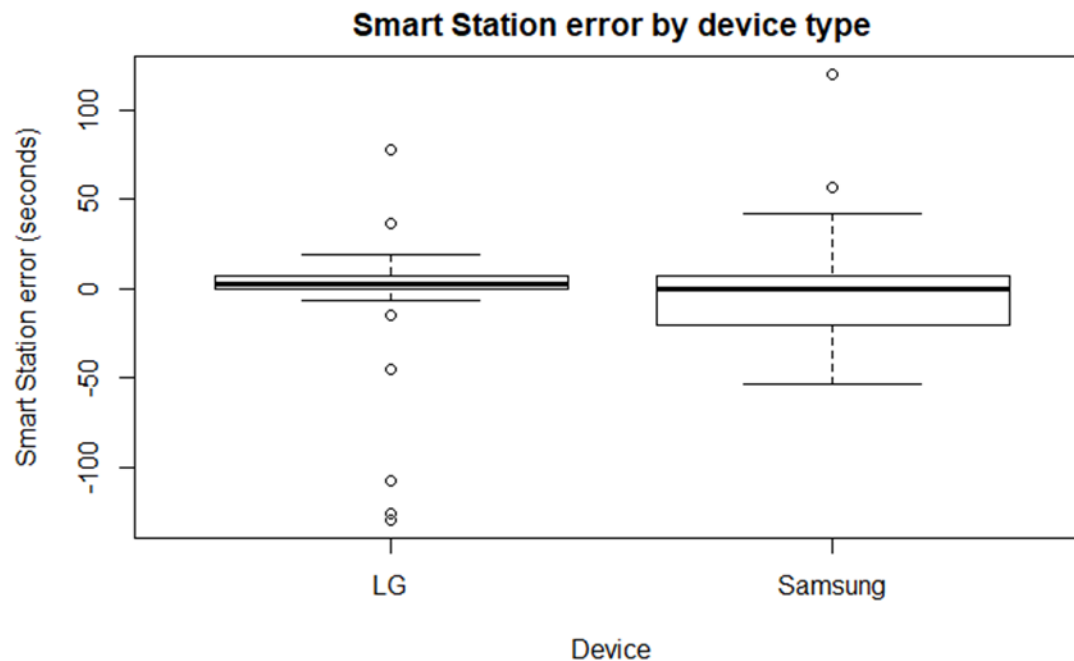


Figure 76: Boxplots of Smart Station time error by device type

Figure 75 shows that the Smart Station error is around zero and the boxplots of all the distances overlap. This indicates that, on average, the error is the same regardless of the distance. Figure 76 exhibits the boxplots of the errors by the device manufacturer. The two boxplots overlap which suggests that the errors by device type are the same on average. Outliers are visible in all the graphs. The error is obtained by subtracting the Smart Station estimate and the actual time the devices had their Wi-Fi turned on. The ideal value of error should be zero. Not only was the error tested statistically to determine if it was different from zero, but the influence of other variables on the error was tested. These variables were the device manufacturer, the distance between the device and the Smart Station, and the time of detection.

To do this, the influence of the variables on the error was modeled with multiple linear regression. Initially, a complex model with three-way interaction was executed. The results of the coefficients of this model are shown in Table 56. In the most complicated model, it can be noted that there is little to no evidence that there is any interaction of the variables. Therefore, the interaction was discarded from the model.

Table 56: Coefficients of the most complicated model (response variable: error)

Coefficients	Estimate	Standard error	t value	Pr(> t)
Intercept	1.55E+01	2.89E+01	0.537	0.594
Actual time	3.92E-03	1.41E-01	0.028	0.978
Distance	-4.96E+00	4.46E+00	-1.112	0.272
Device	-4.17E+01	4.09E+01	-1.019	0.314
Actual time:Distance	2.39E-03	2.18E-02	0.11	0.913

Actual time:Device	1.14E-01	2.00E-01	0.572	0.57
Distance:Device	4.84E+00	6.31E+00	0.772	0.444
Actual time:Distance:Device	9.90E-04	3.08E-02	0.032	0.974

The model without interactions was also implemented. The results of this model are presented in Table 57. There is little to no evidence that the error is affected by the device type (one-sided p-value: 0.603). Therefore, the device variable was excluded from the model and the other variables were tested. The model without the device variable is shown in Table 58.

Table 57: Coefficients of the model with no interactions (response variable: error)

Coefficients	Estimate	Standard error	t value	Pr(> t)
Intercept	-1.10E+01	1.52E+01	-0.724	0.473
Actual time	7.64E-02	5.67E-02	1.348	0.184
Distance	-2.01E+00	1.51E+00	-1.33	0.190
Device:Samsung	5.81E+00	1.11E+01	0.524	0.603

Table 58: Coefficients of the model without device as predictor (response variable: error)

Coefficients	Estimate	Standard error	t value	Pr(> t)
Intercept	-8.08E+00	1.40E+01	-0.576	0.567
Actual time	7.64E-02	5.63E-02	1.358	0.180
Distance	-2.01E+00	1.50E+00	-1.34	0.186

Table 59: Coefficients of the model with actual time as the predictor (response variable: error)

Coefficients	Estimate	Standard error	t value	Pr(> t)
Intercept	-1.88E+01	1.16E+01	-1.616	0.112
Actual time	7.64E-02	5.67E-02	1.348	0.184

There is little to no evidence that actual time has an impact on the time error (one-sided p-value: 0.184). In summary, the device manufacturers tested (LG and Samsung), the distance from the Smart Station (within ten meters), or the time a device is within the range of detection, do not affect the accuracy of the detection time. This is a tremendously important factor that encourages the use of this technology.

On average, the error detected was around -5.0 seconds. However, the 95% confidence interval was -16.2 to 6.2 seconds. In addition, the true mean does not seem to be different from zero (one sample t-test, p-value: 0.3738). Therefore, the Smart Station can be used to estimate the detection time. This is a noteworthy statement for this research because it demonstrates that the detection times estimated are unbiased and can be used for OD flows which are time-dependent. The MSE was 1,685 for the model without correction. Since the mean error does not differ from zero, a correction was not considered.

5.3.2. Experiment to Estimate Wait Time

A t-test was performed on the two groups: the observed wait times and the estimated wait times from the Smart Stations. The observed wait times were obtained by measuring the total time people waited for the bus. The estimated wait times were obtained by implementing the rule-based algorithm on the two Wi-Fi datasets, one from the bus stop and one for the SS on the bus. These values were obtained as explained in chapter four. The analyzed stop was Wal-Mart of the Blueline. Figure 77 displays the histograms of the two groups.

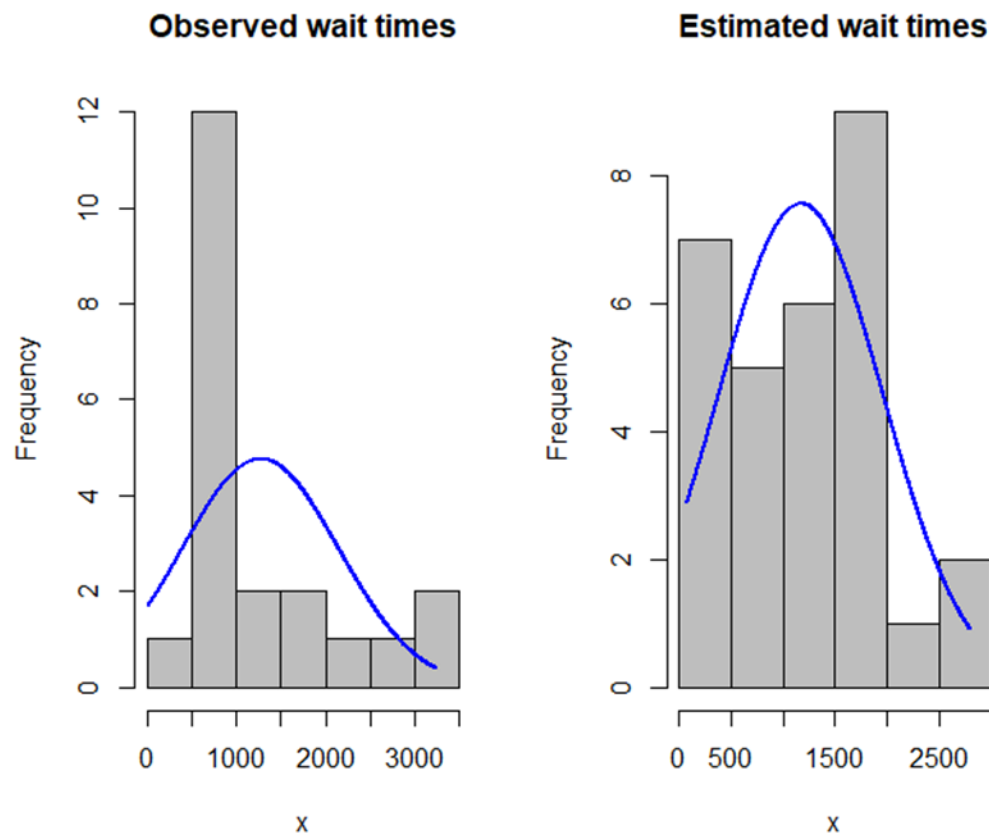


Figure 77: Histograms of the observed and estimated wait times

Figure 78 shows the average of the two groups. It can be noted that the SS provided a smaller value. This is because the SS group has more points. These extra observations may be false positives. In other words, false positives are signals that are considered passengers but are not.

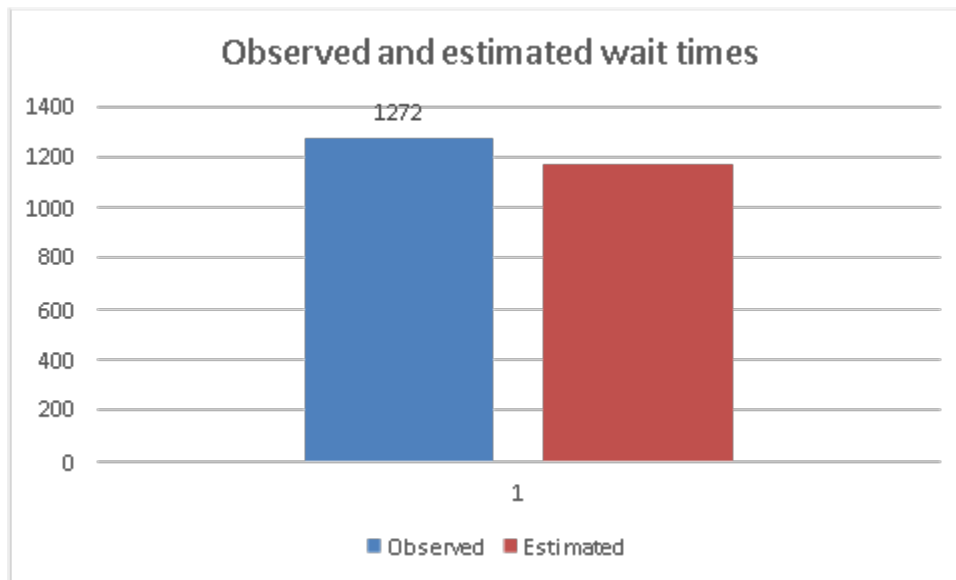


Figure 78: Mean observed and estimated wait time (seconds)

The observed time has a mean of 1,272 seconds ($n=21$) and the estimated wait time has a mean of 1,171 seconds ($n=30$). However, there is no evidence that they are statistically different. A two-sample t-test indicated that there is no difference between the two methods in estimating the wait time (p -value: 0.6778, 95% CI: -385.0 to 586.1 seconds). These results coincide with the previous experiments made on the estimation of the detected time by the Raspberry Pi that was statistically similar to the actual time a device had its Wi-Fi turned on in the detection rate. These results suggest that the Smart Stations can be used to estimate an objective wait time in a cheap and efficient manner.

5.4. Travel Times

A paired t-test was performed on the difference of the manual travel times between bus stops and their estimated values using the Smart Station. The difference was obtained in order to account for lack of dependence of the two groups. The estimated travel times were obtained by implementing the rule-based algorithm developed for this purpose as explained in chapter four. The difference was obtained for all the lines and all their bus stops. Figure 80 shows the difference between the two groups and the density histogram.

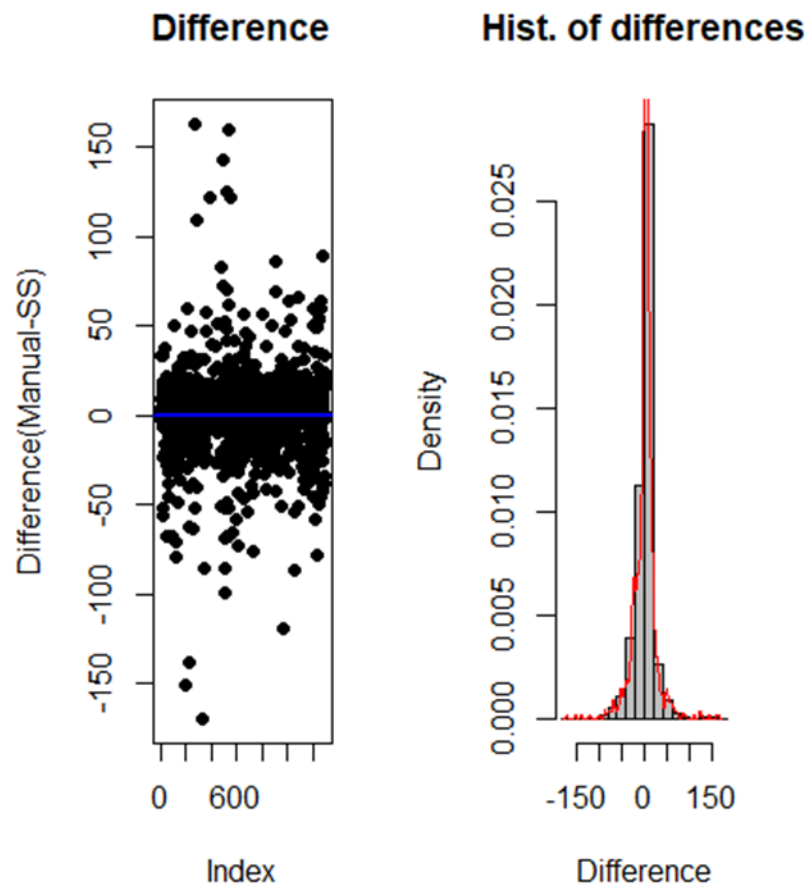


Figure 79: Dispersion and probability histogram of the travel time differences (n=1,292)

From Figure 79, it is noted that the calculated differences centered around the zero- difference line. This is a visual indication that the calculated difference tends to zero, on average. There are some outliers, however, the majority of points seem to have a value closer to zero. In addition, the histogram suggests that the differences follow a normal distribution. Although this is an assumption of the paired t-test method, the large sample size would not require a normal distribution due to the central limit theorem. Figure 80 shows a symmetric and normal distribution of the differences with more clarity.

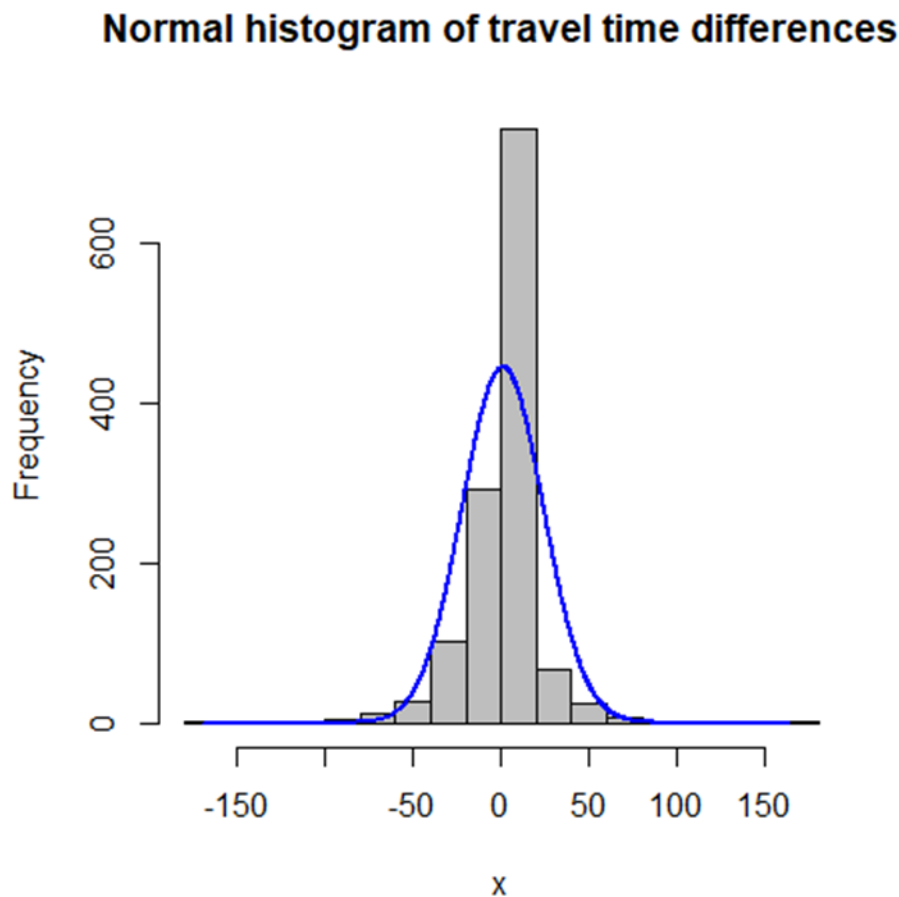


Figure 80: Histogram of the travel time differences (n=1,292)

The mean travel time difference has a value of 1.2 seconds. Nevertheless, there is little evidence to suggest that it is statistically different from zero at the 95 percent confidence level. A paired t-test indicated that the difference could be zero (p-value: 0.06638, 95% CI: -0.1 to 2.4 seconds). These results demonstrate that the algorithm implemented on the GPS data is effective in estimating travel times of a transit vehicle. Furthermore, if there were a significant difference, it would not be enough in magnitude to represent a miscalculation. In other words, the error of fewer than three seconds that could be generated is tolerated for transportation studies.

It was expected that there would be more variability in the results because the algorithm could not perfectly simulate the way the data was collected, because buses stop at varying distances from the location of the bus stop that is presented on the map. In addition to this, the city does not provide separate areas for bus stops; hence, the buses forcedly interact with other vehicles, which increases the stochasticity of the buses' behavior.

In addition to determining if the transit vehicles' travel times could be estimated, the difference between other groups was tested. As mentioned in chapter four, the differences were controlled by lines, days of the week, peak hours, and periods of days (AM and PM). To do this, a four-way

analysis of variance was implemented. The results of this analysis are shown in Table 60. Since the results suggest that there could be a difference between some lines and the peak and off-peak hours, the Tukey HSD plot is shown in Figure 81 in order to estimate a 95 percent confidence interval of the difference of all the assemblies within the four groups.

Table 60: Analysis of variance table (response: travel time difference)

Source	D. of freedom	Sum of sq.	Mean sq.	F-value	Pr(>F)
Lines	4	5,092	1,272.96	2.383	0.04966*
Days	4	759	189.82	0.3553	0.84041
Times	1	187	186.8	0.3497	0.55439
Peak Hours	1	1927	1,927.2	3.6078	0.05773·
Residuals	1,281	684,283	534.18		

Significance codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’

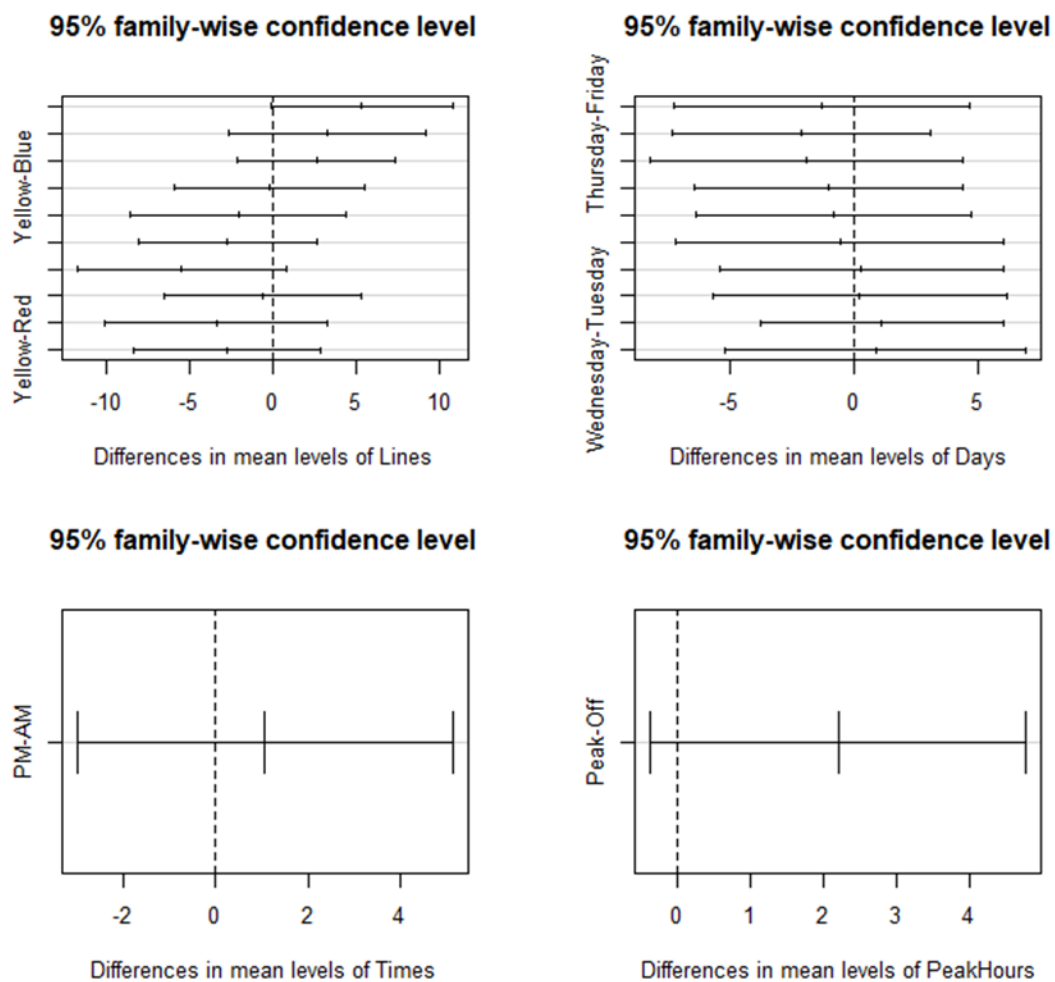


Figure 81: TukeyHSD plot for all groups

Although the analysis of variance indicates that there might be some difference between the different bus lines and the peak and off-peak hours, the Tukey HSD test shows that all the confidence intervals of the differences between every group fall between negative and positive values. Therefore, it is concluded that the travel times estimated with the algorithm run on the Smart Station data are statistically similar to the travel times measured in the field. This means that the Smart Stations can effectively be used for travel time estimation while they are tracking passengers on board a transit vehicle.

The model seems to have heavy outliers. Therefore, an analysis of the influence of the outliers was made as shown in Figure 82. It is noted that although the standardized residuals are high, the leverages and Cook's distances are very low. The outliers may be affecting the accuracy of the model, but they do not affect the results significantly. An improvement of the algorithm should focus on decreasing the number of heavy outliers to have more reliable results. This improvement could provide better-estimated values before implementing this methodology in the industry.

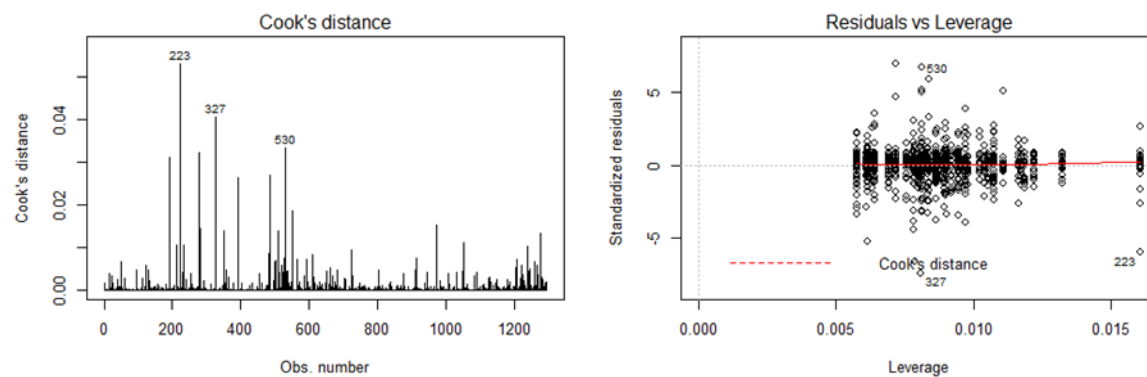


Figure 82: Plot of Cooks' distance standardized residuals vs leverage

6. CONCLUSIONS

This study introduces the use of a wireless Wi-Fi scanning device for transit data collection. Field experiments were designed and conducted in order to obtain ground truths of the current passenger demands of the chosen study area, understand the nature of the Wi-Fi and GPS data that were collected, and implement and evaluate algorithms to estimate transportation data. The estimated transportation characteristics were ridership, OD flows, passengers' wait time, and transit vehicle's travel times.

This innovative device, called the Smart Station, has a low initial investment, and it is simple to program and to use for data collection. For the first time in transportation research, Kismet Wireless software was used and proved to be a powerful tool for real-time monitoring. The richness of the data can be attributed to the combination of Wi-Fi and GPS information. This was used to feed algorithms that were developed for this research and the data were tested to be statistically accurate.

The advantages of using Smart Station over traditional data collection methods include the following: (1) Wireless, automated data collection and retrieval, (2) Real-time observation of passenger behavior, (3) Negligible maintenance after programming and installing the hardware, (4) Low costs of hardware, software, and installation, and (5) Simple and short programming and installation time. The remainder of this chapter provides a summary of the context and analysis of the results of ridership, OD flows, wait time and travel time. In addition, limitations and future research are discussed.

6.1. Ridership

Ridership provides transportation and city planners with a framework to develop or enhance transit projects as the users need it. The number of passengers is a parameter that changes over time and space. Understanding these changes can help optimize the number and capacity of transit vehicles in order to match the offer with the passengers' demand. Real-time estimation of the number of riders can be useful information to make changes in the transit routes that would benefit the riders instantaneously, improving the user experience. With the Smart Stations and the proposed methodologies, it is possible to collect information about trip behavior, like the number of passengers.

This research utilizes Wi-Fi and GPS data obtained with the Smart Station in order to estimate the number of passengers of the Streamline buses. This estimate was obtained using rule-based algorithms to filter out those Wi-Fi signals that do not belong to passengers' devices. There are certain characteristics like network type and the data transfer rate that can be used to rule out devices that do not belong to riders. On the other hand, there are characteristics like detection time and speed that have a complex context and cannot be used directly for classification of riders.

The rule-based and enhanced rule-based methods have an average accuracy of 98% and 115% after Poisson regression correction, respectively. These percentages are more accurate than other studies that use Wi-Fi and Bluetooth that were found in the literature. Therefore, ridership estimation using the Smart Station has proved to be a promising approach. However, more research would shed light on the accuracy of these results before fully implementing this methodology for industrial purposes.

The use of an unsupervised machine learning algorithm was explored in order to classify the devices that would belong to riders. The K-means clustering method was implemented for its simplicity and popularity. A PCA decomposition was used to reduce the dimensionality of the pre-processed data for easier implementation of the K-means clustering algorithm and for graphing the data. After data visualization, a total of two centroids produced more accurate results.

After correction by the Poisson regression model, the accuracy of the classification was 128%. This accuracy is lower than the rule-based methods. In general, the clustering algorithm underestimated ridership before correction.

A problem that may arise with these algorithms is that the models could be overfitting the values to the manual counts that were used for training to obtain the correction factor. Increasing the sample size can provide an idea of whether or not the factors provide good estimates for other days. This becomes one disadvantage for this method because the ground truths are still necessary. This correction factor may differ from place to place. Nevertheless, the Smart Station can decrease the size of the manual counts in industrial applications.

6.2. OD Flow Characteristics

Origin-destination characteristics of passengers provide transportation planners with information to evaluate if the routes satisfy the needs of passengers. In the transit industry, the OD flows are usually estimated with on-board surveys. Nonetheless, this method is time-consuming, labor intensive, and can be prone to response bias. This study combines Wi-Fi and GPS data collected by the Smart Station and uses timestamps and location to determine origin-destination patterns. The preprocessed data needed to be filtered to retain the actual passenger trips and remove detections that do not belong to passengers' devices, however.

The advantage of using the Smart Station is that trip duration can be detected because each device has a unique MAC address. Considering accurate time stamps and bus location monitoring at the same time, it is possible to estimate the precise stop where a passenger device boards or alights a transit vehicle. Using all these captured trips, OD flow patterns are generated.

It is necessary to match the timestamps with the vehicle's GPS coordinates. Automated passenger counts and other methods rely on manual counts to calibrate models of OD flows estimation. However, with the Smart Station, individual passengers are indirectly observed by tracing their MAC addresses. In this case study, the OD matrices generated by the model match the behavior obtained by the manual counts and the survey data. Therefore, the methodology proposed is adequate for OD estimation. Since it is dependent on the bus stop's location and the GPS data collected, this methodology could be implemented on any transit system in an area that has access to the GPS signals. Additionally, there is not a reason to believe passengers would decide certain bus stops over others based on their device's brand. Therefore, even with randomization, a large sample size would provide unbiased results of the population's OD flows.

6.3. Wait Time

Passengers' wait times at bus stops are very difficult values to observe in the field because of the large number of people that may wait at a bus stop; it is difficult to remember and keep track of every passenger while they are waiting. The other common method to obtain passengers' wait times is to perform surveys asking them their perceived wait time. However, the perceived wait time is subject to bias because human beings have a hard time remembering exact numerical

values, especially if they do not keep track of the time. Both methods may have a response biased since not all passengers can be tracked out if there are many people at the bus stop and some passengers are inclined to refuse to fill out surveys.

This study proposes the use of a new device to track passengers and estimate their wait time. The technology proved to provide accurate estimations of the detection time of a device while it was within the range of detection of the Smart Station. This detection range was demonstrated to be at least ten meters; however, Wi-Fi technology has a higher range of detection, which indicates that the devices could be detected at a wider range. Additionally, the accuracy of detection was not affected by different device manufacturers, distance from the Smart Station or total time a device could be detectable. The study case of this research was a bus stop in Bozeman, Montana. Data were collected for over a week. The observed and estimated wait time values differed for over a minute on average and they were not statistically different. Hence the proposed methodology accurately described the wait times of the passengers. If more research provides similar results, the methodology can become a referent for consultants and transit agencies to obtain the behavior of passengers at bus stops in an efficient manner.

6.4. Travel Time

Understanding the travel time of a transit vehicle is key in the evaluation of the quality of service provided to passengers. Additionally, the travel time is directly proportional to the estimated time of arrival of a vehicle, which can be provided to passengers. When passengers know the time they have to wait before boarding a transit vehicle, they feel more satisfied with the service provided. The transit system of the study case analyzed in this study relies on a fixed schedule for passengers to estimate the time of arrival of the buses. Nevertheless, the buses fail to follow the schedule as noted by the great variability of the observed travel times. Also, passengers complained about the actual arrival time of the vehicles.

This study takes advantage of the use of Smart Stations of buses and estimated the travel times based on GPS technology that was used to infer the location of the devices that were being detected based on their MAC addresses. The objective was to demonstrate the versatility of the Smart Stations to collect different types of transportation data while scanning passenger movements.

The error between the ground truths and the estimated travel times was calculated to be 1.2 seconds, on average, for this study case. Additionally, the error was not rejected to be different from zero. Different lines, days, times of days and traffic characteristics did not affect the accuracy of these results. Therefore, the Smart Station can be used to estimate travel times of a transit vehicle while it collects other data. However, more research is recommended in order to reduce several outliers that were observed.

6.5. Limitations of Research

This research provides a tool that has the potential to infer various parameters of passengers in a transit system. Nevertheless, there are factors that may decrease the accuracy of the results with the use of Smart Station. It is important to note these restrictions so they can be taken into consideration by future researchers. The limitations are associated with hardware, Wi-Fi data, filtering methods, and privacy concerns.

6.5.1. Hardware

The Raspberry Pi is a computer that relies on a power supply that may not always be provided in a transit system. In this study, portable batteries were utilized as a power source. They reliably provided power for around twelve hours. A research study that uses the Smart Station for longer times of data collection should be equipped with a permanent source of power to work.

The antenna for GPS data collection was a cheap, but accurate device. Nevertheless, a physical obstruction, such as being contained inside a box or poor extension of the cord, may lead to the imperceptibility of the GPS satellite signals. It is important to note, however, that the device collects accurate GPS data inside the buses and buildings with no need to be in open air.

The computers need an internet connection for time synchronization. Whenever the Raspberry Pi is rebooted, it will adopt the time it had at the last moment it was turned on. It is not until after the internet provides an accurate time based on the location of the connection, that timestamps can be reliable. Therefore, an internet connection is recommended before data collection for the purpose of time synchronization.

Ideally, the Raspberry Pi should always have internet connection for time synchronization, remote data retrieval and for inspection of the computer. Nonetheless, the internet may not always be available in the transit system. In such a case, surveyors need to provide internet connection through hotspots or other methods, which is expensive and against the philosophy of this method. Internet, however, may be provided just before data collection and the computers can measure time accurately without an internet connection. The computers must be retrieved later to obtain the data.

The Smart Stations were rebooted every time a loop was completed to avoid detection time overestimation. A good practice is to turn on the computers at the beginning of a loop and turn them off when the loop ends. Best results will be obtained when the loops do not come back to the original location using the same route. In other words, when a transit vehicle travels from an initial location to a final one, the GPS data is easier to analyze because routers will not be detected again, which exaggerates their detection time.

Although the Smart Station proved to accurately detect phone probes in a ten-meter radius, they can easily detect passengers even if a transit vehicle is larger than that. The premise is to locate the SS near the doors so it will detect the riders when they board and alight the vehicle.

6.5.2. Wi-Fi Data

Wi-Fi technology relies on the air as a medium for sharing information. The air is subject to changing meteorological conditions. Internet providers are susceptible to various sources of failure due to the architecture of the Wi-Fi devices, the medium, and quality of data transferred. Wi-Fi has a wide range of detection, which can cause ubiquitous sources of noise. The GPS data used a different time system than the Wi-Fi data. This must be considered in the development of the algorithms. Finally, random sampling of the Wi-Fi data can only be considered with certainty when there are similarities with the ground truths.

6.5.3. Filtering Methods

The filtering methods can have two types of uncertainty. The filters could be too permissive to keep networks that are not riders or be too antagonistic and rule out networks that are passengers. Therefore, the parameters need to be tuned according to the minimum errors based on the ground

truths. This could lead to the generation of overfitting models. Moreover, the calibration factors may change from one transit system to another.

Additionally, some passengers carry multiple devices. This was confirmed when surveys were implemented to explore the technology penetration rate. Furthermore, the possession of a device is dependent on socio-economic and demographic characteristics. Relying solely on networks can be misleading for different geographical locations.

6.5.4. Privacy Concerns

Some people may have concerns about the use of Wi-Fi technology for public data collection. Smart Stations track and collect several attributes of passengers' detectable devices. Some individuals may not be aware of the amount of information that their handheld electronic devices share when they connect to an internet provider or even when the devices are by default looking for an access point to connect to. However, the MAC addresses and owners cannot be linked by the public without the datasets that only mobile carriers have. Therefore, collecting Wi-Fi data with Smart Stations will not allow surveyors to determine information about specific passengers, not even names. In addition, since the SS will be on buses and surveyors will not be there, privacy is less affected as there is less interaction with riders. Educating the public about this fact may facilitate adoption of this technology. The potential for transportation data collection is of enormous use for transportation planners.

6.6. Future Research

Up to this time, Smart Stations have provided data that have efficiently estimated ridership, OD flows, wait time, and travel time. Further research to estimate these variables in other case studies would strengthen the results obtained in this research. In addition, the methodologies can be modified to understand the role of the different cut off values used in the rule-based methods. These cut off values can be different in other scenarios and surveyors would have to calibrate them based on initial manual data collection.



The exploration of the unsupervised machine learning was simplistic in this research. More methods should be explored to evaluate the feasibility of cluster analysis in passenger classification based on Wi-Fi data. In this research, tuning values were used for the rule based-methods, both by trial and error and by a cost function. Other clustering methods and graphical tools can be used to better understand the intrinsic nature of the data. Due to the large amount of data collected, data mining could be an adequate process of discovering patterns.

Wi-Fi technology has a large radius of detection compared to other wireless sensing methods. This could be beneficial for the use of the Smart Station. If the distance and characteristics of networks are similar to those of drivers' devices, an estimation of the parking lots available could be made. Since the XML archive collected by the SS is a markup language, a lot of text can be analyzed. To simplify the complexity of analyzing standard words, the XML was converted to Excel tables. This inconsistency is time- consuming, and the data are not in their natural form. Some patterns may be obtained only by analyzing the raw data. Therefore, data mining methods could be implemented on the original datasets. This task requires high computational skills.

It is believed that Smart Stations can also be used to estimate traffic flow characteristics such as speed, flow, and density. If drivers of a highway system also carry smartphone devices, their Wi-Fi probes could be detected as they traverse different segments of the highway. Two Smart Stations

can be installed along a highway segment and provide information on detection time and an estimation of devices' speed. These traffic characteristics are an essential requirement in the planning, design, and operation of transportation systems. Therefore, the use of Smart Stations can become an incredible new tool for transportation researchers and professionals.

7. APPENDIX A: SAMPLE SURVEY IMPLEMENTED IN THE PILOT STUDY (BLUELINE)

<div style="display: flex; justify-content: space-around; align-items: center;"> <div style="text-align: center;">  <p>MONTANA STATE UNIVERSITY</p> </div> <div style="text-align: center;"> <p>2018 Streamline Rider Survey</p> </div> <div style="text-align: center;">  <p>A PROGRAM OF HRDC</p> </div> </div>																																						
Time:	Date:	Bus Line: Blue																																				
<p>This survey is intended to collect data on the Streamline passengers' characteristics to perform an analysis of the Weekday Service, in terms of the number of users and origin-destination information. This study is independent of Streamline.</p>																																						
<p>1. What devices are you currently carrying (check all that apply)? If carrying multiple of the same device, please indicate by circling the number of that device that you are carrying.</p> <div style="display: flex; flex-wrap: wrap;"> <div style="width: 50%;"> <input type="checkbox"/> Smartphone (x2) (x3) </div> <div style="width: 50%;"> <input type="checkbox"/> Portable gaming device (x2) (x3) </div> <div style="width: 50%;"> <input type="checkbox"/> Tablet (e.g. iPad) (x2) (x3) </div> <div style="width: 50%;"> <input type="checkbox"/> Laptop (x2) (x3) </div> <div style="width: 50%;"> <input type="checkbox"/> Another phone (no Wi-Fi) (x2) (x3) </div> <div style="width: 50%;"> <input type="checkbox"/> Other: _____ </div> </div>																																						
<p>2. Check the brand name(s) of the device(s) that you are carrying.</p> <div style="display: flex; flex-wrap: wrap;"> <div style="width: 33%;"> <input type="checkbox"/> Samsung </div> <div style="width: 33%;"> <input type="checkbox"/> LG </div> <div style="width: 33%;"> <input type="checkbox"/> HTC </div> <div style="width: 33%;"> <input type="checkbox"/> Apple </div> <div style="width: 33%;"> <input type="checkbox"/> Motorola </div> <div style="width: 33%;"> <input type="checkbox"/> Other: _____ </div> </div>																																						
<p>3. Please select the bus stop at which you boarded, and the bus stop you plan to alight. (You may mark with an X or a check mark)</p> <table border="1" style="width: 100%; border-collapse: collapse; margin-top: 10px;"> <thead> <tr> <th style="width: 60%;">Bus stops</th> <th style="width: 10%;">Boarded</th> <th style="width: 10%;">De-board</th> </tr> </thead> <tbody> <tr><td>MSU Depart</td><td></td><td></td></tr> <tr><td>6th & Garfield</td><td></td><td></td></tr> <tr><td>Garfield & Wilson</td><td></td><td></td></tr> <tr><td>Wilson & College N</td><td></td><td></td></tr> <tr><td>Wilson & Curtis N</td><td></td><td></td></tr> <tr><td>Babcock & Tracy</td><td></td><td></td></tr> <tr><td>Mendenhall & Bozeman (Clinic)</td><td></td><td></td></tr> <tr><td>Mendenhall & Black (Downtown Transfer)</td><td></td><td></td></tr> <tr><td>Wilson & Lamme</td><td></td><td></td></tr> <tr><td>Tamarack & Tracy (Fairgrounds) W</td><td></td><td></td></tr> <tr><td>Tamarack & 5th W</td><td></td><td></td></tr> </tbody> </table>			Bus stops	Boarded	De-board	MSU Depart			6th & Garfield			Garfield & Wilson			Wilson & College N			Wilson & Curtis N			Babcock & Tracy			Mendenhall & Bozeman (Clinic)			Mendenhall & Black (Downtown Transfer)			Wilson & Lamme			Tamarack & Tracy (Fairgrounds) W			Tamarack & 5th W		
Bus stops	Boarded	De-board																																				
MSU Depart																																						
6th & Garfield																																						
Garfield & Wilson																																						
Wilson & College N																																						
Wilson & Curtis N																																						
Babcock & Tracy																																						
Mendenhall & Bozeman (Clinic)																																						
Mendenhall & Black (Downtown Transfer)																																						
Wilson & Lamme																																						
Tamarack & Tracy (Fairgrounds) W																																						
Tamarack & 5th W																																						

7th @ M Town Plaza		
Wal-Mart (Lawn & Garden door)		
Oak @ Days Inn		
Oak & 15th		
Bridger Peaks Town Center (North Center)		
19th & Baxter (Town Pump)		
Cattail @ City Brew (by driveway)		
27th & Catron (Social Security Office)		
Catron @ Target (driveway)		
Gallatin Center (Staples lot)		
19th & Baxter (after light)		
Oak & 15th (at pullout)		
Oak & 12th		
7th & Hemlock		
Tamarack & 5th		
Tamarack & Tracy (Senior Center) E		
Tamarack & Rouse (before light) E		
Rouse & Cottonwood		
Mendenhall & Bozeman (Clinic)		
Mendenhall & Black (Downtown Transfer)		
Wilson & Curtis		
Wilson & College S		
Grant & Wilson S		
MSU Arrive		

4. Do you have feedback or ideas that would improve the Streamline service? If so, leave a comment.

8. APPENDIX B: ORIGINS AND DESTINATIONS DURING MANUAL COUNTS

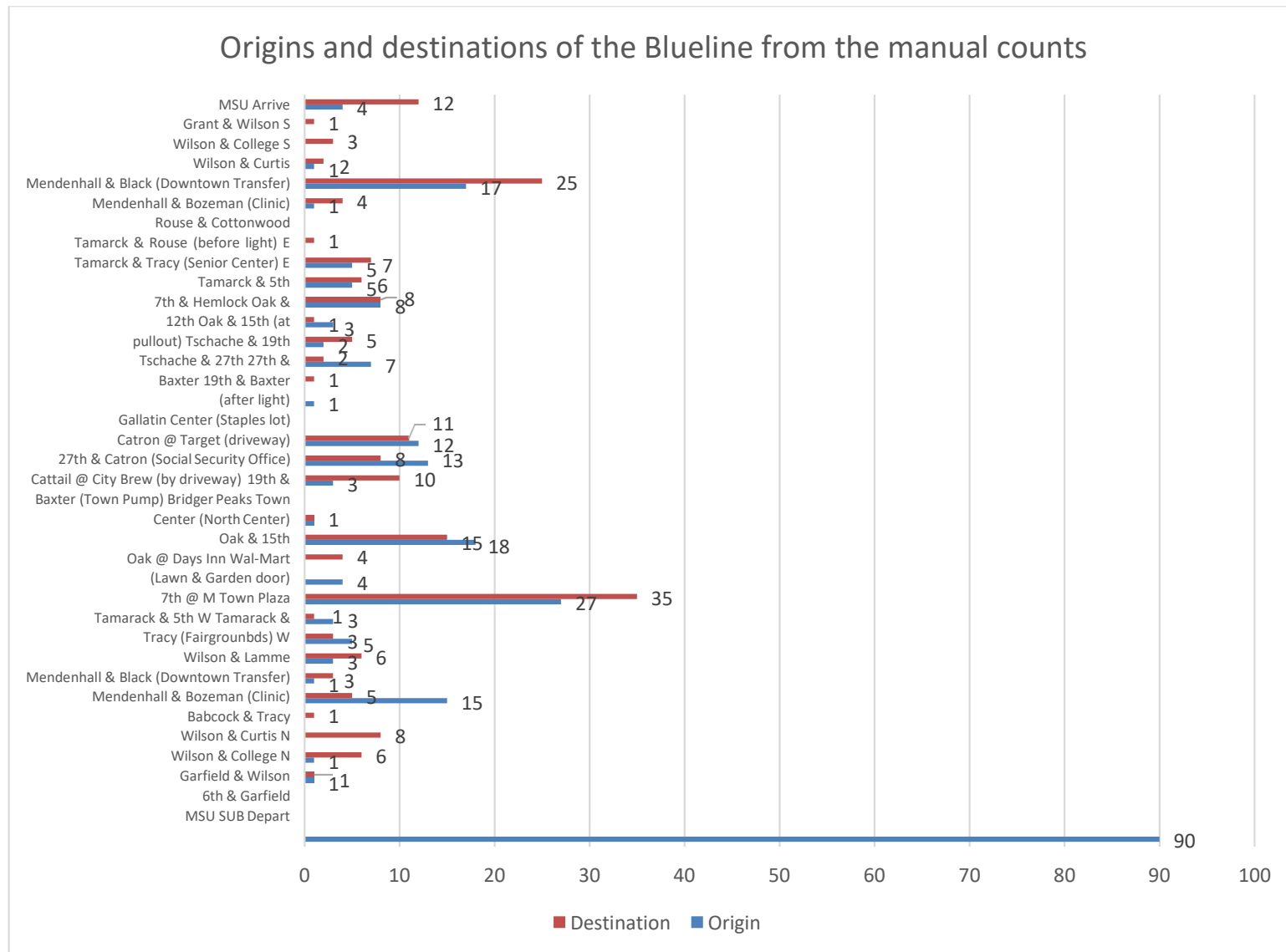


Figure 83: Origins and destinations of the Blueline from the manual counts

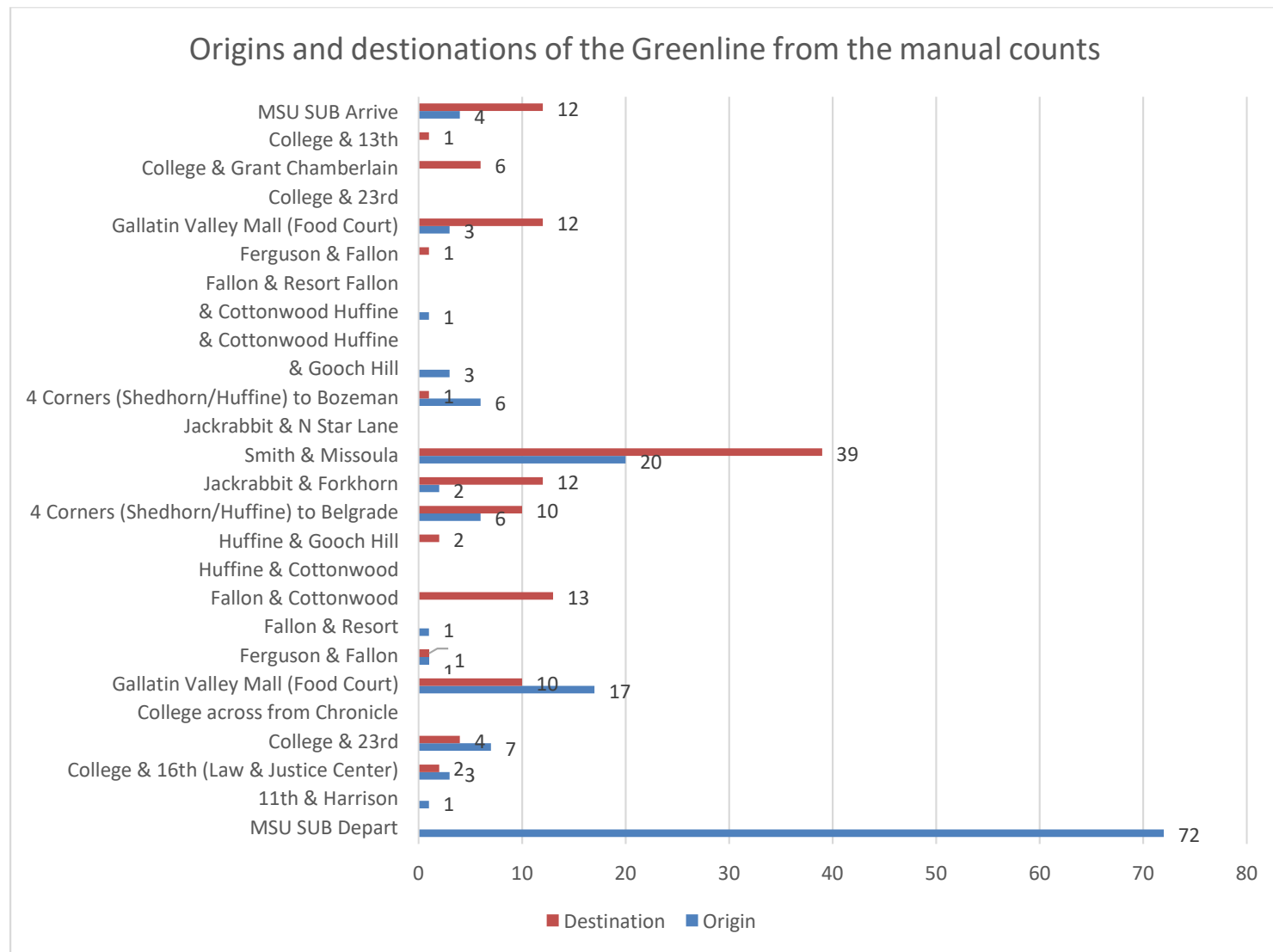


Figure 84: Origins and destinations of the Greenline from the manual counts

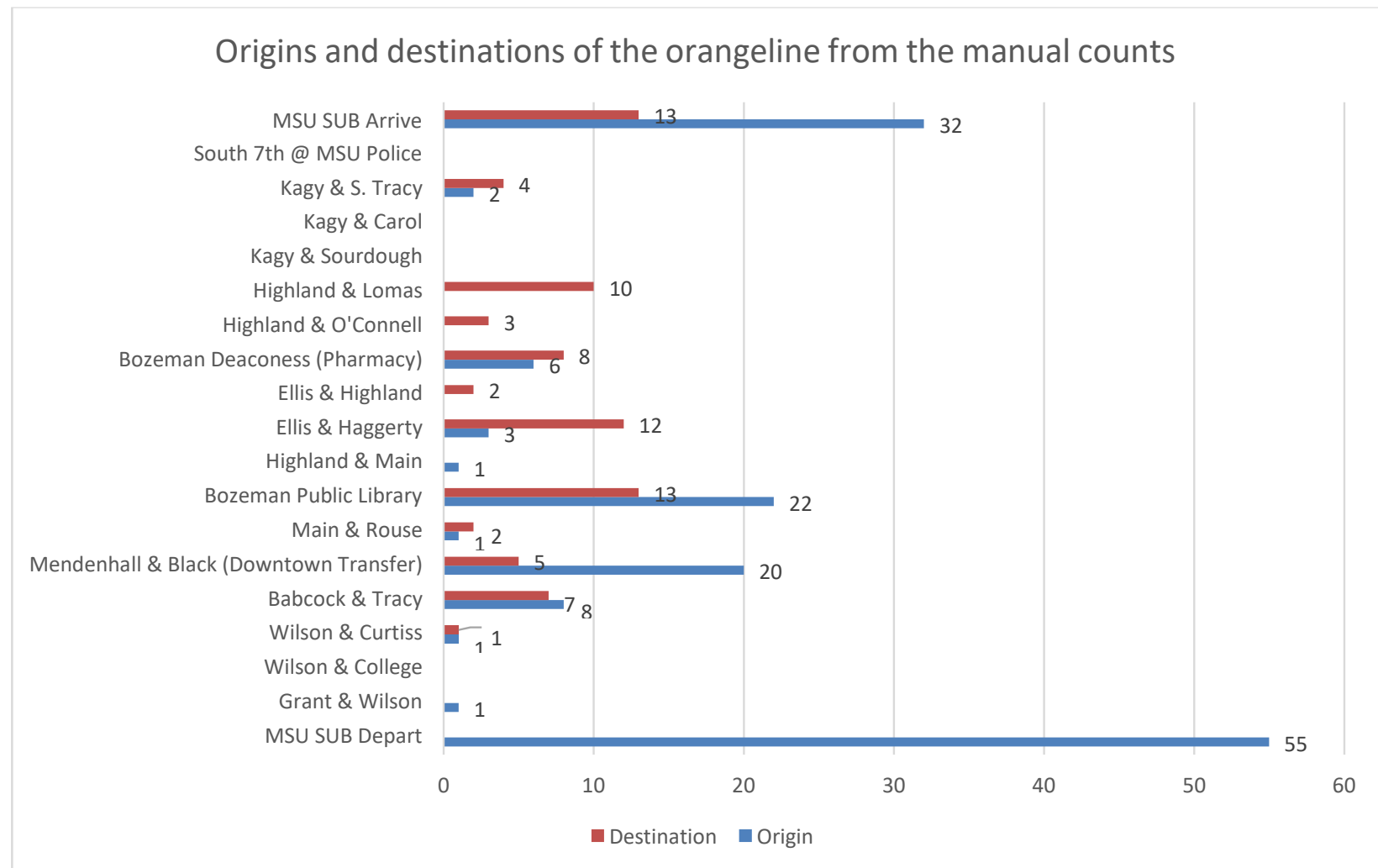


Figure 85: Origins and destinations of the Orangeline from the manual counts

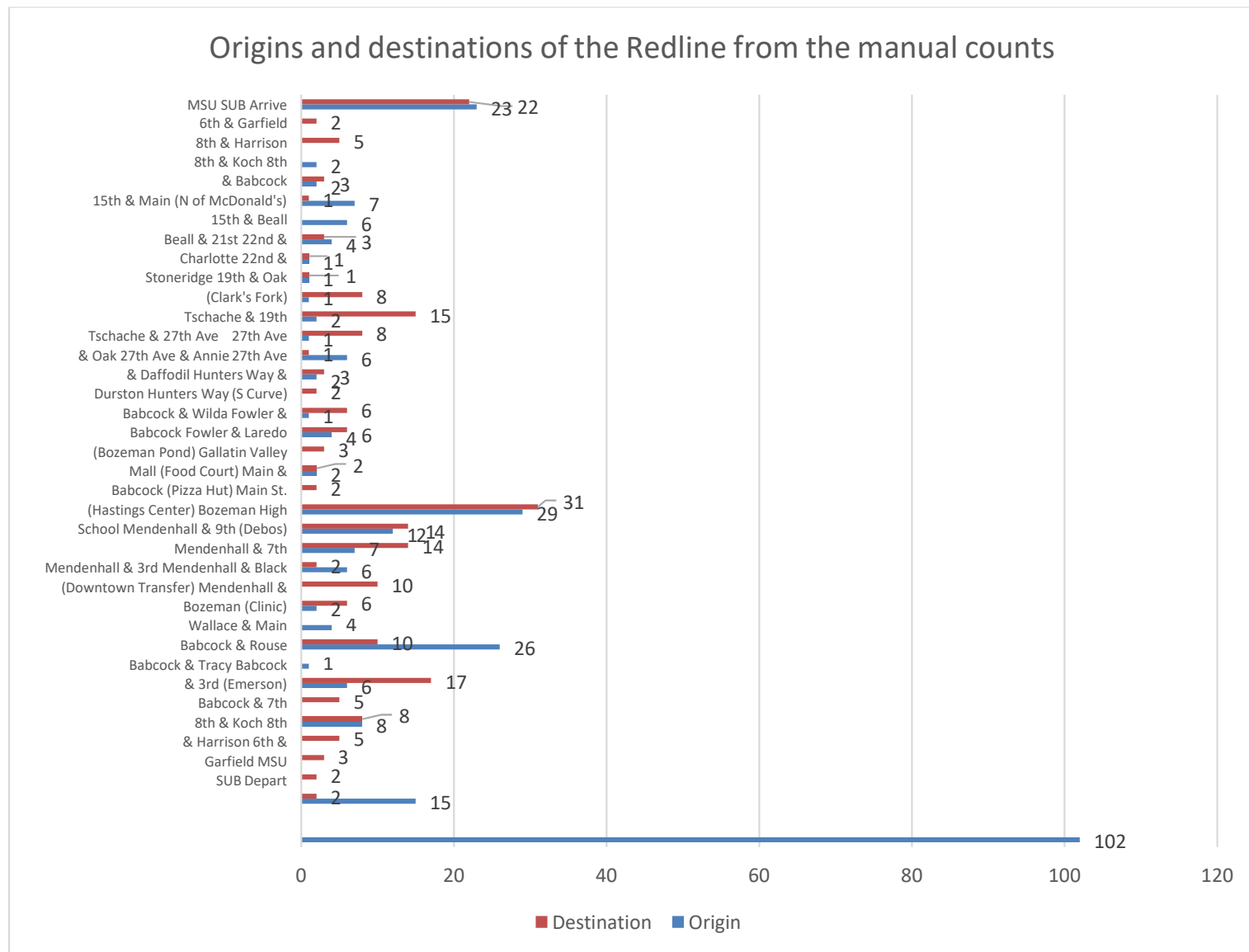


Figure 86: Origins and destinations of the Redline from the manual counts

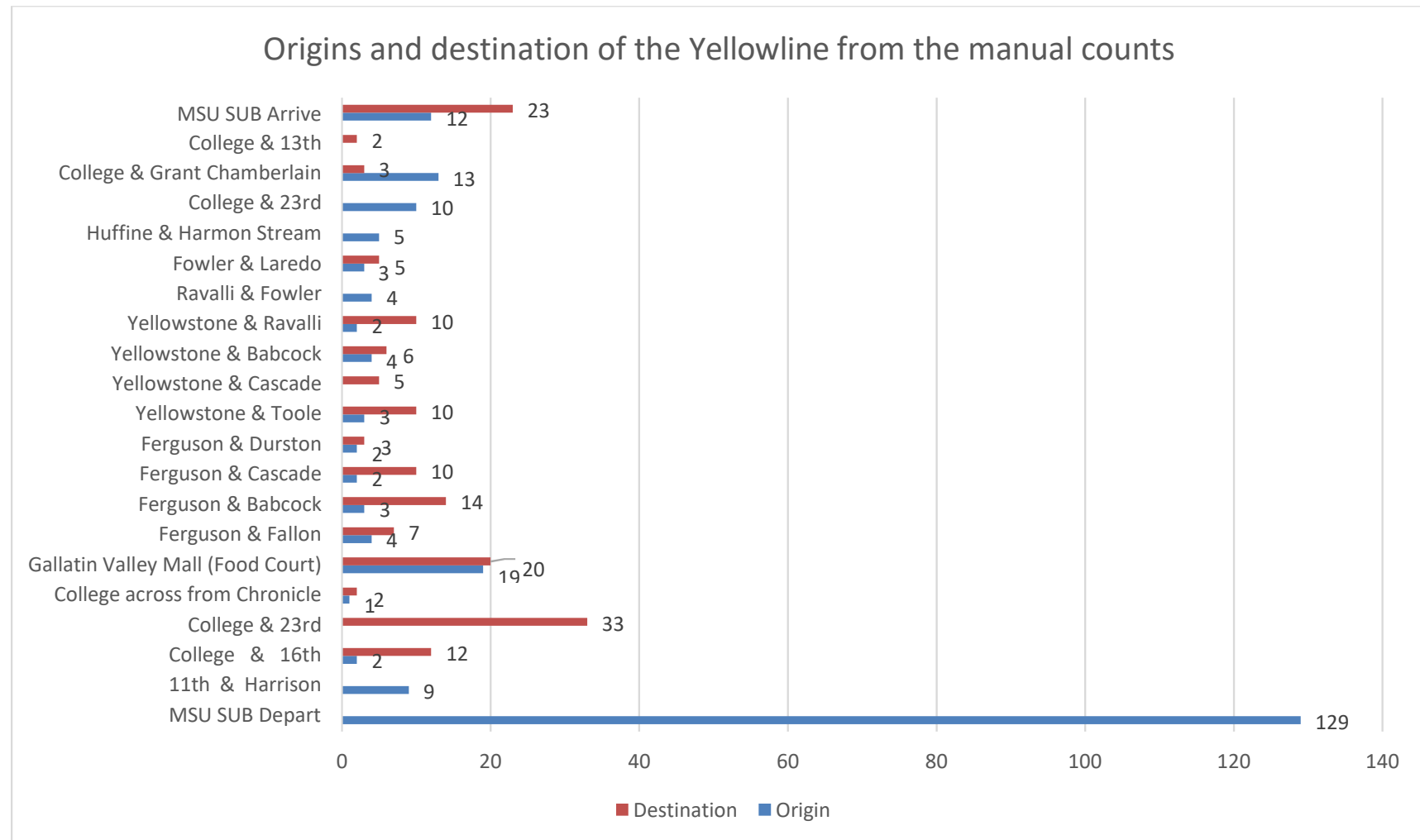


Figure 87: Origins and destinations of the Yellowline from the manual counts

9. APPENDIX C: DESCRIPTION OF THE ATTRIBUTES COLLECTED BY THE SMART STATION

List of attributes:

1. Wireless Network ID
2. First time of detection
3. Last time of detection
4. Type of wireless network
5. MAC address
6. Wireless network channel
7. Average latitude
8. Average longitude
9. Manufacturer
10. Maximum rate of data transfer
11. Maximum speed
12. Minimum speed
13. Maximum signal strength
14. Minimum signal strength

Number: 1

Title: /wireless-network/#id

Name: Wireless Network ID

Variable Type: Numeric

Description: This is a list of all the different networks detected in a numerically ascending manner.

Possible values: Set of natural numbers (1,2, ..., ∞)

Number: 2

Title: /wireless-network/@first-time

Name: First time of detection

Variable Type: String

Description: This is a string containing the date and time a Wireless Network ID was first detected.

Possible values: Day of the week + Date + Time of day + Year

Number: 3

Title: /wireless-network/@last-time

Name: Last time of detection

Variable Type: String

Description: This is a string containing the date and time a Wireless Network ID was last detected.

Possible values: Day of the week + Date + Time of day + Year

Number: 4

Title: /wireless-network/@type

Name: Type of Wireless Network

Variable Type: String

Description: This describes the type of wireless signal that is being sent by a device and the hardware that is sending the network.

Possible values:

- Probe: This is a signal that is sending data packets everywhere to then receive a response from a device that connects to the Internet. All devices that connect to the internet are sending probes when they are not connected to the internet (Liang, Qing and Dongxia 2014). These probes can be seen by the Smart Station.
- Infrastructure: This is a device that is capable of sending information to a mobile device based on the requested information.

- Ad hoc: This is a network that is being sent by various nodes. It is an infrastructure type of network that does not rely on a single access point.
- Data: This is wireless data being transmitted from nodes, generally of commercial use (e.g. printers and local networks).

Number: 5

Title: /wireless-network/BSSID

Name: MAC address

Variable Type: String

Description: This parameter contains the unique identifier code which is needed by a router to send the requested information to the correct device.

Possible values: A string composed of 16 alphanumeric values separated into two with a colon.

Number: 6

Title: /wireless-network/channel

Name: Wireless Network Channel

Variable Type: String

Description: This is a list of all the different channels that are used to transfer signals detected. The different channels correspond to a different frequency of the electromagnetic signals sent by the devices. Channels 13 and 14 are illegal in the United States (Data Alliance Inc. 2004).

Possible values:

- Channel 0: 2407 Megahertz
- Channel 1: 2412 Megahertz
- Channel 2: 2417 Megahertz
- Channel 3: 2422 Megahertz
- Channel 4: 2427 Megahertz
- Channel 5: 2432 Megahertz
- Channel 6: 2437 Megahertz
- Channel 7: 2442 Megahertz
- Channel 8: 2447 Megahertz
- Channel 9: 2452 Megahertz
- Channel 10: 2457 Megahertz
- Channel 11: 2462 Megahertz
- Channel 12: 2467 Megahertz
- Channel 13: 2472 Megahertz
- Channel 14: 2482 Megahertz

Number: 7

Title: /wireless-network/gps-info/avg-lat

Name: Average latitude

Variable Type: Numeric

Description: This is the angular distance between the equator and the average position of a detected device. The value is expressed in degrees.

Possible values: The minimum and maximum values for the city of Bozeman and Belgrade are 45.6 and 45.8 degrees, respectively.

Number: 8

Title: /wireless-network/gps-info/avg-lon

Name: Average longitude

Variable Type: Numeric

Description: This is the angular distance between the meridian of Greenwich and the average position of a detected device. The value is expressed in degrees.

Possible values: The minimum and maximum values for the city of Bozeman and Belgrade are the following: -110.9 and -111.2 degrees, respectively.

Number: 9

Title: /wireless-network/manuf

Name: Manufacturer of Device

Variable Type: String

Description: It contains the name of the company that manufactures the hardware of a wireless able device.

Possible values: Names of manufacturers (e.g. SamsungE, Apple, Zte, TctMobil, MurataMa, etc.).

Number: 10

Title: /wireless-network/maxseenrate

Name: Maximum rate of data transfer

Variable Type: Numeric

Description: This is the maximum rate of kilobits per second at which a device is sending and receiving information. A router has the capacity of sending 54 megabytes per second. Generally, a phone will be sending around 1 and 6 megabytes per second.

Possible values: 1000 to 54000 Bytes per second.

Number: 11

Title: /wireless-network/gps-info/max-spd

Name: Maximum speed

Variable Type: Numeric

Description: This is the maximum relative speed between the detected device and the Smart Station. The values are provided in meters per second. Three decimals are provided.

Possible values: Set of positive real numbers. Values larger than 26.8 m/s (60 mph) should not be found because it exceeds the speed limit, however, some errors or poor signal may introduce other values.

Number: 12

Title: /wireless-network/gps-info/min-spd

Name: Minimum Speed

Variable Type: Numeric

Description: This is the minimum relative speed between the detected device and the Smart Station. The values are given in meters per second. Three decimals are provided.

Possible values: Set of positive real numbers. Values larger than 26.8 m/s (60 mph) should not be found because it exceeds the speed limit.

Number: 13

Title: /wireless-network/snr-info/max_signal_dbm

Name: Maximum Signal Strength

Variable Type: Numeric

Description: This is the maximum signal strength in decibel milliwatts (-dBm) at which a device was seen. The greater the absolute value, the farther the device is located, which is translated into a poor signal. The value signifies the exponent that the base 10 is being raised to. The largest value possible would be 10^0 , that would yield a value of 1 milliwatt (mW) (Moyers 2015).

Possible values: Signal strength is represented in -dBm format (0 to -100)

Number: 14

Title: /wireless-network/snr-info/min_signal_dbm

Name: Minimum Signal Strength

Variable Type: Numeric

Description: This is the minimum signal strength in decibel milliwatts (-dBm) at which a device was seen. The greater the absolute value, the farther the device is located, which is translated into

a poor signal. The value signifies the exponent that the base 10 is being raised to. The largest value possible would be 10^0 , that would yield a value of 1 milliwatt (mW) (Moyers 2015).

Possible values: Signal strength is represented in -dBm format (0 to -100)

References for Appendix C:

Data Alliance Inc. 2004. Legal and Illegal Frequencies & Channels. Accessed February 12, 2018. <http://en.data-alliance.net/legal-illegal-frequencies/>.

Liang, Ming, Miao Qing, and Wang Dongxia. 2014. "Research on monitoring probe deployment in large scale network." 2014 International Conference on Information and Network Security 110-114.

Moyers, Eric. 2015. Why is almost everything negative in Wireless? July 17. Accessed February 2, 2018. <https://supportforums.cisco.com/t5/small-business-support-documents/why-is-almost-everything-negative-in-wireless/ta-p/3159743>.

10. APPENDIX D: ESTIMATION OF NUMBER OF CLUSTERS FOR THE K-MEANS ALGORITHM

Table 61: Number of clusters by the elbow, silhouette, and gap statistic methods

Date	Line	Elbow	Silhouette	Gap statistic
1/14/2019	Blue	2	2	2
1/15/2019	Blue	2	2	2
1/16/2019	Blue	2	3	2
1/17/2019	Blue	2	2	2
1/18/2019	Blue	2	2	2
10/8/2018	Blue	4	2	1
10/10/2018	Blue	5	2	1
10/11/2018	Blue	2	1	2
10/12/2018	Blue	2	2	3
10/18/2018	Blue	2	2	2
2/4/2019	Green	2	2	2
2/5/2019	Green	2	2	2
2/6/2019	Green	2	2	2
2/7/2019	Green	2	2	2
2/8/2019	Green	2	2	2
10/17/2018	Green	4	2	2
10/18/2018	Green	2	2	1
10/19/2019	Green	2	2	4
10/19/2019	Green	3	2	2
1/7/2019	Orange	2	2	2
1/8/2019	Orange	2	2	1
1/9/2019	Orange	3	3	3
1/10/2019	Orange	2	2	1
1/11/2019	Orange	3	1	4
10/8/2018	Orange	4	2	1
10/10/2018	Orange	4	2	1
10/18/2018	Orange	5	3	1
1/7/2019	Red	2	2	1
1/8/2019	Red	2	2	1
1/9/2019	Red	3	2	1
1/10/2019	Red	4	2	2
1/11/2019	Red	2	2	2
10/16/2018	Red	2	2	1
10/17/2018	Red	2	2	2
10/17/2018	Red	4	2	1
10/18/2018	Red	2	2	2
10/19/2018	Red	2	2	1
1/14/2019	Yellow	4	2	1
1/15/2019	Yellow	4	2	1

Date	Line	Elbow	Silhouette	Gap statistic
1/16/2019	Yellow	4	2	1
1/17/2019	Yellow	2	2	2
1/18/2019	Yellow	2	2	1
10/10/2018	Yellow	2	2	2
10/18/2018	Yellow	2	2	1
10/18/2018	Yellow	2	2	2
10/19/2018	Yellow	2	2	1

11. REFERENCES

- Abbott-Jard, M., Shah, H., & Bhaskar, A. (2013). Empirical evaluation of Bluetooth and Wifi scanning for road transport. *Australasian Transport Research Forum (October, 2013)*, Brisbane, Australia., (October), 1–14.
- Abdi, H., & Williams, L. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(August), 433–459. <https://doi.org/10.1016/B978-0-08-044894-7.01358-0>
- Abdulhai, B., Pringle, R., & Karakoulas, G. J. (2003). Reinforcement Learning for True Adaptive Traffic Signal Control. *Journal of Transportation Engineering*. [https://doi.org/10.1061/\(asce\)0733-947x\(2003\)129:3\(278\)](https://doi.org/10.1061/(asce)0733-947x(2003)129:3(278))
- Abedi, N., Bhaskar, A., & Chung, E. (2013). Bluetooth and Wi-Fi MAC Address Based Crowd Data Collection and Monitoring : Benefits , Challenges and Enhancement. *Australasian Transport Research Forum 2013 Proceedings 2*, (October), 1–17. <https://doi.org/10.1080/02640410701348669>
- Ahmed, H., El-Dariby, M., Morgan, Y., & Abdulhai, B. (2008). A wireless mesh network-based platform for ITS. *IEEE Vehicular Technology Conference*, 3047– 3051. <https://doi.org/10.1109/VETECS.2008.329>
- Aldein Mohammed, Z. K., & Ali Ahmed, E. S. (2017). Internet of Things Applications , Challenges and New Technologies. *World Scientific News*, (978), 126–148.
- Amit, Y., Geman, D., & Wilder, K. (1997). Joint induction of shape features and tree classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(11), 1300–1305. <https://doi.org/10.1109/34.632990>
- Arciszewski, T., Khasnabis, S., Khurshidulhoda, S., & Ziarko, W. (1994). Machine learning in transportation engineering: A feasibility study. *Applied Artificial Intelligence*, 8(1), 109–124. <https://doi.org/10.1080/08839519408945434>
- Arel, I., Liu, C., Urbanik, T., & Kohls, A. G. (2010). Reinforcement learning-based multi-agent system for network traffic signal control. *IET Intelligent Transport Systems*. <https://doi.org/10.1049/iet-its.2009.0070>
- Bahoken, F., & Raimond, A.-M. O. (2013). Designing Origin-Destination Flow Matrices from Individual Mobile Phone Paths The effect of spatiotemporal filtering on flow measurement. *26th International Cartographic Conference*. Retrieved from http://www.researchgate.net/publication/257101224_Designing-Origin-Destination_Flow_Matrices_from_Individual_Mobile_Phone_Paths_-The_effect_of_spatiotemporal_filtering_on_flow_measurement/file/72e7e52ca7f2da5737.pdf
- Bai, L., Ireson, N., Mazumdar, S., & Ciravegna, F. (2017). Lessons learned using wi-fi and Bluetooth as means to monitor public service usage. *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers on - UbiComp '17*, 432–440. <https://doi.org/10.1145/3123024.3124417>

- Baranski, W. M., Wytyczak-Partyka, A., & Walkowiak, T. (2008). Computational complexity reduction in PCA-based face recognition. *Sixth International Conference on Soft Computing Applied in Computer and Economic Environments*, (1), 2–5. Retrieved from <https://facedetect-f-spot.googlecode.com/files/2007KunoviceMBAWPTW.pdf>
- Barry, J., Newhouser, R., Rahbee, A., & Sayeda, S. (2002). Origin and Destination Estimation in New York City with Automated Fare System Data. *Transportation Research Record: Journal of the Transportation Research Board*, 1817(02), 183– 187. https://doi.org/10.1007/978-3-642-30529-0_12
- Bates, J., Polak, J., Jones, P., & Cook, A. (2001). The valuation of reliability for personal travel. *Transportation Research Part E: Logistics and Transportation Review*. [https://doi.org/10.1016/S1366-5545\(00\)00011-9](https://doi.org/10.1016/S1366-5545(00)00011-9)
- Bazzan, A. L. C. (2009). Opportunities for multiagent systems and multiagent reinforcement learning in traffic control. *Autonomous Agents and Multi-Agent Systems*, 18(3), 342–375. <https://doi.org/10.1007/s10458-008-9062-9>
- Beard, C., & Stallinds, W. (2016). *Wireless Communication Networks and Systems* (First). Pearson Higher Education, Inc. Retrieved from <https://books.google.nl/books?id=LJ5VCwAAQBAJ>
- Ben-Akiva, M. E., & Lerman, S. R. (1987). *Discrete Choice Analysis: Theory and Application to Predict Travel Demand*. *Journal of the Operational Research Society*.
- Ben-akiva, M., & Morikawa, T. A. I. A. (1985). Data Combination and Updating Methods for Travel Surveys. *Transportation Research Record* 1203, (6), 40–47.
- Bhaskar, A., Qu, M., & Chung, E. (2015). Bluetooth vehicle trajectory by fusing bluetooth and loops: Motorway travel time statistics. *IEEE Transactions on Intelligent Transportation Systems*, 16(1), 113–122. <https://doi.org/10.1109/TITS.2014.2328373>
- Bihorel, S. (2018). *The neldermead Package - version 1.0-9*.
- Bock, H. H. (1996). Probabilistic models in cluster analysis. *Computational Statistics and Data Analysis*, 23(1), 5–28. [https://doi.org/10.1016/0167-9473\(96\)88919-5](https://doi.org/10.1016/0167-9473(96)88919-5)
- Böhm, M. F. (2016). Digital based Pedestrian Counting, (June).
- Bongiorno, N., Bosurgi, G., Pellegrino, O., & Sollazzo, G. (2017). How is the Driver's Workload Influenced by the Road Environment? *Procedia Engineering*, 187, 5–13. <https://doi.org/10.1016/j.proeng.2017.04.343>
- Bozeman City Commission. (2009). *Bozeman Community Plan*. Bozeman.
- Breiman, L. (1994). *Bagging predictors: Technical Report No. 421*. Department of Statistics University of California.
- Breiman, L. (2001). Random Forests. *Machine Learning*, 5–32. <https://doi.org/10.1023/A:1010933404324>
- Bude, C., & Kervefors, A. (2015). Internet of Things Internet of Things Exploring and Securing a Future Concept Industrial adviser. *Degree Project in Communication Systems, First Level Stockholm, Sweden*, 19–28.

- Bullock, D., Haseman, R., Wasson, J., & Spitler, R. (2010). Automated Measurement of Wait Times at Airport Security. *Transportation Research Record: Journal of the Transportation Research Board*, 2177(2177), 60–68. <https://doi.org/10.3141/2177-08>
- Buluswar, S. D., & Draper, B. A. (1998). Color machine vision for autonomous vehicles. *Engineering Applications of Artificial Intelligence*, 11(2), 245–256. [https://doi.org/10.1016/S0952-1976\(97\)00079-1](https://doi.org/10.1016/S0952-1976(97)00079-1)
- Bylander, T. (2002). Estimating generalization error on two-class datasets using out-of-bag estimates. *Machine Learning*, 48(1–3), 287–297. <https://doi.org/10.1023/A:1013964023376>
- Cambridge Systematics Inc. (2013). *Bozeman Community Transportation Safety Plan*. City of Bozeman. <https://doi.org/10.1016/B978-0-12-397829-5.00013-2>
- Cameron, A. C., & Trivedi, P. K. (1999). *Essentials of Count Data Regression*.
- Carrel, A., Lau, P. S. C., Mishalani, R. G., Sengupta, R., & Walker, J. L. (2015). Quantifying transit travel experiences from the users' perspective with high-resolution smartphone and vehicle location data: Methodologies, validation, and example analyses. *Transportation Research Part C: Emerging Technologies*, 58, 224–239. <https://doi.org/10.1016/j.trc.2015.03.021>
- Cathedral and the Bazaar. (2019). GPSd — Put your GPS on the net! Retrieved February 15, 2019, from <http://www.catb.org/gpsd/>
- Chan, K. Y., Dillon, T. S., Singh, J., & Chang, E. (2012). Neural-network-based models for short-term traffic flow forecasting using a hybrid exponential smoothing and levenberg-marquardt algorithm. *IEEE Transactions on Intelligent Transportation Systems*. <https://doi.org/10.1109/TITS.2011.2174051>
- Chang, G. L., & Su, C. C. (1995). Predicting intersection queue with neural network models. *Transportation Research Part C*. [https://doi.org/10.1016/0968-090X\(95\)00005-4](https://doi.org/10.1016/0968-090X(95)00005-4)
- Chang, N., Rashidzadeh, R., & Ahmadi, M. (2010). Robust indoor positioning using differential Wi-Fi access points. *IEEE Transactions on Consumer Electronics*. <https://doi.org/10.1109/TCE.2010.5606338>
- Chen, L. J., & Hung, H. H. (2011). A two-state markov-based wireless error model for bluetooth networks. *Wireless Personal Communications*, 58(4), 657–668. <https://doi.org/10.1007/s11277-009-9899-5>
- Cheng, C. (2006). *Symmetric and trimmed solutions of simple linear regression*. University of Southern California.
- Chow, W. (2014). *Evaluating Online Surveys for Public Transit Agencies Using a Prompted Recall Approach*. Massachusetts Institute of Technology.
- Cisco Press. (2017). WiFi Networking: Radio Wave Basics | IT Infrastructure Advice, Discussion, Community - Network Computing. Retrieved February 9, 2019, from <https://www.networkcomputing.com/wireless-infrastructure/wifi-networking-radio-wave-basics>
- Cortés, C. E., Jara-Díaz, S., & Tirachini, A. (2011). Integrating short turning and deadheading in the optimization of transit services. *Transportation Research Part A: Policy and Practice*, 45(5), 419–434. <https://doi.org/10.1016/j.tra.2011.02.002>

- Cunche, M. (2014). I know your MAC address: targeted tracking of individual using Wi- Fi. *Journal of Computer Virology and Hacking Techniques*, 10(4), 219–227. <https://doi.org/10.1007/s11416-013-0196-1>
- Curran, K., Furey, E., Lunney, T., Santos, J., Woods, D., & McCaughey, A. (2011). An evaluation of indoor location determination technologies - Journal of Location Based Services. *Journal of Location Based Services*, 99999(1), 1–18. <https://doi.org/10.1080/17489725.2011.562927>
- Danielsson, P. E. (1980). Euclidean distance mapping. *Computer Graphics and Image Processing*, 14(3), 227–248. [https://doi.org/10.1016/0146-664X\(80\)90054-4](https://doi.org/10.1016/0146-664X(80)90054-4)
- Dataplicity. (2019). Dataplicity: Remotely control your Raspberry Pi. Retrieved February 24, 2019, from <https://www.dataplicity.com/>
- Deloitte. (2017). *2017 Global Mobile consumer Survey: US edition*.
- Dimitrova, D. C., Alyafawi, I., & Braun, T. (2012). Experimental comparison of bluetooth and wifi signal propagation for indoor localisation. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 7277 LNCS(5533), 126–137. https://doi.org/10.1007/978-3-642-30630-3_11
- Ding, C., & He, X. (2004). K-means Clustering via Principal Component Analysis. In *Proceedings of the 21st International Conference on Machine Learning* (pp. 1–9). Banff. <https://doi.org/10.1007/s10339-009-0337-0>
- Ding, C., Wang, D., Ma, X., & Li, H. (2016). Predicting short-term subway ridership and prioritizing its influential factors using gradient boosting decision trees. *Sustainability (Switzerland)*. <https://doi.org/10.3390/su8111100>
- Douangphachanh, V., & Oneyama, H. (2013). Estimation of road roughness condition from smartphones under realistic settings. In *2013 13th International Conference on ITS Telecommunications, ITST 2013*. <https://doi.org/10.1109/ITST.2013.6685585>
- Dougherty, M., Kirby, H., & Boule, R. (1993). *The use of neural networks to recognise and predict traffic congestion*.
- Duflot, M., Kwiatkowska, M., Norman, G., & Parker, D. (2006). A formal analysis of bluetooth device discovery. *International Journal on Software Tools for Technology Transfer*, 8(6), 621–632. <https://doi.org/10.1007/s10009-006-0014-x>
- Dunlap, M., Li, Z., Henrickson, K., & Wang, Y. (2016). Estimation of Origin and Destination Information from Bluetooth and Wi-Fi Sensing for Transit. *Journal of the Transportation Research Board*, 11–17. <https://doi.org/10.3141/2595-02>
- El-Tawab, S., Oram, R., Garcia, M., Johns, C., & Park, B. B. (2017). Data analysis of transit systems using low-cost IoT technology. *2017 IEEE International Conference on Pervasive Computing and Communications Workshops, PerCom Workshops 2017*, 497–502. <https://doi.org/10.1109/PERCOMW.2017.7917613>
- Erkan, İ., & Hastemoglu, H. (2016). Bluetooth as a traffic sensor for stream travel time estimation under Bogazici Bosphorus conditions in Turkey. *Journal of Modern Transportation*, 24(3), 207–214. <https://doi.org/10.1007/s40534-016-0101-y>

- Etter, A. (2002). *Information Security Reading Room A Guide to Wardriving and Detecting Wardrivers*.
- Ezechina, M. A., Okwara, K. K., & Ugboaja, C. A. U. (2015). The Internet of Things (Iot): A Scalable Approach to Connecting Everything. *The International Journal Of Engineering And Science*, (2014), 9–12. Retrieved from https://figshare.com/articles/The_Internet_of_Things_Iot_A_Scalable_Approach_to_Connecting_Everything/1329665
- Ferris, B., Hähnel, D., & Fox, D. (2006). Robotics: Science and Systems sss6 Philadelphia, PA, USA, August t6-Gaussian Processes for Signal Strength-Based Location Estimation. *Robotics: Science and Systems*. Retrieved from <http://www.roboticsproceedings.org/rss02/p39.pdf>
- Ferro, E., & Potorti, F. (2004). Bluetooth and Wi-Fi Wireless Protocols: *IEEE Wireless Communications Magazine*, 1–24.
- Gao, M., Zhu, T., Wan, X., & Wang, Q. (2013). Analysis of travel time patterns in urban using taxi GPS data. *Proceedings - 2013 IEEE International Conference on Green Computing and Communications and IEEE Internet of Things and IEEE Cyber, Physical and Social Computing, GreenCom-IThings-CPSCoM 2013*, 512–517. <https://doi.org/10.1109/GreenCom-iThings-CPSCoM.2013.101>
- Gao, Q., Ahn, M., & Zhu, H. (2014). Cook's Distance Measures for Varying Coefficient Models with Functional Responses. *Cook's Distance Measures for Varying Coefficient Models with Functional Responses. Technometrics: A Journal of Statistics for the Physical, Chemical, and Engineering Sciences*, 57(2), 268–280., 52(2), 268–280. <https://doi.org/10.1177/0333102415576222>.Is
- Haghani, A., Hamed, M., Sadabadi, K., Young, S., & Tarnoff, P. (2010). Data Collection of Freeway Travel Time Ground Truth with Bluetooth Sensors. *Transportation Research Record: Journal of the Transportation Research Board*, 2160(2160), 60–68. <https://doi.org/10.3141/2160-07>
- Han, B., Hui, P., Kumar, V. S. A., Marathe, M. V., Shao, J., & Srinivasan, A. (2012). Mobile data offloading through opportunistic communications and social participation. *IEEE Transactions on Mobile Computing*, 11(5), 821–834. <https://doi.org/10.1109/TMC.2011.101>
- Harwood, M. (2011). *CompTIA Network+ (N10-004) Certification Guide*. Indianapolis: Pearson Education, Inc.
- Haseman, R., Wasson, J., & Bullock, D. (2010). Real-Time Measurement of Travel Time Delay in Work Zones and Evaluation Metrics Using Bluetooth Probe Tracking. *Transportation Research Record: Journal of the Transportation Research Board*, 2169, 40–53. <https://doi.org/10.3141/2169-05>
- Hedemalm, E. (2017). *Emil Hedemalm Online Transportation Mode Recognition and an Application to Promote Greener Transportation*. Luleå University of Technology.
- Heydt-Benjamin, T. S., Chae, H. J., Defend, B., & Fu, K. (2006). Privacy for public transportation. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 4258 LNCS, 1–19. https://doi.org/10.1007/11957454_1

- Holst, E., & Thyregod, P. (1999). A statistical test for the mean squared error. *Journal of Statistical Computation and Simulation*, 63(4), 321–347. <https://doi.org/10.1080/00949659908811960>
- Hora, J., Dias, T. G., Camanho, A., & Sobral, T. (2017). Estimation of Origin-Destination matrices under Automatic Fare Collection: The case study of Porto transportation system. *Transportation Research Procedia*, 27, 664–671. <https://doi.org/10.1016/j.trpro.2017.12.103>
- HRDC. (2019). Streamline Bus. Retrieved February 15, 2019, from <https://streamlinebus.com/>
- Hu, N., Legara, E. F., Lee, K. K., Hung, G. G., & Monterola, C. (2016). Impacts of land use and amenities on public transport use, urban planning and design. *Land Use Policy*. <https://doi.org/10.1016/j.landusepol.2016.06.004>
- Huh, S. (2014). Coding practice of the Journal Article Tag Suite extensible markup language. *Science Editing*, 1(2), 105–112. <https://doi.org/10.6087/kcse.2014.1.105>
- IEEE Standards Association. (2012). *IEEE Std 802.11ae-2012 Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications IEEE Computer Society*. New York: IEEE.
- International Telecommunication Union. (2013). *SERIES Y: GLOBAL INFORMATION INFRASTRUCTURE, INTERNET PROTOCOL ASPECTS AND NEXT-GENERATION NETWORKS Next Generation Networks – Frameworks and functional architecture models. ITU-T Recommendations*. Geneva. <https://doi.org/http://handle.itu.int/11.1002/1000/11559>
- Internet Society Organization. (2017). Paths to Our Digital Future. Retrieved from <https://future.internetsociety.org/wp-content/uploads/2017/09/2017-Internet-Society-Global-Internet-Report-Paths-to-Our-Digital-Future.pdf>
- Iszaidy, I., Ngadiran, R., Ahmad, R. B., Jais, M. I., & Shuhaizar, D. (2017). Implementation of raspberry Pi for vehicle tracking and travel time information system: A survey. *Proceedings of 2016 International Conference on Robotics, Automation and Sciences, ICORAS 2016*, 1–4. <https://doi.org/10.1109/ICORAS.2016.7872605>
- Jaffe, E. (2015). How San Francisco Got Its New Rider-Friendly Transit Map. Retrieved March 4, 2019, from <https://www.citylab.com/transportation/2015/09/how-san-francisco-got-its-new-rider-friendly-transit-map/403738/>
- Janitza, S., & Hornung, R. (2018). On the overestimation of random forest's out-of-bag error. *PLoS ONE* (Vol. 13). <https://doi.org/10.1371/journal.pone.0201904>
- Ji, Y., Mishalani, R. G., & McCord, M. R. (2014). Estimating Transit Route OD Flow Matrices from APC Data on Multiple Bus Trips Using the IPF Method with an Iteratively Improved Base: Method and Empirical Evaluation. *Journal of Transportation Engineering*. [https://doi.org/10.1061/\(ASCE\)TE.1943-5436.0000647](https://doi.org/10.1061/(ASCE)TE.1943-5436.0000647)
- 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. <https://doi.org/10.1155/2017/6374858>
- Julio, N., Giesen, R., & Lizana, P. (2016). Real-time prediction of bus travel speeds using traffic shockwaves and machine learning algorithms. *Research in Transportation Economics*. <https://doi.org/10.1016/j.retrec.2016.07.019>

- Kalaputapu, R., & Demetsky, M. J. (1995). Modeling Schedule Deviations of Buses Using Automatic Vehicle-location Data and Artificial Neural Networks. *Transportation Research Record: Journal of the Transportation Research Board*.
- Kikuchi, S., & Perincherry, V. (1992). Model To Estimate Passenger Origin- Destination Pattern on a Rail Transit Line. *Transportation Reserach Record* 1349, (2), 54–61.
- Kim, H. (2014). Analysis of variance (ANOVA) comparing means of more than two groups. *The Korean Academy of Conservative Dentistry*, 7658, 74–77.
- Kim, Y., Shin, H., & Cha, H. (2012). Smartphone-based Wi-Fi pedestrian-tracking system tolerating the RSS variance problem. *2012 IEEE International Conference on Pervasive Computing and Communications, PerCom 2012*, (March), 11–19. <https://doi.org/10.1109/PerCom.2012.6199844>
- Kismet Wireless. (2019). Kismet - Kismet. Retrieved February 15, 2019, from <https://www.kismetwireless.net/>
- Kostakos, V., Camacho, T., & Mantero, C. (2013). Towards proximity-based passenger sensing on public transport buses. *Personal and Ubiquitous Computing*, 17(8), 1807–1816. <https://doi.org/10.1007/s00779-013-0652-4>
- Kuderer, M., Gulati, S., & Burgard, W. (2015). Learning driving styles for autonomous vehicles from demonstration. *Proceedings - IEEE International Conference on Robotics and Automation, 2015–June*(June), 2641–2646. <https://doi.org/10.1109/ICRA.2015.7139555>
- Kumar, S. V., Vanajakshi, L., & Subramanian, S. C. (2011). A model based approach to predict stream travel time using public transit as probes. *IEEE Intelligent Vehicles Symposium, Proceedings*, (Iv), 101–106. <https://doi.org/10.1109/IVS.2011.5940413>
- Kurkcu, A., & Ozbay, K. (2017). Estimating Pedestrian Densities, Wait Times, and Flows with Wi-Fi and Bluetooth Sensors. *Transportation Research Record: Journal of the Transportation Research Board*, 2644, 72–82. <https://doi.org/10.3141/2644-09>
- Kwak, S. G., & Kim, J. H. (2017). Introduction Basic Concepts of Central Limit Theorem Central limit theorem: the cornerstone of modern statistics KJA. *Korean Journal of Anesthesiology*, 70(2), 144–156. <https://doi.org/10.4097/kjae.2017.70.2.144>
- Langston, J. (2016). Bluetooth and Wifi sensing from mobile devices may improve bus service. University of Washington website. Retrieved January 20, 2016. <http://www.washington.edu/news/2016/01/20/bluetooth-and-wi-fi-sensing-from-mobile-devices-may-help-improve-bus-service/>
- Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A., Christakis, N., ... Roy, D. (2009). Computational Social Science, 323(5915), 721–723. <https://doi.org/10.1126/science.1167742>
- Lee, R. J., & Sener, I. N. (2016). Transportation planning and quality of life: Where do they intersect? *Transport Policy*, 48, 146–155. <https://doi.org/10.1016/j.tranpol.2016.03.004>
- Lefloch, D., Cheikh, F. A., Hardeberg, J. Y., Gouton, P., & Picot-Clemente, R. (2008). *Real-time people counting system using a single video camera*. Gjøvik University College. <https://doi.org/10.1117/12.766499>

- Letchner, J., Fox, D., & LaMarca, A. (2005). Large-scale localization from wireless signal strength. *Proceedings of the National Conference on Artificial Intelligence*, 1, 15–20. <https://doi.org/10.1016/j.colegn.2010.05.004>
- Li, B. (2009). Markov models for Bayesian analysis about transit route origin-destination matrices. *Transportation Research Part B: Methodological*. <https://doi.org/10.1016/j.trb.2008.07.001>
- Li, Y., & Cassidy, M. J. (2007). A generalized and efficient algorithm for estimating transit route ODs from passenger counts. *Transportation Research Part B: Methodological*. <https://doi.org/10.1016/j.trb.2006.04.001>
- Li, Y., Li, X., & Yoshie, O. (2014). Traffic engineering framework with machine learning based meta-layer in software-defined networks. *Proceedings of 2014 4th IEEE International Conference on Network Infrastructure and Digital Content, IEEE IC-NIDC 2014*, 121–125. <https://doi.org/10.1109/ICNIDC.2014.7000278>
- Lim, C. H., Wan, Y., Ng, B. P., & See, C. M. S. (2007). A real-time indoor WiFi localization system utilizing smart antennas. *IEEE Transactions on Consumer Electronics*, 53(2), 618–622. <https://doi.org/10.1109/TCE.2007.381737>
- Litman, T. A. (2017). Autonomous Vehicle Implementation Predictions. *Traffic Technology International*. <https://doi.org/10.1613/jair.301>
- Liu, C. (2016). *eMarketer's Updated Estimates and Forecast for 2015-2020. US Ad Spending 2015-2020*.
- Liu, H., Darabi, H., Banerjee, P., & Liu, J. (2007). Survey of wireless indoor positioning techniques and systems. *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*, 37(6), 1067–1080. <https://doi.org/10.1109/TSMCC.2007.905750>
- Louppe, G. (2014). *Understanding Random Forests: From Theory to Practice*. University of Liège. <https://doi.org/10.13140/2.1.1570.5928>
- Manzoni, V., Maniloff, D., Kloeckl, K., & Ratti, C. (2010). Transportation mode identification and real-time CO2 emission estimation using smartphones. *SENSEable City Lab, Massachusetts Institute of Technology, Nd*.
- Martin, J., Mayberry, T., Donahue, C., Foppe, L., Brown, L., Riggins, C., ... Brown, D. (2017). A Study of MAC Address Randomization in Mobile Devices and When it Fails. <https://doi.org/10.1515/popets-2017-0054>
- McCord, M. M., Mishalani, R. G., & Wirtz, J. (2006). Passenger Wait Time Perceptions at Bus Stops. *Journal of Public Transportation*, 9, 89–106.
- McCord, M. R., & Mishalani, R. G. (2016). Determining Transit Passenger Boarding-to- Alighting Flows using Mobile Device Wi-Fi Signals : Empirical Results Background : Route-level Passenger Origin-Destination (OD) Flows. In *Ohio Transportation Engineering Conference* (pp. 1–23).
- Meehan, B. (2005). *Travel Times on Dynamic Message Signs*.
- Miao, L. (2015). *Comparative Analysis of Two Clustering Algorithms : K-means and FSDP (Fast Search and Find of Density Peaks)*. San Jose State University.
- Michau, G., Borgnat, P., Pustelnik, N., Abry, P., Nantes, A., & Chung, E. (2015). Estimating link-dependent Origin-Destination matrices from sample trajectories and traffic counts. *ICASSP, IEEE*

- International Conference on Acoustics, Speech and Signal Processing - Proceedings, 2015–Augus*, 5480–5484. <https://doi.org/10.1109/ICASSP.2015.7179019>
- Mishalani, R., Ji, Y., & McCord, M. (2011). Effect of Onboard Survey Sample Size on Estimation of Transit Bus Route Passenger Origin-Destination Flow Matrix Using Automatic Passenger Counter Data. *Transportation Research Record: Journal of the Transportation Research Board*, 2246, 64–73. <https://doi.org/10.3141/2246-09>
- Monsere, C., & Breakstone, A. (2006). Validating dynamic message sign freeway travel time messages with ground truth geospatial data. ... *Record: Journal of ...*, (12), 19– 27. <https://doi.org/10.3141/1959-03>
- Musa, A. B. M., & Eriksson, J. (2012). Tracking Unmodified Smartphones Using Wi-Fi Monitors.pdf. In *ACM Sensys 2012* (pp. 281–294). <https://doi.org/10.1145/2426656.2426685>
- Oransirikul, T., Nishide, R., Piumarta, I., & Takada, H. (2014). Measuring Bus Passenger Load by Monitoring Wi-Fi Transmissions from Mobile Devices. *Procedia Technology*, 18(September), 120–125. <https://doi.org/10.1016/j.protcy.2014.11.023>
- Oyelade, O., Oladipupo, O., & Obagbuwa, I. (2010). Application of k-Means Clustering algorithm for prediction of Students' Academic Performance. *International Journal of Computer Science and Information Security*, 7, 292–295. <https://doi.org/10.1007/s10570-018-1977-y>
- Perk, V., & Kamp, N. (2003). *Handbook of Automated Data Collection Methods for the National Transit Database*. Tallahassee.
- Petre, A., Chilipirea, C., & Baratchi, M. (2016). WiFi tracking of pedestrian behavior Principles of WiFi tracking, 1–24.
- Piech, C. (2013). K Means. Retrieved March 2, 2019, from <http://stanford.edu/~cpiech/cs221/handouts/kmeans.html>
- Pokrajac, D., Borcean, C., Johnson, A., Hobbs, A., Agodio, L., Nieves, S., ... Austin, J. (2009). Evaluation of automated license plate reader accuracy. *9th International Conference on Telecommunications in Modern Satellite, Cable, and Broadcasting Services, TELSIKS 2009 - Proceedings of Paper*, 217–220. <https://doi.org/10.1109/TELSKS.2009.5339422>
- Pomerleau, D. A. (1991). Efficient Training of Artificial Neural Networks for Autonomous Navigation. *Neural Computation*, 3(1), 88–97. <https://doi.org/10.1162/neco.1991.3.1.88>
- Porter, J. D., Kim, D. S., Magaña, M. E., Poocharoen, P., & Arriaga, C. A. G. (2013). Antenna characterization for bluetooth-based travel time data collection. *Journal of Intelligent Transportation Systems: Technology, Planning, and Operations*, 17(2), 142–151. <https://doi.org/10.1080/15472450.2012.696452>
- Praveena, M., & Jaiganesh, V. (2017). A Literature Review on Supervised Machine Learning Algorithms and Boosting Process. *International Journal of Computer Applications*, 169(8), 32–35. <https://doi.org/10.5120/ijca2017914816>
- Psarros, I., Kepaptsoglou, K., & Karlaftis, M. (2015). An Empirical Investigation of Passenger Wait Time Perceptions Using Hazard-Based Duration Models. *Journal of Public Transportation*, 14(3), 109–122. <https://doi.org/10.5038/2375-0901.14.3.6>

- Purser, K. (2016). *Exploring Travel Time Reliability Using Bluetooth Data*. California Polytechnic State University.
- Quayle, S., Koonce, P., DePencier, D., & Bullock, D. (2010). Arterial Performance Measures with Media Access Control Readers. *Transportation Research Record: Journal of the Transportation Research Board*, 2192, 185–193. <https://doi.org/10.3141/2192-18>
- Ranganai, E., Van Vuuren, J. O., & De Wet, T. (2014). Multiple case high leverage diagnosis in regression quantiles. *Communications in Statistics - Theory and Methods*, 43(16), 3343–3370. <https://doi.org/10.1080/03610926.2012.715225>
- Raspberry Pi Foundation. (2019). Download Raspbian for Raspberry Pi. Retrieved February 15, 2019, from <https://www.raspberrypi.org/downloads/raspbian/>
- Rekimoto, J., Miyaki, T., & Ishizawa, T. (2007). LifeTag: WiFi-based continuous location logging for life pattern analysis. *Science*, 4718, 35–49. https://doi.org/10.1007/978-3-540-75160-1_3
- Remias, S. M., Hainen, A. M., Mathew, J. K., Vanajakshi, L., Sharma, A., & Bullock, D. M. (2014). Travel Time Observations Using Bluetooth MAC Address Matching : A Case Study on the Rajiv Gandhi Roadway : Chennai , India. *Sādhana*.
- Robert Peccia & Associates, & Alta Planning + Design. (2017). *Bozeman Transportation Master Plan*. Bozeman.
- Rodrigue, J.-P., Comtois, C., & Slack, B. (2017). *The geography of transport systems* (Fourth). Retrieved from https://transportgeography.org/?page_id=8589
- Rouse, M. (2007). What is GSM (Global System for Mobile communication)? - Definition from WhatIs.com. Retrieved February 11, 2019, from <https://searchmobilecomputing.techtarget.com/definition/GSM>
- Salek, M.-D., & Machemehl, R. B. (1999). *Characterizing Bus Transit Passenger Waiting Times*. University of Texas at Austin (Vol. SWUTC/99/1). <https://doi.org/10.1007/s10969-008-9058-3>
- Salyers, D. C., Striegel, A. D., & Poellabauer, C. (2008). Wireless reliability: Rethinking 802.11 packet loss. *2008 IEEE International Symposium on A World of Wireless, Mobile and Multimedia Networks, WoWMoM2008*, 1–4. <https://doi.org/10.1109/WOWMOM.2008.4594875>
- Schneider, A., Hommel, G., & Blettner, M. (2010). Linear regression analysis. *Dtsch Arztebl International*, 3(2), 776–782. <https://doi.org/10.5455/jmood.20130624120840>
- Shahid, N., Perraudin, N., Kalofolias, V., Puy, G., & Vanderghenst, P. (2016). Fast Robust PCA on Graphs. *IEEE Journal on Selected Topics in Signal Processing*, 10(4), 740–756. <https://doi.org/10.1109/JSTSP.2016.2555239>
- Statista. (2019). Smartphone penetration in the US (share of population) 2010-2021 | Statistic. Retrieved February 16, 2019, from <https://www.statista.com/statistics/201183/forecast-of-smartphone-penetration-in-the-us/>
- Tanioka, K., & Yadohisa, H. (2012). Effect of Data Standardization on the Result of k- Means Clustering. In *Proceedings of the 34th Annual Conference of the Gesellschaft für Klassifikation e. V.* (p. 613). <https://doi.org/10.1007/978-3-642-24466-7>
- Taunya Fagan. (2019). Bozeman Demographics: Population, Weather, Economy, Gallatin County. Retrieved February 14, 2019, from <https://www.taunhafagan.com/bozemandemographics/>

- Toqué, F., Côme, E., Mahrsi, M. K. El, & Oukhellou, L. (2016). Forecasting dynamic public transport origin-destination matrices with long-short term memory recurrent neural networks. In *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC* (pp. 1071–1076). Institute of Electrical and Electronics Engineers Inc. <https://doi.org/10.1109/ITSC.2016.7795689>
- Varun, S. S., Singh, C. V., & Nagaraj, R. (2013). Conceptualization of an Integrated Positioning & Time Synchronization System for GPS / INS Using VHDL / FPGA Tools for Marine Applications. *The International Journal Of Engineering And Science*, 90–96.
- Videa Martinez, A. A. (2019). Inference of passenger ridership, O-D flows, wait times, and travel times using Wi-Fi and GPS signals (Thesis – available on Montana State University Library website).
- Wang, C. S., & Yan, Y. D. (2002). The IEEE 5* International Conference on Intelligent Transportation Systems 3 - 6 September 2002, Singapore. *Transformation*, (September), 904–909.
- Wang, H., Calabrese, F., Di Lorenzo, G., & Ratti, C. (2010). Transportation Mode Inference from Mobile Phone Call Detail Records. In *Annual Conference on Intelligent Transportation Systems* (pp. 318–323). Madeira Island: IEEE. <https://doi.org/10.1109/ITSC.2010.5625188>
- Wang, Y., Malinovskiy, Y., Street, N. E., Neeley, M., Hammond, P. J., & Hall, M. (2011). Error Modeling and Analysis for Travel Time Data Obtained from Bluetooth MAC Address. *Department of Civil and Environmental Engineering, University of Washington.*, (1), 82.
- Wasson, J. S., Sturdevant, J. R., & Bullock, D. M. (2008). Real-time travel time estimates using media access control address matching. *ITE Journal (Institute of Transportation Engineers)*, 78(6), 20–23.
- Wiering, M. (2000). Multi-Agent Reinforcement Learning for Traffic Light Control. *Proceedings of the 17th International Conference on Machine Learning*.
- World Population. (2018). Bozeman, Montana Population 2019 (Demographics, Maps, Graphs). Retrieved February 14, 2019, from <http://worldpopulationreview.com/us-cities/bozeman-population/>
- Yang, Z., Wu, C., & Liu, Y. (2012). Locating in Fingerprint Space: Wireless Indoor Localization with Little Human Intervention. In *Mobicom '12 Proceedings of the 18th annual international conference on Mobile computing and networking*. <https://doi.org/10.1145/2348543.2348578>
- Yasin, A. M., Karim, M. R., & Abdullah, A. S. (2010). Travel time measurement in real- time using automatic number plate recognition for Malaysian environment. *Journal of the Eastern Asia Society for Transportation Studies*, 8(January), 1738–1751. <https://doi.org/10.1016/j.ijhm.2014.10.013>
- Zanca, G., Zorzi, F., Zanella, A., & Zorzi, M. (2008). Experimental comparison of RSSI- based localization algorithms for indoor wireless sensor networks. *Proceedings of the Workshop on Real-World Wireless Sensor Networks - REALWSN '08*, 1. <https://doi.org/10.1145/1435473.1435475>
- Zha, H., Ding, C., Gu, M., He, X., & Simon, H. (2001). Spectral relaxation for k-means clustering. *Advances in Neural Information Processing Systems*, 14, 1057–1064. Retrieved from <papers2://publication/uuid/16FFAB42-9D62-4F7D-889E-9A63C19B5A10>

- Zhang, G. Y., Zhang, C. X., & Zhang, J. S. (2010). Out-of-bag estimation of the optimal hyperparameter in SubBag ensemble method. *Communications in Statistics: Simulation and Computation*, 39(10), 1877–1892. <https://doi.org/10.1080/03610918.2010.521277>
- Zhang, H., Ritchie, S. G., & Lo, Z.-P. (2007). Macroscopic Modeling of Freeway Traffic Using an Artificial Neural Network. *Transportation Research Record: Journal of the Transportation Research Board*. <https://doi.org/10.3141/1588-14>