# Mobile Device Identifier Data Collection and Analysis for Transportation Intelligence Purposes: Applications, Uncertainty and Privacy

Yegor Malinovskiy

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington
2013

Reading Committee:
Yinhai Wang, Chair
Scott Rutherford
Phillip Hurvitz

Program Authorized to Offer Degree:
Department of Civil and Environmental Engineering

University of Washington

University of Washington

## Abstract

Mobile Device Identifier Data Collection and Analysis for Transportation Intelligence Purposes: Applications, Uncertainty and Privacy

Yegor Malinovskiy

Chair of the Supervisory Committee:

Professor Yinhai Wang

Department of Civil and Environmental Engineering

Travel evaluation metrics have been historically biased towards motorized modes, which dominate land transportation choices and are partially responsible for numerous environmental and health issues facing our society today. Encouraging active travel solutions is seen as a means of improving sustainability, health and cohesiveness of a community. Unfortunately, information regarding volume, trip origin and destination, travel time and personal interactions is difficult to obtain due to a lack of sensor infrastructure and unrestricted movement of these modes. Therefore, information is often limited to annual surveys and model estimates which are insufficient to address the increasing needs of sustainable planning and large scale behavior

studies.

The ubiquity of mobile devices, coupled with their need to communicate wirelessly, provides a wealth of data that, if properly handled, can be used to quickly enhance understanding and recognition of transportation patterns. This data provides an opportunity to create a very low maintenance sensor infrastructure that is readily scalable and is easy to deploy and use for a number of transportation purposes, from long-term city planning to day to day traffic operations. Of particular interest are the spatial and temporal patterns that evolve as a result of daily human activity in very dense urban cores and campuses, where non-motorized modes dominate and mobile devices are highly prevalent. Traditional sensing approaches have often failed to capture non-motorized travel movements, resulting in data bias. Mobile device data has been viewed as a potential solution to these bias issues. The research conducted within this dissertation focuses on discussing and developing Bluetooth Media Access Control (MAC) based travel data collection approaches and their implications. The challenges of working with opportunistically collected data and the resulting uncertainties are discussed and a number of approaches for mitigating them are proposed. Specifically, the work contained provides a MAC data collection analysis framework, develops algorithms and techniques for the reduction of uncertainty in MAC address-based travel data collection approaches and proposes and evaluates a novel pedestrian data collection approach using an app-based MAC sensing approach.

# Table of Contents

Table of Contents ..... 1

List of Figures ..... 7

List of Tables ..... 14

ACKNOWLEDGMENTS ..... 15

Chapter 1 Introduction ..... 18

1.1 Problem Statement ..... 18

1.1.1 Data Acquisition Costs ..... 19

1.1.1 Personalized Transportation ..... 20

1.1.2 Data Bias ..... 20

1.1.1 Non-motorized Transportation Data Needs ..... 21

1.2 Background ..... 23

Chapter 2 State of the Art ..... 26

2.1 Bluetooth Travel Measurement ..... 26

2.1.1 Motorized Travel ..... 26

2.1.2 Non-motorized Travel ..... 28

2.2 Mobile Sensing ..... 28

2.2.1 Devices as Location Probes ..... 29

2.2.2 Devices as Environment Sensors ..... 30

2.2.3 Trajectory Inference ..... 31

2.3 Privacy and Public Participation ..... 32

2.3.1 Privacy Preservation in Trajectory Data ..... 32

2.3.2 Public-driven Data Collection Efforts ..... 32

Chapter 3 Study Significance ..... 34

3.1 Point Sensor Approaches ..... 34

3.2 Re-Identification Approaches ..... 35

3.3 Implications of Re-identification ..... 37

3.4 Research Objectives ..... 38

3.5 Study Scope ..... 39

3.6 Dissertation Organization ..... 40

Chapter 4 Collection and Applications of Mobile Device Identifier Data ..... 42

4.1 Introduction to Device Sensing ..... 42

4.1.1 Detection Basics ..... 43

4.1.2 System Design ..... 44

4.1.3 Current Design Overview ..... 47

4.1.4 Communications Design ..... 49

4.2 Single-node Data Collection Paradigm Applications ..... 51

4.2.1 Data collection approach ..... 51

4.2.2 Passenger Wait Time Estimation Application ..... 51

4.2.3 Pedestrian Dwell Behavior Application ..... 63

4.2.4 Single Sensor Paradigm Issues ..... 73

4.3 Corridor Data Collection Paradigm Applications ..... 75

4.3.1 Data collection approach ..... 75

4.3.2 Pedestrian Travel Behavior Application ..... 75

4.3.3 Highway Segment Application ..... 84

4.3.4 Corridor Paradigm Issues ..... 90

4.4 Mobile-node Data Collection Paradigm Applications ..... 91

4.4.1 Introduction ..... 91

4.4.2 Mobile Node Approach Simulation ..... 92

4.4.3 Pedestrian Route Estimation Application ..... 98

4.4.4 Mobile Paradigm Issues ..... 107

Chapter 5 Strategies for Systematic Reduction of Population, Temporal and Spatial Data Uncertainty 108

5.1 Data Uncertainty in MAC sensing ..... 108

5.2 Analysis Tools ..... 110

5.2.1 MAC Matching and Filtering Engine ..... 110

5.2.2 Corridor sensor comparison tab ..... 112

5.2.3 Mobile device routing tab ..... 114

5.3 Reducing Population Uncertainty ..... 115

5.3.1 Site selection and MAC-brand based filtering ..... 116

5.3.2 Outliers and Filtering ..... 117

5.4 Reducing Temporal Uncertainty ..... 120

5.4.1 Study Site ..... 120

5.4.2 MAC Address Data Acquisition ..... 125

5.4.3 Test Configurations ..... 128

5.4.4 License Plate Data Acquisition ..... 131

5.4.5 Experiment Results ..... 131

5.4.6 Error Analysis Westbound ..... 133

5.4.7 Error Analysis Eastbound ..... 140

5.4.8 Configuration Comparison ..... 144

5.4.9 Configuration Comparison Summary ..... 152

5.4.10 Conclusions and Recommendations ..... 152

5.5 Reducing Spatial Uncertainty ..... 155

5.5.1 Inference of plausible paths ..... 156

5.5.2 Verification and Effectiveness ..... 165

Chapter 6 Privacy Concerns in Mobile Device Data ..... 176

6.1 Overview of Inherent Privacy Issues ..... 176

6.2 Data sensitivity (identity disclosure) and indefinite storage ..... 178

6.3 Ease of collection ..... 180

Chapter 7 Conclusions and Future Research..... 182

7.1 Summary of Research ..... 182

7.2 Research Contributions..... 183

7.3 Future Research ..... 186

Bibliography ..... 189

Reprint Permissions ..... 204

APPENDIX A ..... 205

# List of Figures

Figure 4-1: Detection paradigms ..... 42

Figure 4-2: MACAD evolution..... 46

Figure 4-3: STAR Lab Bluetooth detectors (MACAD device) used in this study ..... 48

Figure 4-4: Bluetooth data collection and distribution diagram ..... 50

Figure 4-5: Short-range Bluetooth detector. ..... 52

Figure 4-6: a) Redmond Park and Ride test site b) Northgate Park and Ride test site ..... 53

Figure 4-7: Device Manufacturer Populations at Each Transit Center ..... 57

Figure 4-8: Wait Times by Transit Stop ..... 59

Figure 4-9: Wait Times by Hour of Day ..... 61

Figure 4-10: a) Downtown Montreal site b) Seattle UW Campus site ..... 64

Figure 4-11: Device brand distributions at both study sites..... 66

Figure 4-12: Dwell times at the Montreal Chinatown location ..... 69

Figure 4-13: Dwell time distributions at the Montreal sensor locations..... 70

Figure 4-14: Dwell times at the University of Washington, Seattle location ..... 72

Figure 4-15: Dwell time distributions at the Seattle sensor locations ..... 73

Figure 4-16: Travel times between the two sensor locations in downtown Montreal ..... 77

Figure 4-17: Montreal travel time distributions in both directions..... 78

Figure 4-19: Distribution of travel times to and from Husky Stadium in Seattle..... 81

Figure 4-20: Travel time and dwell time on a day before Graduation on the UW Campus. ..... 84

Figure 4-21: a) Selected freeway test corridor on SR-520. b) Bluetooth sensor (left) and portable ALPR (right) used to collect travel time data at the 24th Ave location. ..... 85

Figure 4-22: SR-520 freeway test ..... 89

Figure 4-24: Mobile sensing simulation architecture ..... 94

Figure 4-25: Simulation of 892 entities on a simple grid network ..... 95

Figure 4-26: Relationship between the number of observers and the proportion of visible population detected ..... 97

Figure 4-27: A Motorola Droid handset running the Mobile Monitor application ..... 99

Figure 4-28: Collected trajectories on UW campus on 4/20/2011 1:10pm to 2:00pm using 4 observers ..... 101

Figure 4-29: Unfiltered Device Distance vs. Travel Times ..... 103

Figure 4-30: Filtered Device Distance vs. Travel Times ..... 104

Figure 4-31: Filtered Speed Distribution of Discovered Devices ..... 106

Figure 4-32: Devices Detected by Minute of Experiment ..... 106

Figure 4-33: Manual Data Collection Results ..... 107

Figure 5-1: Inherent issues in opportunistic MAC sensing paradigms..... 109

Figure 5-2: MAC Matching and filtering engine schema and interface. .....112

Figure 5-3: Corridor Sensor Comparison tab in DriveNET .....114

Figure 5-4: Mobile Device Routing tab in DriveNET .....115

Figure 5-5: Isolating pedestrian mode from MAC data.....119

Figure 5-6: Study route on SR-522 [Image from maps.google.com] ..... 121

Figure 5-7: Spectrum average for 170th St NE. ..... 123

Figure 5-8: Spectrum average for NE 61st Ave. ..... 124

Figure 5-9: Spectrum noise image ..... 125

Figure 5-10: Lane-ft coverage of a 12 dBi directional sensor at NE 170th St..... 127

Figure 5-11: Sensor configurations [Background images from maps.google.com] ..... 130

Figure 5-12: Sensor detection zones ..... 132

Figure 5-13: Travel time comparison westbound SR-522 (ALPR – blue, BT – red + orange) (1hr averages) ..... 135

Figure 5-14: Westbound SR-522 error and volume (1hr averages) ..... 137

Figure 5-15: Travel time comparison eastbound SR-522 (ALPR – blue, BT – red + orange) (1hr averages) ..... 141

Figure 5-16: Eastbound SR-522 error and volume (1hr averages) ..... 143

Figure 5-17: a) Westbound detection rates normalized by ALPR volume b) Westbound matching rates normalized by ALPR volume ..... 148

Figure 5-18: a) Eastbound detection rates normalized by ALPR volume b) Eastbound matching rates normalized by ALPR volume ..... 151

Figure 5-19: Inference of plausible paths ..... 158

Figure 5-20: Diagram of route imputation system..... 160

Figure 5-21: Distance threshold (in meters) for certain/uncertain path discrimination.. 161

Figure 5-22: Imputed plausible paths from campus experiment conducted on 04/20/2011 ..... 165

Figure 5-23: Static sensor mounting locations on University of Washington campus ... 167

Figure 5-24: Comparison of heatmaps of MAC devices detected on campus..... 169

Figure 5-25: Ray charts depicting pairwise flows for each static sensor location..... 171

Figure 5-26: Percent of correctly matched MACs without and with path reconstruction ..... 174

Figure 5-27: a) Percent of correctly matched MACs by distance threshold with popularity weights of 1250 to 7500..... 175

# List of Tables

Table 4-1: Comparison of Bluetooth Estimated Wait Time to Measured Wait Times 58

Table 5-1: Bluetooth device mounting and antenna configurations ..... 128

Table 5-2: Westbound error regression model results..... 139

Table 5-3: Eastbound error regression model results..... 143

Table 5-4: Westbound 15-minute aggregate error statistics by configuration ..... 146

Table 5-5: Eastbound 15-minute aggregate error statistics by configuration ..... 149

Table 5-6: Observer sensor visit itineraries ..... 168

Table 5-7: Relative errors in pairwise flows for mobile and static Bluetooth data... 172

# ACKNOWLEDGMENTS

Although the final few months of this dissertation have been a complete blur, this page allows me to pause for a moment and remember all those that helped me get here. It goes without saying that my adviser, Dr. Yinhai Wang deserves much credit in not only helping me achieve what is contained in the following pages, but also for leading me through some of the most formative years of my life. The opportunities granted by working with STAR Lab have been nothing short of life-changing and I can only hope to continue this partnership in the years to come.

I am particularly grateful to the members of my reading committee, who have served as additional role models and collaborators throughout the years, guiding and assisting me not only in this dissertation, but a host of other pursuits as well.

I would also like to thank the members of STAR Lab, each of whom I have been incredibly fortunate to have met and the many that I have had the pleasure to work with. I have witnessed the lab grow from a few members to nearly two dozen members and it has been incredibly exciting to be a part of this journey. Without the help of lab members like Ms. Sa Xiao, Mr. Xiaolei Ma, Ms. Xiaoyue Liu, Dr. Kari Watkins and Dr. Runze Yu, many of the aspects of this dissertation would not have been possible and my time spent would not have been as productive (and fun!)

I am also deeply indebted to many STAR Lab visiting scholars; Dr. Un Kun Lee, Dr. Hua Wang, Dr. Xiaofeng Chen and Ms. Bahar Araghi for their invaluable feedback, collaboration and friendship through the numerous phases of this work.

I am also grateful to the Valle Program, run by Dr. Scott Rutherford and Dayna Cole, which allowed me to reflect and refine many of the ideas contained in this dissertation while hosted by Dr. Ingmar Andreasson and Dr. Oded Cats at the Kungliska Tekniska Hogskolan.

Finally, I would like to thank my family and loved ones for supporting me and encouraging me on every step of the way. It is their love that fuels my passion.

# DEDICATION

To all who have supported me through these years.

Thank you.

# Chapter 1 Introduction

## 1.1 Problem Statement

Three primary issues motivate the proposed research. The first, and perhaps most obvious, is the increasing costs and dwindling resources of the nation's transportation network. The capability to reduce data collection costs by outsourcing much of the work to the end user is an attractive concept, both from a scaling and maintenance perspective. The second issue addressed is the bias currently present in much of transportation data, which tends to be primarily available for cars and trucks. Meanwhile, the transportation system is continuing to diversify, with many US cities seeking to increase shares on non-motorized and transit travel. The ability to collect data from individuals and the devices they carry presents the opportunity to represent all modes and means of travel, fully capturing the complexity of the transportation system. This, in turn allows for appropriate representation of these modes enabling a more balanced resource investment scheme. The third issue is the increasing interest in personal accessibility solutions, or the ability to tailor the existing transportation options to individuals given sufficient information. This concept allows the end users to benefit from the ubiquitous data directly, by providing them with immediate, relevant information to optimize their traveling experience. Each of the issues presented is described in greater detail in the following subsections.

### 1.1.1 Data Acquisition Costs

Over sixty percent of the world's population owns a mobile telephone, with many developed countries reaching near one hundred percent ownership rates (Mohavedi Naini, 2010). As mobile devices become more complex and their need to communicate with each other grows, there is a growing stream of information that is generated around each mobile device owner. The ubiquity of such devices creates interesting opportunities in human behavior and travel characteristics evaluations, a concept that is quickly gaining momentum in a number of scientific communities. For example, this flow of information is opening new means of analyzing public spaces such as shopping malls, zoos and airports (Bullock et al., 2010), as well as entire towns (Kiukkonen et al., 2010). Of the several available data exchange protocols available, Bluetooth and Wi-Fi have become by far the most popular. Bluetooth is a short-range communication technology that has been widely used in our daily life for mobile-to-mobile device communications. Wi-Fi has, in turn, ensured that many mobile devices have access to the Internet. Most personal electronic devices, such as personal digital assistants (PDAs) and cell phones, have embedded Bluetooth and Wi-Fi modules that can communicate with other peripheral electronic devices and the Internet.

Analogous to most communicating devices, each Bluetooth and WiFi device has a globally unique 48-bit Media Access Control (MAC) address. It is this unique identifier that provides potential for an easy means of obtaining travel characteristics of a given human-populated network (Malinovskiy et al., 2009, Wasson et al., 2008). Not only can the overall

population of the system be estimated (Mohavedi Naini, 2010), but the travel times, routing choices and interactions can be analyzed to provide an overall understanding of spatial and temporal patterns. Moreover, this approach is as universal as the communication protocols that power it – data collection can still be performed in locations that have no infrastructure.. This allows for a standard method of evaluation of urban core travel across the globe.

### 1.1.1 Personalized Transportation

As the new device-based data collection approach enables data aggregation across the population, it also allows for data dissemination across the very same channel as well. The users that have used their devices to submit their information can now receive information not only about themselves, but also about their community as a whole. This enables a more personalized experience, as it allows for interaction between the transportation system and the user. Although this area is still very new, research regarding what these services will (or should) look like and their effect on the community has already began (Tuominen and Ahlqvist, 2010; Passos and Rosetti, 2009)

### 1.1.2 Data Bias

Relying on mobile devices for travel data acquisition also provides an opportunity to pursue information parity for sustainable modes. For nearly a century, transportation metrics have favored motorized means by focusing on measures like mobility (number of miles travelled) and traffic volume counts, which do not describe the entirety of the transportation

system (Litman, 2008). As a result, little attention has been paid to pedestrian and bicycle facilities until recent efforts have highlighted these modes as sustainable alternatives. The development of infrastructure-free information gathering techniques based on mobile devices allows communities to begin to collect data on these sustainable modes. However, the lack of a data collection framework and a systematic approach stymies these efforts. It is the aim of this study to develop a precedent and framework for such data gathering and analysis.

### 1.1.1 Non-motorized Transportation Data Needs

To understand the need for a non-motorized transportation data framework, we can first consider which questions are being asked by communities nationwide. A 2005 FHWA report lists the following two questions as being of primary interest (Schneider et al., 2005):

(1) Where is pedestrian and bicycle activity taking place?

(2) What effect does facility construction have on levels of bicycling and walking?

While these questions can be answered with a mobile device network, it can be argued that the questions themselves are a product of the available means of data collection. Furthermore, these questions represent only a small subset of interested users; other questions arise from economists, real-estate developers and business owners. When considering a new wave of mobile sensor infrastructure, new questions can be asked that consider route choice and

travel time (Li and Tsukaguchi, 2005), which can be used for a myriad of purposes, from walk-in customer estimation (Borgers and Timmermans, 1986) to large event planning and evacuations (Gräßle, F. and Kretz, 2010). Pedestrian and bicycle accessibility has also become a factor in home buying and renting decisions, as demonstrated by the efforts of online real estate analysis products such as WalkScore (Cortright, 2009). In all, a means of easily estimating the quantity and spatial and temporal distribution of pedestrians and cyclists has broad implications across a number of fields.

Present data collection approaches are limited to surveys, which are either administered on location or via a broad distribution (sometimes including a mobile GPS-based component), manual counts, which involve field data collection by personnel and automatic spot counts, achieved by either infra-red trip-line sensors, or, in the case of cyclists – inductance loops. Video-based data collection methods that are capable of counts as well as localized route choice are also under development (Kong et al., 2006, Malinovskiy et al., 2008). Aside from expensive, stated preference-based, surveys, none of these approaches provide network-wide travel information. Furthermore, due to the costs of many of these approaches, communities often conduct studies on an annual basis, picking a particular location set and day of the year to act as a surrogate for overall performance (Alta Planning + Design, 2006). This approach not only is capable of producing non-representative results due to climate variations, but also does not provide a clear trend line that can be analyzed for effective improvements in infrastructure or policy. Furthermore, the limited spatial scope of the studies often leaves much of the network

without reliable data.

Development of a cost-effective data collection framework that relies on existing mobile phone infrastructure would alleviate many of these concerns, providing continuous, rich data. This is a chance to quickly address the current disconnect between community planning ambitions and available active travel knowledge, while opening doors for additional investigations that could highlight epidemiological issues, cultural behavior, economic impacts and community evacuation strategies.

## 1.2 Background

Although the MAC address-based collection techniques are becoming more prevalent, there are some drawbacks to their use. Relatively small sample size is an issue for some purposes – most studies using MAC address matching have found that they are able to match somewhere between five to ten percent of the total volumes (Malinovskiy et al., 2010). An additional and perhaps most serious issue is the ambiguity of accuracy due to the inherent properties of the MAC address broadcast protocols. One of the most common protocols is known as Bluetooth, published by Special Interests Group (SIG). This protocol is common in mobile telephones and has been the focus of MAC address-based travel time estimation. The ambiguity of accuracy of

the use of the Bluetooth protocol for travel time measurement comes from the random frequency hopping characteristic of the protocol. As the protocol was designed to function in the same 2.4 GHz band as WiFi, a frequency hopping mechanism was implemented to prevent interference (Special Interests Group, 2010). The constantly changing frequency mandated by the Bluetooth protocol could delay the device connection time by up to 10.24 seconds, leaving room for ambiguity in exact device arrival time, characterized as the tem,poral uncertainty component below. This "connection time" complication is further exacerbated by the variety of ranges that a receiving Bluetooth sensor device may have.

In short the issues that currently affect device-based data collection are as follows:

- Sample (population) uncertainty: This type of error results from the sampling process of the devices present. First, multiple devices may belong to a single entity, biasing results. Second, since devices may be present on pedestrians, cyclists, in vehicles as well as on busses, there is potential for error in mode determination – it is not possible to tell which mode a record belongs to by MAC address alone.

- Temporal uncertainty: A device can be detected anytime in a given time range after it enters the detection zone. It can also be missed entirely or be detected multiple times depending on the time it stays in the detectable area. The time until its first detection is determined by numerous factors, such as protocol specifics, antenna strength and

interference. The location of detection within the detectable range is a property of the device speed, antenna range and temporal error.

- Spatial uncertainty: Since a device is re-identified only at certain locations, there is a certain amount of uncertainty that is involved in-between detections – when sensors are mounted along a short corridor with few alternate paths, the ambiguity is low, however, when there are a number of alternative paths, or the sensor locations are no longer static, the actual routes of the detected devices are not known and must be imputed.

# Chapter 2 State of the Art

Research regarding the use of ubiquitous devices for data collection has been primarily focused on three directions: 1) using devices as location probes to relay position and speed for mobility analysis; 2) treating devices as environment sensors, extending the location probe paradigm to record information about the device's surroundings and; 3) detecting bypassing device communications for the purposes of counting and travel time collection via matching.

## 2.1 Bluetooth Travel Measurement

### 2.1.1 Motorized Travel

In the transportation field, much of the focus for collecting data from ubiquitous devices has been on the MAC identifiers broadcast by the Bluetooth protocol. Many Bluetooth devices, such as headsets, are, by default, set in the discovery mode and can be discovered by other Bluetooth devices inquiring for Bluetooth connections. In particular, the motorized transportation community has become increasingly interested in Bluetooth tracking for the collection of travel time data using dedicated, static sensors (Ahmed et al., 2008; Wasson et al., 2008; Tarnoff et al.,

2009; Haseman et al., 2010, Haghani et al., 2010 and Quayle et al., 2010). Many research papers focused on utilizing Bluetooth for communications in Intelligent Transportation Systems (ITS) (Bhagwat, 2001, Bechler et al, 2001, Aloi et al, 2003, and Sugiura and Dermawan, 2005). Ahmed et al. (2008) may be the first group that used Bluetooth MAC address for vehicle traffic monitoring. The Bluetooth MAC address associated with a probe vehicle was tracked by a Bluetooth and Wi-Fi-based mesh network. A Bluetooth MAC address matching method for travel time collection using static sensors was developed and tested by the Indiana Department of Transportation and Purdue University (Wasson et al., 2008). Tarnoff et al. (2009) demonstrated and evaluated a Bluetooth MAC address detector developed by the University of Maryland at the 88th Annual Meeting of Transportation Research Board in 2009. Evaluation of accuracy of Bluetooth-based travel time measurements using static sensors have also been conducted, with encouraging results, most travel times were well within 10% of the ground truth (Malinovskiy et al. 2010, Malinovskiy et al. 2011).

Additional work is emerging regarding using Bluetooth sensors mounted in vehicles, in particular work by Ruppe et al. (2012) and Versichele et al., (2012). Versichele employs a Bluetooth sensor mounted within a mobile vehicle to monitor crowd movement in the 2011 Tour of Flanders and provides interesting insights into mapping crowds from a single mobile platform. Ruppe on the other hand, envisions a network of vehicle-based mobile Bluetooth sensors that are able to capture other bypassing devices. This emerging work helps assert the utility and future of some of the mobile Bluetooth-based approaches investigated in this dissertation.

### 2.1.2 Non-motorized Travel

Research regarding pedestrian and bicyclist travel data collection via Bluetooth is far more limited. In one of the earliest studies, O'Neill et al. (2006) focused on correlating "gatecounts," or trip-line counts of pedestrians and Bluetooth devices detected in the area of the count. In this study, it was found that about 7% of pedestrians detected were carrying Bluetooth devices. The number of devices detected grew linearly with the amount of pedestrians present. Network approaches to multi-modal data collection using Bluetooth have also received relatively little attention. A conference proceeding by Barberis et al. (2006) outlines the concept of Bluetown, a fully integrated data collection network based on Bluetooth beacons. The authors suggest creating an ad-hoc network of Bluetooth sensors that are tied into groups by central nodes, capable of relaying the acquired travel time info from each sensor into a main database. Although the possibility of collecting data from multiple modes is mentioned, the authors do not delve deeper.

## 2.2 Mobile Sensing

As the concept of ubiquitous computing develops from its nascent vision (Weiser, 1991), the applications of such an infrastructure are quickly coming to light. Conceptual work was begun laying the necessary foundation that will be used to support the structure, functions and limitations of the ubiquitous network. Specifically, mobile device sensors carried by users are

envisioned to become important sources of data for anything from traffic conditions and noise pollution to air quality and population health (Cuff et al., 2008; Abdelzaher et al., 2007; Lane et al., 2010; Kanjo, 2009; Kansal, 2007.). Although the full potential of ubiquitous sensing may be far on the horizon, a few applications and experiments using ubiquitous computing devices have begun to appear. Most current approaches focus solely on internal features of the device, treating it as a probe, primarily collecting information about its location and speed. However, extensions to this approach begin to treat the devices as sensors, able to collect external environment data such as noise, air quality and surrounding devices. Thus, the devices are examined as primarily location probes and, increasingly, environment sensors.

### 2.2.1 Devices as Location Probes

Although some work focuses on general mobility, examining city-wide profiles (Bayir et al., 2009), the majority of probe-based analysis is currently done for vehicular movements. Probe vehicle based analysis has recently become much more affordable due to increased use of GPS among fleet vehicles as well as the capability of purchasing GPS data from routing service providers such as TomTom or Google. While individual, representative, "pilot" vehicle results can be very accurate, results coming from fleet services such as taxis and delivery trucks may be very different, depending on the number of stops the driver makes. Additional concerns can be raised for GPS data coming from freight trucks, as their speeds tend to be different from passenger cars under identical conditions. Another potential drawback of using GPS probe vehicle data is the relatively small sample size that can be attained. Test vehicle runs often

represent an insignificant fraction of the total volumes and fleet-based GPS penetration rates are also quite low if one considers the size of the whole traffic population. However, commercial data provides such as Inrix, have begun to synthesize historical data with current probe vehicle information in an attempt to improve quality (Inrix, 2011).

### 2.2.2 Devices as Environment Sensors

Some of the first uses of ubiquitous data have been approaches to try study social spaces and interactions (Eagle and Pentland, 2005). These approaches used WiFi or Bluetooth protocols to determine and record daily encounters with surrounding individuals. Research conducted by Paulos et al. looked at a study of "familiar strangers," where commuters were given a chance to recognize the recurring strangers that surround them from day to day using Bluetooth-enabled sensors and phone applications. The sole purpose of the study was to illuminate the anonymous relationships we have with our unknown neighbors, but demonstrates the potential for such applications (Paulos and Goodman, 2004). More recent work done by the Lausanne Data Collection Campaign in Switzerland continues to explore social spaces (Aad and Niemy, 2010; Montoliu and Gatica-perez, 2010). Working with Nokia, the group has been able to demonstrate social behavior metrics can be obtained from device activities (taking photos, sending text messages, sampling audio) and inter-device communication behavior (Kiukkonen et al., 2010). Chaintreau et al. looked at the inter-contact time between devices as a measure of transfer opportunities, and found that the average number of Bluetooth contacts per day across three separate datasets was around 6 at the time of publication in 2006. Perhaps the closest research in

spirit is the work done by Whitbeck et al., (2010), who looked at reconstructing plausible mobility only from Bluetooth address matches, reconstructing which object must have been next to what others at a particular point in time. This research also used a simulation to verify the interactions obtained. While these efforts begin to scratch at the surface of the available data regarding human mobility and movement, they do not yet try to frame ubiquitous devices as a consistent and reliable source of transportation data.

### 2.2.3 Trajectory Inference

In addition to the work regarding mobile sensing listed above, general work regarding the spatial uncertainty of sparse trajectories is beginning to emerge. In particular, Chen et al. (2011), and Wei et al. (2012), have focused on reconstructing trajectories from sparse data from taxis and foursquare check-ins in Beijing. Zheng and Zhou (2011) give a complete overview of the potential trajectory computation applications for data from vehicular GPS, WLAN networks and mobile phones. Many of the concepts explored in the above work can be applied directly to transportation applications and establish a foundation for reducing the spatial uncertainty of opportunistically collected mobile device data.

## 2.3 Privacy and Public Participation

### 2.3.1 Privacy Preservation in Trajectory Data

Besides the numerous technical and theoretical challenges, there are also a number of ethical questions to consider. As computers are slowly "pushed into the background" (Weiser, 1991) to become part of our daily lives, their communications leave traces, breadcrumbs, "digital scents" (Paulos and Goodman, 2004) behind that undermine traditional values of privacy. This is a recurring trend and can be seen in license plate reader data collected over the freeways, credit card transactions and store club membership cards. "Surveillance society" is a term that is quickly gaining momentum as mass data collection becomes easier and easier due to an increasing number of supporting platforms, such as ubiquitous devices. Furthermore, ubiquitous computing is quickly gaining traction as a means of reaching a specific demographic for marketing purposes, as well (Krumm, 2011), which will likely create further value tensions and frustration.

### 2.3.2 Public-driven Data Collection Efforts

Efforts that involve the public in a large scale are beginning to emerge. Most focus on using mobile GPS devices, such as the CycleTracks system developed by the San Francisco County Transportation Authority (Hood et al., 2011). These efforts require the subjects to either carry a GPS-enabled logger or register their device (as is the case with CycleTracks) with a data

collection service, which collects and aggregates individual trajectory information. These efforts have been successful largely due to two factors: (1) clear and transparent communication about the goals and purpose of the data collection and (2) targeting a specific demographic (i.e. cyclists) that has a strong view towards improving the very facilities they use. The success of the proposed study, and future studies that wish to rely on crowdsourced planning data will depend on how well these two requirements are represented. Privacy is not an issue to be taken lightly and the best approach will be to directly convey to the community the type of data being collected and its eventual purpose. While there may be some debate about the use of ubiquitous data, it is unlikely that such a rich data set will continue to be ignored in the future, especially considering the tremendous growth that it may experience. A responsible framework for working with such datasets that protects privacy and maintains anonymity is the most plausible positive outcome.

# Chapter 3 Study Significance

As shown above, there is an increasing number of emerging detection and tracking technologies that are becoming available. As these additional sensing technologies come into play, it becomes imperative to keep track of the opportunities and implications afforded by each. To simplify the task at hand, let us organize the technologies into two groups – ones capable of sensing only at a single point (point sensors) and ones capable of determining something unique about the object they are sensing, thus becoming capable of re-identification, and, by extension tracking. Traditional sensing approaches have been point detectors – loops, pneumatic tubes, even video and radar technologies have focused on counts and speeds at a given point. License plate and toll tag readers were some of the first widely-utilized technologies to focus on re-identification and thus capture corridor and network-wide data. The emergence of MAC sensing makes this information even more accessible and easier to collect. The implications of this development and discussed in the following sections.

## 3.1 Point Sensor Approaches

Detection at a single point is limited to knowledge of events at an isolated point – counts (volumes), speeds and delay (dwell times). These can be collected in a number of ways, using traditional approaches like inductance loops and pneumatic tubes, to computer vision approaches

that emulate presence detection using a "virtual loop" emulator. The end result is a stream of information regarding presence. In a binary sense, when the point in question is occupied, a series of ones is being produced and when it's not, a series of zeroes is produced. Although more advanced approaches can determine the type of object present (and thus provide some means of classification), the overall concept remains the same. This has a number of implications – 1) there is a constant stream of data, it is always known whether there are objects of interest present or not. 2) The information collected cannot be easily extended beyond the point in question.

Although this type of information may appear inherently limited from a transportation network knowledge perspective, the predominant amount of available transportation data has historically been in this form. Along with the limitations of the data however, come some benefits, in particular with respect to individual location privacy. Because the information collected is so limited, there is little privacy risk to any individual, no way of attaching a particular time and place to a given individual object. Due to the dominance of such data within the transportation domain, there has been relatively little discussion of privacy in transportation data prior to the arrival of re-identification capable approaches, discussed below.

## 3.2 Re-Identification Approaches

Re-identification relies on capturing sufficient information regarding the object in question to guarantee its uniqueness (or near-uniqueness) within the population in question.

License plate readers and toll tag readers, the most established types of such sensors use uniquely assigned identification numbers (license plate numbers and toll IDs) to achieve this. Other approaches rely on either a pattern (in some spectrum) or wireless communication identifiers as markers of uniqueness.

Pattern-based approaches include video camera networks that rely on object color distributions, size and other patterns to match one camera's sighting of the object to another's. In addition to exploiting the properties of the inherent object of interest, other surrounding objects within the flow may be used to improve accuracy. For example, if a red car is surrounded by three white cars, in one scene and in another, it is highly likely that that is the same red car. This approach is also used for approximate re-identification used in magnetometer-based sensing approaches, such as Sensys. Since these approaches rely on an inherently inconsistent and potentially not globally-unique cue, their risk for locational privacy is relatively low, compared to communication-based re-identification techniques.

Communication-based re-identification relies on the fact that all devices wishing to communicate (wirelessly or not) require some means of distinguishing themselves from other devices – a unique identifier. MAC addresses fall into this category – their primary purpose is to act as a unique signature for the communicating device. To communicate wirelessly, each device must broadcast this unique identifier, in order to reveal themselves to the interested party. It is listening to this unique identifier that allows for re-identification at any other sensor location.

This approach in inherently more invasive, since the uniqueness of the identifier is guaranteed, furthermore, the relative ease of collection of such data contributes to this concern. The implications of tracking are discussed in the following section.

## 3.3 Implications of Re-identification

Point sensor data provides a limited view of the network as a whole.. Re-identification approaches allow for analysis of entire links, effectively allowing one to study the entire network as a whole. Some of the most prolific and important re-identification based data products include origin-destination pairing data, which is a key component to both long and short-term forecasting efforts. This data has traditionally been collected with surveys, however, the increasing capacity to reliably re-identify individuals automatically using the approaches described above is allowing for this information to be collected without the subject's knowledge or input. This allows for collection of observed preference (instead of stated preference), but relies on implied consent (at best) to collect such data. Because many of the identifiers collected are unique, it becomes relatively easy to tie a particular device to a particular point in space-time. Furthermore, since collecting origin-destination data primarily involves determination of home and work locations, it becomes increasingly easy to tie an individual to a particular device, thus violating their locational privacy.

Besides origins and destinations, imputation of intermediate points is also of interest, in

particular when studying route choice, infrastructure effectiveness and road pricing questions. Imputation of intermediate points allows one to create trajectories, or travel diaries for each observed entity within the network. This information has great potential for use in the new generation activity-based models currently being built and used as transportation and land use forecasting tools. However, the imputation of trajectories yields more issues with respect to compromising individual privacy – in addition to know home and work locations, it is potentially possible to impute place of worship, shopping habits and a host of other individual behavior characteristics. Because many models rely on a variety of indicators to improve predictive power, there is a strong value tension between building accurate models and imputing or otherwise obtaining increasingly privacy invasive data.

As the possibility of MAC-based, network-wide re-identification becomes more apparent, several issues must be addressed. In particular, the inherent uncertainties within the data collection method must be mitigated, all while preserving an acceptable level of privacy risk. The following section describes how the work presented herein addresses these primary issues.

## 3.4 Research Objectives

The objectives of this study include:

- To investigate the feasibility of ubiquitous computing devices as reliable transportation

data sources:

- For vehicle travel times;
- For transit data collection;
- For non-motorized data collection;

- To develop a novel means of collecting non-motorized travel data via mobile device sensors;
- To develop a simulation evaluating the necessary device prevalence and density for representative non-motorized data collection using mobile device sensors;
- To evaluate mobile device-based means of collecting non-motorized travel data;
- To develop a data collection framework for device-based transportation data collection.

## 3.5 Study Scope

This research focuses on the exploration of mobile device MAC address data, collected from Bluetooth devices, and the use of such data transportation intelligence purposes. A number of applications are considered and tested – vehicle travel times, bus stop wait times, pedestrian corridor travel and dwell times, as well as pedestrian route inference. In the process of exploring the space, a number of strategies are proposed to deal with the inherent uncertainty of opportunistically collected data – specifically, dealing population (or sample) uncertainty, temporal uncertainty and spatial uncertainty. These strategies encompass everything from physical sensor placement and data filtering, to more advanced techniques like plausible route reconstruction. In addition to developing these appropriate strategies, this research concurrently

develops the necessary tools to capture and analyze MAC address data – from robust static and scalable app-based sensors for data capture to a visualization and analysis framework capable of making sense of the obtained data. An overview of the research and the organization of this dissertation are given in the next section.

## 3.6 Dissertation Organization

Most of the research contained within this dissertation was conducted in an exploratory manner – much of the field was developing as the research continued and many of the experiments conducted were the first of their kind. However, the main theme of the research has always been the reduction of uncertainty in MAC-based travel analysis applications. This research is then organized with respect to describing the potential applications by increasing uncertainty (single sensor, dual sensor and mobile sensors) and then proceeding to describe the necessary strategies for mitigating this uncertainty. As such, the remainder of this dissertation is organized as follows. Chapter 4 describes in detail the experiments conducted during the research, beginning with simple single sensor tests that focused primarily on population characteristics and continuing to corridor tests conducted in a number of locations. Finally, the mobile sensing approach is introduced and the tests conducted are described. Chapter 5 discusses the specific strategies for uncertainty reduction for the three main groups, first focusing on single sensor data, and thus population or sample uncertainty. The discussion then shifts to temporal uncertainty, manifesting itself as travel time error within the corridor tests. Spatial uncertainty is

discussed using the mobile sensors tests as the primary example – imputation of plausible trajectories is done to reduce spatial uncertainty within the dataset. After the uncertainties have been discussed, privacy concerns, and their relationship with the uncertainties are addressed in Chapter 6. Finally, Chapter 7 concludes with recommendations for future data collection and analysis efforts, as well as directions for future research.

# Chapter 4 Collection and Applications of Mobile Device Identifier Data

## 4.1 Introduction to Device Sensing

As mentioned in the previous section, mobile device sensing is achieved by listening to unique MAC addresses broadcast by wirelessly communicating devices. Although a variety of protocols for wireless communication exist, this research focuses on Bluetooth MAC address broadcasts. Three paradigms of detection are discussed – 1) single sensor detection, or point sensors; 2) dual sensors considered in tandem to provide travel time along a corridor and 3) mobile device-based sensors that are installed as an application on a mobile device platform (i.e. Android). The three paradigms are illustrated in Figure 4-1.

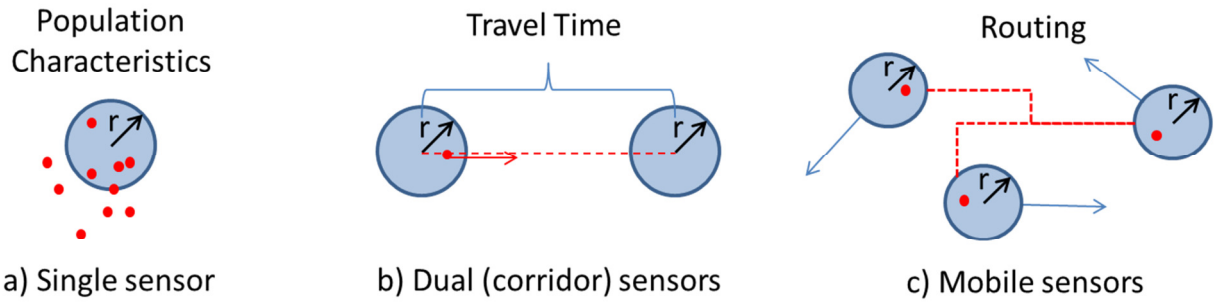



**Figure 4-1: Detection paradigms**

### 4.1.1 Detection Basics

Bluetooth is a short-range communications protocol developed by Special Interests Group (SIG) for inter-device communications. Presently, most electronic devices such as cell phone handsets/headsets, laptop computers, and electronic organizers support the Bluetooth protocol. The protocol itself consists of the device to broadcasting a unique MAC address to devices within range. The broadcast happens at varying frequencies and random intervals (frequency-hopping within a 10.24 second time window), allowing for multiple devices to connect to each other. This protocol was designed for short-range, multi device communications, and is thus optimized for such purposes, creating some challenges for its use for additional purposes, such as travel time collection based on Bluetooth MAC address matching.

Detecting Bluetooth devices is subject to several sources of error, which undermine the overall travel-time measurement accuracy. First, the frequency hopping protocol allows a random delay of up to 10.24 seconds in device discovery time, which may result in a location error of approximately 170 meters (558 ft) at 30 km/h to 570 meters (1870 ft) at 100 km/h (62 mph) at each detection point for highway travel-time data collection. These errors can impact the travel-time data accuracy very significantly if the link is short because the location errors are relatively high for the link distance. A second error factor comes from the variety of Bluetooth devices, antenna types, and geometric configurations that are possible. Depending on configuration of the sensors, Bluetooth devices may be detected sooner or later – for example, two detectors working in tandem in the same location can cut the maximum detection delay in

half, to just over five seconds. Finally, additional errors result from other miscellaneous devices within the analyzed corridor - these could be modes that are not of interest or static devices, but still are recorded due to the nature of the data collection process.

### 4.1.2 System Design

Throughout the project, the in-house designed Bluetooth sensor, hereafter referred to as the Media Access Control Address Detection system, or MACAD for short, has gone through three version changes and a number of upgrades. Figure 4-2 outlines the evolution of the device throughout the year-long project. The first version of the device was designed based on a Gumstix platform. The Gumstix platform provides a full Linux-based operating system running on a 600 MHz processor, all on a footprint about the size of a stick of gum (Gumstix, 2010). The device was powered by eight “D” cell batteries which allowed it to function continuously for 40 hrs. At the time an 8 dBi “rubber duck” external antenna and a 12 dBi in-lid antenna was used with a DCE-ANT NEMA 6 rated enclosure. Although this setup provided ample processing power and functioned well, there were concerns about the relatively short running time as well as the use of “D” cell batteries in wet environments, which was not recommended by WSDOT field engineers.

To reduce power consumption, a 60 MHz processor was chosen for the second version of the device (V2.0). This greatly increased run time, allowing the device to operate for 5 days on just six “D” cell batteries. However, concerns about oxidation of the batteries, as well as the

general wastefulness of single-use batteries prompted a rechargeable battery-based system. Version 2.1 of the system included a Lithium-Iron (LiFE) rechargeable battery and an N-Male interface that allowed for a variety of waterproof, external omni-directional Laird antennae to be mounted on the device.

After completing V2.1, questions arose about data communication – previous versions have been saving the data onto MicroSD cards which had to be extracted prior to data analysis. Although this was convenient for short tests, additional information during longer tests was seen as an advantage. Eventual practical deployment of the device also would require a means to transfer data in real time, allowing for use in conjunction with user information systems. A GSM/GPS module was added to the device to allow for communications and resolve clock synchronicity issues. Finally, a custom board was designed to hold all of the components and yet another battery was chosen. The reasoning behind switching from LiFE to Lithium Polymer (LiPo) batteries was mainly practical – LiPo batteries could be charged much faster, on the order of hours, compared to days when using LiFE batteries. With the design finalized, four units were produced for field testing; these were used in the majority of the experiments discussed in the following sections. However, development of MACAD continued, to incorporate new features – Bluetooth 4.0 compatibility and WiFi MAC collection capability, in a addition to two lower power, 16Mhz ARM processors (one per board). A dual component design was followed to allow for production of two different unit types – ones with full communication capabilities and ability to relay data back in real time, and a “light” version designed for temporary studies that only

contained Bluetooth and WiFi functionalities (along with basic processing and storage capabilities) to reduce unit cost. Each board is autonomous – one processor takes care of communications and another takes care of the sensing tasks. The exact end product is described in greater detail in the following section.

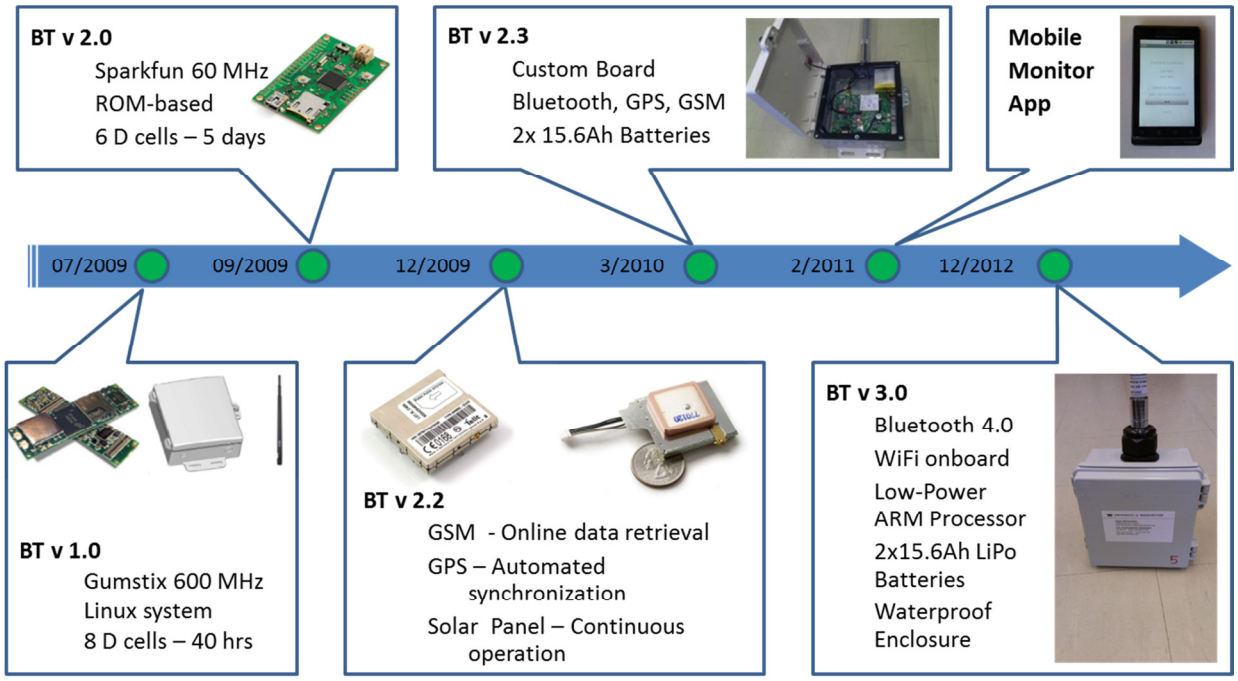



**Figure 4-2: MACAD evolution**

### 4.1.3 Current Design Overview

As alluded to in the previous section, the current device design consists of five main components; (1) a Bluetooth chipset that constantly scans the available 79 channels, (2) a WiFi chipset that scans the WiFi spectrum, (3) a 16 MHz ARM processor that records MACs (4) another 16 MHz ARM processor that takes care of communications and (5) a communications module that synchronizes to the UTC time and transmits data in near real-time (GPS + GSM). The device is enclosed in a weatherproof enclosure which provides a port for an external antenna, as shown in Figure 4-3. This provides an excellent base for testing mounting locations and various antennae as it can be mounted to signposts and signal posts and will accept a wide range of antenna types, as also demonstrated in Figure 4-3. The current design allows the device to function for up to a week without external power using one 6-cell LiPo pack (15.6Ah capacity @ 3.7V), running the sensing board only and storing on a local SD card. The device allows for up to two battery packs at a time, resulting in a maximum runtime of two weeks without external or solar power.

**Figure 4-3: STAR Lab Bluetooth detectors (MACAD device) used in this study**

Solar power compatibility was also considered in the design and a solar power module has been designed and tested. The device operates using the power provided by the battery which is, in turn, charged by the solar panel. Preliminary testing indicates that the discharge rate is lower than the received solar power input rate, indicating that continuous operation is possible. However, solar power was not used during any of the data collection experiments contained in this dissertation.

### 4.1.4 Communications Design

Once mounted, the device synchronizes to UTC time using the communications module. In addition to synchronizing over the GPS network, the system also sends its exact coordinates via GSM. These coordinates are then used to automatically locate the deployed sensor units on the network. This initialization routine is repeated at regular intervals to prevent clock drift (Quayle et al., 2010) and ensure that the device is functioning properly and has not been tampered with. Once the synchronization and location recording is complete, the device begins data collection, recording the bypassing MAC addresses and their respective timestamps. As data is collected, it is sent over the GSM network to a server in University of Washington (UW) Smart Transportation Applications Research (STAR) Lab, where the MACs are kept for a specified time period (currently 60 minutes). If a matching MAC is received during this time period, it is processed, deleted and the resulting information is uploaded to the Digital Roadway Interactive Visualization and Evaluation Network, (DRIVE Net) (accessible at www.uwdrive.net) developed by the STAR Lab at the University of Washington for data sharing, modeling, and online analysis (Ma et al., 2011). This approach to data collection allows for real-time information flow to the users while maintaining a level of privacy. Figure 4-4 illustrates the overarching structure of the data collection effort.

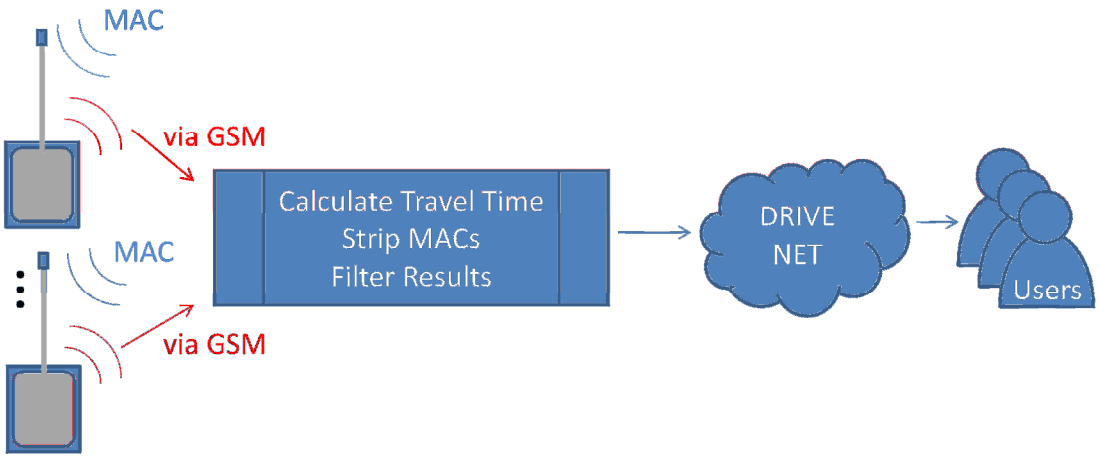



**Figure 4-4: Bluetooth data collection and distribution diagram**

## 4.2 Single-node Data Collection Paradigm Applications

### 4.2.1 Data collection approach

A single-point data collection effort involves collecting data at a specific location without tying the data to that of other sensors. This approach is the most prolific in transportation data gathering. There are a number of applications under this paradigm that are possible using MAC address sensing. Any applications where sensing the presence time (dwell time) of a given entity is of interest can be accomplished using this paradigm. Two specific applications are discussed and tested in the following sections: (1) Bus stop passenger wait time estimation and (2) dwell time estimation for pedestrians in tourist areas and special events.

### 4.2.2 Passenger Wait Time Estimation Application

#### *4.2.2.1 Introduction*

For this study, short range Bluetooth detectors (shown in Figure 4-5) were installed at high-volume transit stops for extended periods of time to detect enabled Bluetooth devices in the short range of the detector. The detectors consisted of a Motorola Droid phone running the Android operating system attached to a 15Ah lithium battery, which allowed the devices to function for over 2 weeks. Scans for Bluetooth devices were performed continuously using the built-in antenna, a Class 2 Bluetooth device with a range of approximately 10m.

As transit users approached the stops to wait for the bus, the detectors noted the arrival time of the MAC address broadcasted by the user's Bluetooth device. The user was assumed to be present and waiting at the bus stop as long as continuous check-ins were made by their device (for example, every 60 seconds). The user was assumed to have left the stop once the devices failed to check in for the second consecutive time.

**Figure 4-5: Short-range Bluetooth detector.**

#### *4.2.2.2 Study Sites*

Locations with minimal interference from bypassing vehicles were selected to minimize data noise. Park and ride lots present nearly ideal locations due to their singular purpose. Furthermore, there is less ambiguity regarding collection of waiting times at park and rides, as there are no passengers on the bus, thus their devices will not interfere with the data collection process. Two park and ride lots were chosen in this study, one in Redmond, WA (Redmond P&R) and one in Seattle, WA (Northgate P&R). These locations differ by their population

characteristics but are both served by King County Metro (KCM), the transit agency for greater Seattle. Sensors were mounted directly on the bus shelters, along the roof beam to prevent vandalism and theft. No cases of either were encountered during the experiment. The sensor locations considered in this study are shown in Figure 4-6 below.


a) Redmond Park and Ride Instrumented Bays


b) Northgate Park and Ride Instrumented Bays

**Figure 4-6: a) Redmond Park and Ride test site b) Northgate Park and Ride test site**

#### 4.2.2.3 Data Acquisition and Filtering

Although the chosen locations had few outside influences and were predominantly populated by transit users, data filtering was still necessary to estimate approximate wait times. One of the biggest issues with the obtained data was the apparent abundance of very short wait times, usually under one minute long. These times are usually not a result of just-in-time arrivals, but are outliers that can occur in a few different situations. The most common is the occurrence of Bluetooth devices from passengers already on a bus that is stopping at a stop – this creates the impression that one or many devices have appeared and left with the bus. With higher Bluetooth prevalence rates, it may be possible to detect such events, comparing large arrival events with the transit schedule to detect likely transit riders. However, due to the low sample sizes encountered, a single bus may have only one or two discoverable devices, making such discrimination more difficult. Another common source is bypassing vehicles and pedestrians, although this was minimized in the chosen locations. If these cases are considered to be waiting times as well, there will be a strong bias towards shorter wait times. Due to the "blind" nature of the data collection technology, there is no means of knowing which device was a true just-in-time arrival and which was a bypassing passenger, pedestrian or driver.

As mentioned before, the arrival and departure times of a transit user were estimated using a continuous check-in (detection) procedure. In order for a device to be considered a waiting transit user, the MAC address had to be seen at continuous intervals no more than 60 seconds apart. Furthermore, a valid record had to contain at least two check-ins. Although there

is a case for bias against shorter wait times, this approach greatly reduces the number of irrelevant and erroneous records. The locations chosen did not have many vehicular delays, thus most bypassing MACs are ignored.

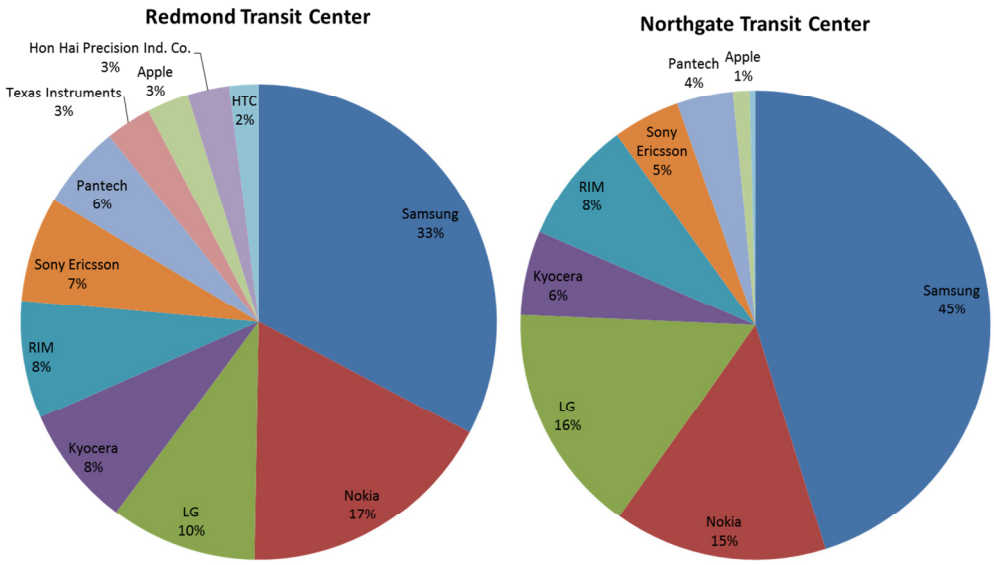
In Redmond, some of the bays were sufficiently close to one another that some MACs may have been detected at two or more stops. In this case, the MAC address was assigned to the stop with the most check-ins. Wait times of over 100 minutes were discarded and considered to be static devices, as there are no buses with headways greater than 90 minutes.

#### *4.2.2.4 Site Characteristics*

As mentioned above, the locations considered differ in population characteristics – Redmond P&R provides transit access to many high-tech firms and it was thus hypothesized that the location may not only have more devices in general, but also that there was a higher chance of the encountered device being a smartphone or other internet-capable device. Northgate P&R is situated in a less technically-oriented neighborhood and provides access to the University of Washington and downtown Seattle.

Brand analysis was performed to examine the characteristics of the devices detected at these locations. Some information regarding the type of the devices present can be gleamed from the manufacturer – for example if there are many devices with in-vehicle system manufacturer codes (such as Garmin, Parrot, etc.) then vehicles are likely to be a large part of the captured

data, which may render it unsuitable for transit delay estimation, without proper filtration techniques. Device manufacturers of the detected MAC addresses can be obtained by looking at the first three hex digits, which specify the manufacturer of the Bluetooth chipset in the device. Figure 4-7 shows the relative brand presence of each site. It can be seen that Samsung, Nokia and LG dominate the landscape. There are a number of reasons for this – one is that all of these manufacturers produce headsets, which are popular accessories and often do not have an option to disable the "discoverable" function, thus continuously broadcasting their MAC address while on. Another reason is that the Samsung, Nokia and LG handsets have the option of leaving the "discoverable" mode on indefinitely, unlike phones running the Android or iPhone OS, which by default allow for only 120 second intervals of device discoverability.




a) Device Brand Populations at Redmond TC

b) Device Brand Populations at Northgate TC

**Figure 4-7: Device Manufacturer Populations at Each Transit Center**

Overall 341 devices were determined to be waiting at bays 2,4,5 and 6 of the Redmond Park and Ride from the evening (6pm) of 6/12/2011 to the evening (6pm) of 6/17/2011 (120 hours). At Northgate, 259 devices were found from the evening (6pm) of 6/12/2011 to the evening (6pm) of 6/15/2011(72 hours) just in Bay 1 of the Park and Ride. Thus, the Northgate site had a much higher device rate (3.6/hour) than the Redmond site (2.84/hour), even though only one bay was examined. KCM reports that the Northgate Park and Ride experiences roughly 417 boardings per hour vs. 900 at Bays 2,4,5 and 6 in Redmond. This equates to a fairly low sample rate of 0.86 and 0.29 percent at Northgate and Redmond, respectively. These rates are lower than the 2-3% sample rates found among pedestrians at the University of Washington and the 5% sample rate found among pedestrians in Montreal, QC (*13*). One of the reasons for this may be the shorter range of the employed devices. The rate difference between Northgate and Redmond is also notable, as it is somewhat counterintuitive. However, there is evidence that the Redmond location has a larger share of smartphones in general– both locations have an 8% Blackberry prevalence, but HTC and Apple hold an additional 5% in Redmond vs. just 1% for Apple devices in Northgate.

#### *4.2.2.5 Experimental Results*

Examining the data in greater detail, some general insights can be gained about transit wait times. Looking at the transit wait times by the various stops, boxplot shown in Figure 4, can

give us an insight into how particular operating strategies (headway durations) can affect the average wait time and thus the general satisfaction of the users. The average wait time at Northgate was 4.3 minutes, while the average wait times at Redmond for bays 2,4,5 and 6 were 11.5, 8.9, 11.5 and 4.9 minutes, respectively. Ground truth data was collected throughout the three days, in the morning hours (8-9 am for Redmond and 10-11 am for Northgate) and is shown in comparison to the average wait times in Table 4-1 below. Ground truth data was collected manually by an observer noting the arrival and departure times of every individual per bus bay. Averages were computed across all days for the morning peak (6am-10am) (to have sufficient data) and compared to ground truth data collected on site.

**Table 4-1: Comparison of Bluetooth Estimated Wait Time to Measured Wait Times**

| | AM Headways | Average Wait (mins) | Ground Truth (mins) | Error % |
|---|---|---|---|---|
| Northgate Bay 1 | 30 mins | 3.78 | 6.6 | 43% |
| Redmond Bay 2 | 30 mins | 3.22 | 0.0* | N/A |
| Redmond Bay 4 | 30 mins | 8.17 | 5.2 | -58% |
| Redmond Bay 5 | 30 mins | 12.77 | 11.6 | -10% |
| Redmond Bay 6 | 10-30 mins | 4.89 | 3.9 | -26% |

*single data point

Some locations showed higher error rates than others, in particular, the Northgate and Redmond Bay 4 locations. Looking at Figure 4-8, the sample size (shown on right axis) is fairly high for these bays, indicating that lack of data may not be the problem. The issue may be noise

from adjacent bays, with people standing in between the bays while waiting for their bus (as observed). Comparing the average morning headways (shown in Table 4-1) to the wait time vectors encountered at each bay, it can be seen that the measured times are plausible, but include times over 30 minutes, which may occur during off-peak hours. Hourly data shown in Figure 6 confirms that longer wait times occur during night and mid-day service lulls.

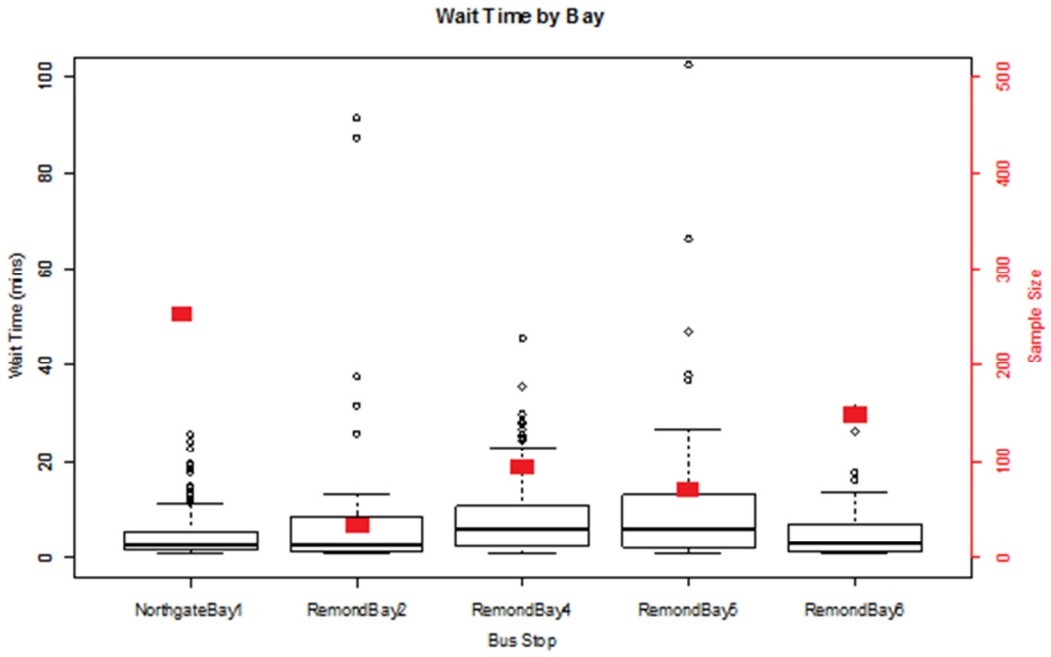



**Figure 4-8: Wait Times by Transit Stop**

Additional comparison of wait times by device manufacturer provides some interesting insights into potential time savings using mobile transit information services such as OneBusAway, a service that provides real-time bus arrival information in the greater Seattle area. OneBusAway is available online as a website, but also as an app for the Apple and Android

platforms. This suggests that internet-capable (smartphone) device users may have greater access to real-time bus information and can reduce their wait times. Table 4-2 shows the respective delays experienced by each manufacturer. Apple devices experience the lowest average wait times – 3.3 minutes on average, compared to 6.5 minutes experienced by all other devices. This result comes with one caveat. Due to the limited discoverability inherent to Apple iPhones – only 120 seconds, phones with longer waiting times may not be discovered. HTC and Samsung manufacture phones for the Android platform, which also has a 120 second discoverability cutoff. However, both manufacturers build other devices as well, thus their results are not necessarily indicative of smartphones in general. This is true for Nokia as well, which makes phones capable of continuous discoverability, but not necessarily well suited for internet browsing. Of the data collected, only Research In Motion (RIM) manufacturers exclusively smartphones (and accessories which can be assumed to be bundled – a person with a Blackberry headset is likely to have a Blackberry phone). RIM brand devices experience a delay of 5.5 minutes on average, one minute (18%) less than the average of all other manufacturers.

**Table 4-2: Average Wait**

| Brand | Average Wait |
|---|---|
| Apple | 3.3 |
| RIM | 5.5 |
| Nokia | 5.9 |
| Samsung | 5.9 |
| Sony Ericsson | 6.1 |
| LG | 6.3 |
| Kyocera | 7.1 |
| HTC | 8.3 |
| Pantech | 9.1 |
| Hon Hai | 17.4 |
| Overall | 7.49 |

Taking an even broader look, Figure 4-9 shows the wait times at all locations by hour of day. The two commuter peaks are also clearly visible in the sample sizes encountered by the hour, showing that the aggregated data is indicative of the overall trends. This shows the

variability of passenger wait times based on operation mode – peak times (6-10am and 4-7pm) have lower wait times than off peak wait times, particularly in the evenings. This is expected, as the headways during peak times are notably shorter.

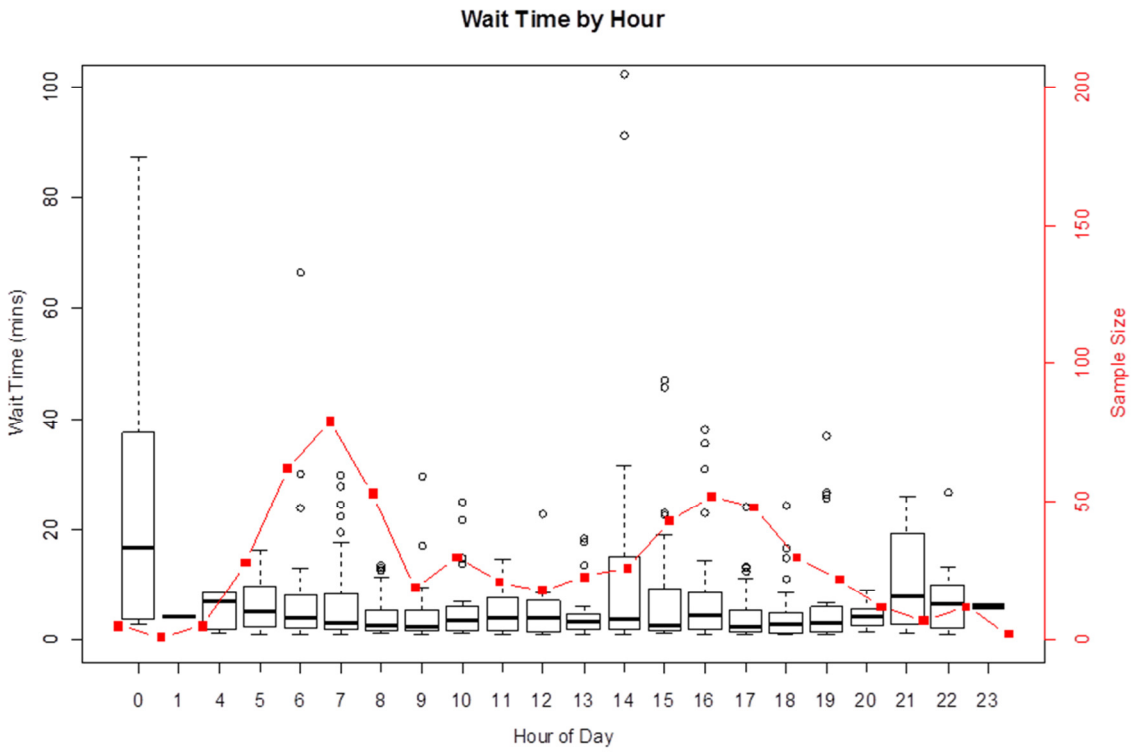



**Figure 4-9: Wait Times by Hour of Day**

#### *4.2.2.6 Summary and Conclusions*

One of the main variables in transit level of service is passenger wait time, or the amount of time passengers spend at a particular stop. Lower wait times imply better transit service, as they reduce total travel delay. Wait times are generally a function of the headways at which the transit service operates. Shorter headways mean shorter wait times, but are expensive to

maintain, particularly if rider volumes do not justify such frequent service. Thus, many agencies are interested in monitoring the overall wait times at transit stops and gauge these against their operational strategies. However, monitoring wait times is usually a manual endeavor, which is costly and time consuming. Surveys asking passengers for wait time information may not be accurate, as perceived wait times often differ from actual ones. Automated means of collecting actual wait times would allow agencies to better gauge and evaluate their operational strategies by greatly reducing the cost of data collection.

A proof of concept study investigating the use of Bluetooth detectors for transit stop wait time data collection is presented in this paper. Two separate transit centers in the Seattle area were instrumented with Bluetooth detectors at individual stops. Wait times were collected by examining the amount of time a detected Bluetooth device spent in range of the sensors. The overall sample rates were 0.86 and 0.29 percent at the two sites.

The use of Bluetooth sensors for transit data collection opens some additional opportunities for transit operations analysis and planning. Although the sample size collected was small, the data collected still demonstrated big-picture trends and could be compared to hourly ground truth averages. However, finer resolution data is still difficult to interpret due to the "blindness" of the technology – the inability to determine the mode of individual records. The collected wait times also contain some potential bias – one source being bypassing vehicles and pedestrians and the other containing devices with limited discoverability settings.

Furthermore, the wait times represent only individuals with Bluetooth devices, which may not be fully representative of the entire population. As more and more "smart" devices enter the market, the sample size is likely to grow, which may allow for better analysis options, such as individual bus route analysis and finer time resolutions that can mitigate some of the issues described above.

### 4.2.3 Pedestrian Dwell Behavior Application

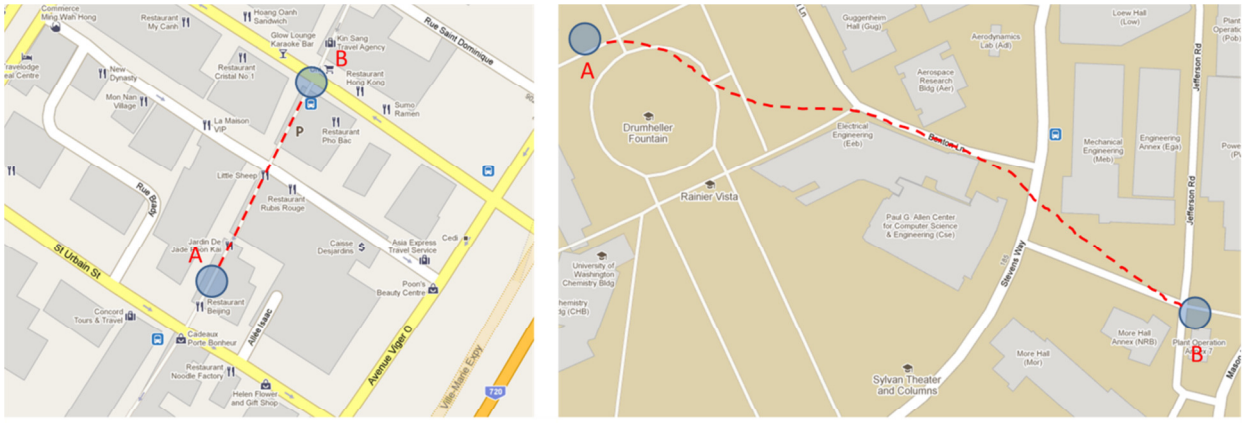
#### *4.2.3.1 Introduction*

In addition to measuring bus stop waiting time, the single point paradigm can be used to determine residence (dwell) times for pedestrians as they traverse through an area. This problem is interesting from a tourism and event organization perspective, as it allows for determination of where individuals may spend extra time in a pedestrian environment. Two locations were examined in this application - a short, pedestrian-only corridor in downtown Montreal, QC (Montreal) and a section of the University of Washington (UW) campus in Seattle (Seattle), also restricted to pedestrian access. Both studies were conducted in the summer – testing in Montreal was performed August 18-19$^{th}$ 2010, and testing in Seattle was performed during the UW Graduation Ceremony weekend on June 11-12$^{th}$, 2011. Data from both locations was truncated to 24 hours starting from 9:30am. Two sensors were mounted at each location, allowing for a

corridor sensing paradigm as well, however, the point sensing approach will be discussed first. The travel time results obtained in this study follow in Section 4.3.

#### *4.2.3.2 Study Sites*

The Montreal site consisted of a corridor that was about 100 m long, passing from point "A" in the middle of a block (where no vehicles were present) to point "B", the end of the next block at an intersection with an arterial. The UW site consisted of a path about 350m long, leading from "A," the Drumheller fountain (a popular picture spot) down to "B," Husky Stadium, where the Graduation Ceremony was taking place. Figure 4-10 shows the maps of both locations, along with sensor placements (shown in blue) and shortest path outlines (shown in red).




a) Chinatown Test Site, Montreal. Walking Distance = 100m (328ft)

b) University of Washington Test Site, Seattle. Walking Distance = 350m (1148ft)

**Figure 4-10: a) Downtown Montreal site b) Seattle UW Campus site**

The selected sites represent two common pedestrian areas, dense urban centers

(downtowns) and campuses (malls, parks, etc.). There are a number of travel characteristics that are of interest at these types of locations. Count data (volumes) was the most basic data collected, as examining Bluetooth MAC address readings as a population estimate is of interest. Furthermore, dwell time, was collected to determine the amount of time an individual spends within a particular area. However, prior to discussing the obtained results regarding these three metrics, it is important to establish the characteristics of each site with respect to device presence rates and types.

One of the reasons that research regarding pedestrian detection using Bluetooth has been limited is due to the relatively low sample sizes attainable. While vehicles may have a number of Bluetooth-equipped accessories, and are subject to hands-free laws which require drivers to wear headsets (or be on "speaker mode"), pedestrians have a limited number of devices that they may be carrying in the current environment. Primarily, we can expect pedestrians to carry cell phones, headsets and computing devices such as tablets or laptops. Additional devices such as heart-rate monitors and pedometers are beginning to enter the market, but do not yet have a notable share of the total device market. The shares of each device manufacturer per site are shown in Figure 4-11. Overall, 2520 unique devices were seen in Montreal, while only 534 were seen in Seattle over a 24-hour period. Both sites had similar brand populations, comprising primarily of headset/handset manufacturers such as Nokia, Samsung and LG. These devices are especially prevalent due to the support of continuous discoverability (broadcast) in handsets. In contrast, Apple, Motorola and HTC handsets use a limited discoverability protocol, where the device

remains visible for only up to 120 seconds. Other devices, however, such as tablets, can support continuous broadcast modes.

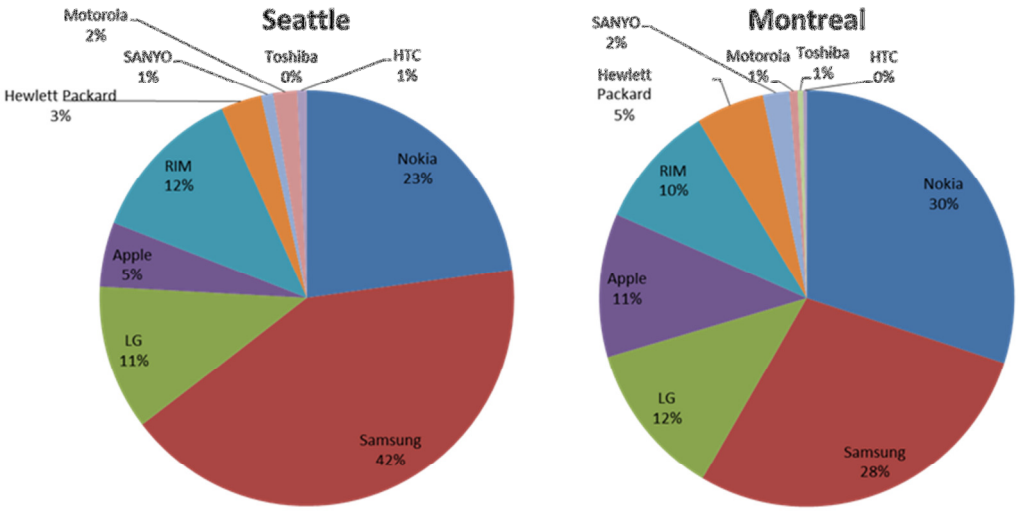


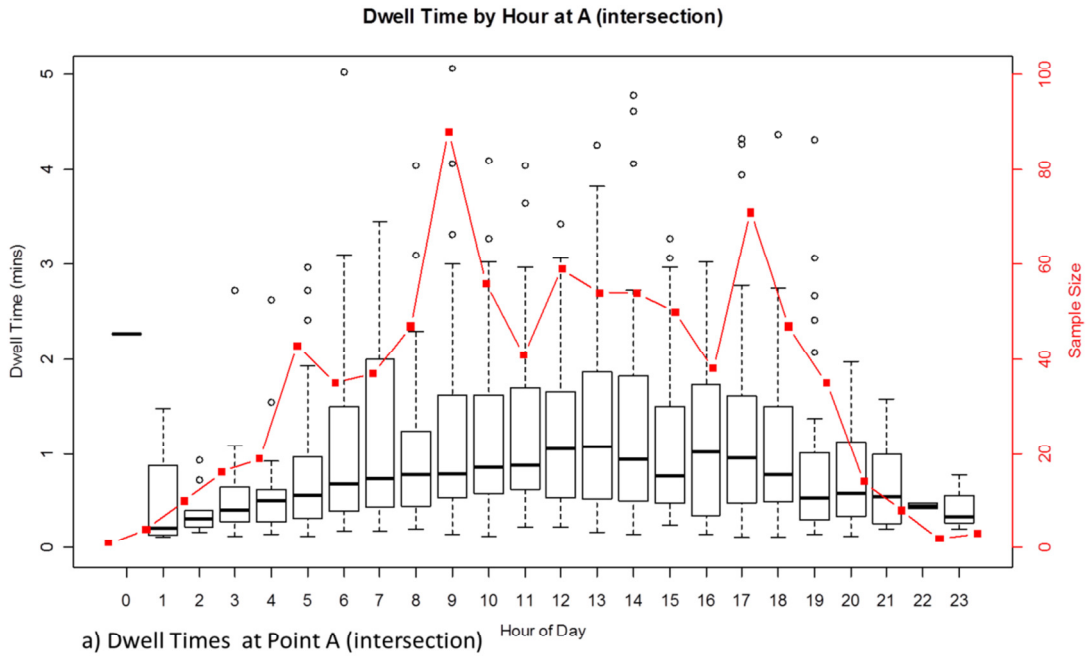

**Figure 4-11: Device brand distributions at both study sites**

Short, one-hour manual counts were done within range of the installed sensors to determine the approximate sample rates at each location. In downtown Montreal, which had extremely high pedestrian flows of around 2687 pedestrians/h, the sensors were able to capture roughly 5 % of the population. In the UW Seattle location 2.25 % of the population was represented by a MAC address (Malinovskiy, 2012). These figures are much lower than the vehicle sample rates of roughly 10 % that were obtained at the data collection points in other studies (Malinovskiy, 2010 and 2011).

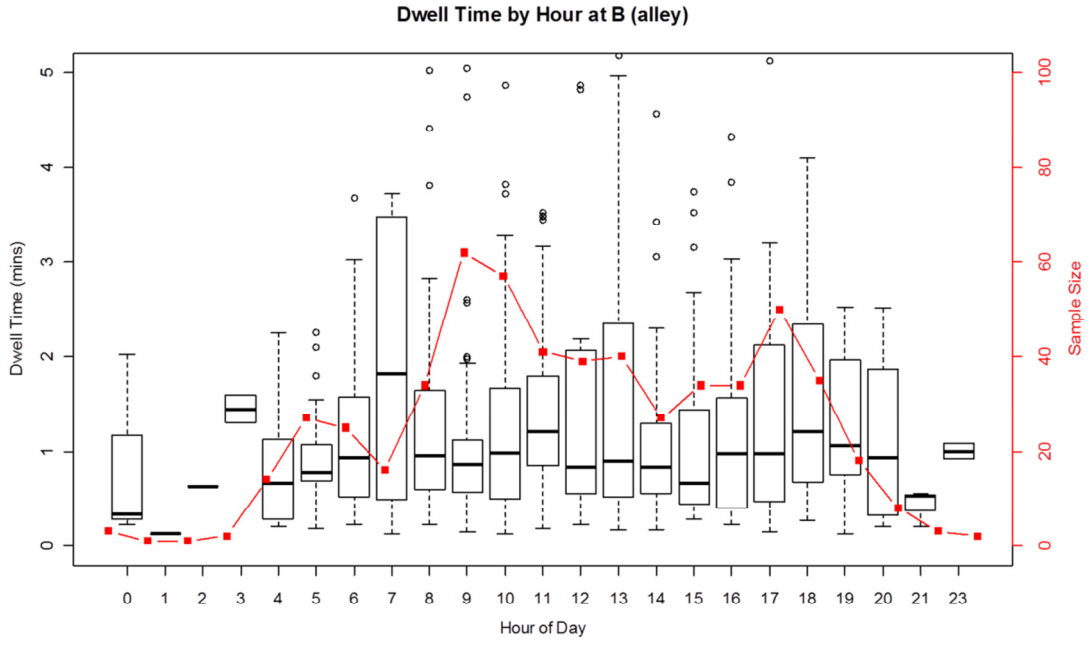
#### 4.2.3.3 Experiment Results

Dwell times at both locations were examined to see how these values may be of value in describing a particular location. Dwell times were calculated on the basis of continuous presence, as for the transit user study – a device had to check in every 60 seconds to be considered continuously present but needed just one interval to be a valid record. This is a reasonable arbitrary threshold that attempts to filter out briskly passing individuals (over 5ft/s) so they would not be considered "dwelling" and pass the detection range (100 meters) in under the 60 second limit. The longest interval time with a continuous presence is considered to be the dwell time. In Montreal, the two locations chosen for sensor mounting were only 100 m apart but had different dwell time characteristics; with an average dwell time of 1.28 minutes (76.8 seconds) at location "A", near the intersection and 1.70 minutes (102 seconds) at location "B", in the middle of the alley. As expected, the sensor mounted at the intersection discovered more unique Bluetooth MAC addresses - 832 versus the 573 unique MACs discovered at the mid-alley location where only pedestrians were present. Figure 4-12 shows the recorded dwell times at each location at the Montreal site. The data is presented by hour of day, displaying the variability encountered during each hour. The 9 o'clock hour is wrapped around two days, since testing began at 9:30 am. Hours with no data are not shown. Looking at Figure 4-12, most pedestrians spend up to about 3 minutes passing through the detectable zone of 100 m. Median dwell times seem consistently larger at mid-alley than at the intersection, and have different temporal evolution, e.g., globally decreasing after 16:00 at the intersection while increasing from 15:00 to

18:00 and then decreasing at mid-alley, which may reflect more restaurant activity (based on observation of available restaurants) near this location. It's also apparent that the pace quickens at night when the numerous shops and stands are closed. The sample sizes collected during each hourly interval are also shown on the right axis. The two commuter peaks are clearly seen at both locations, but are more evident at the intersection, possibly due to increased vehicle traffic. The morning peak appears in two stages as well, with a portion of devices appearing at 5am and the dominant majority at 9am.




a) Dwell Times at Point A (intersection)




b) Dwell Times at Point B (alley)

**Figure 4-12: Dwell times at the Montreal Chinatown location**

To further examine the difference between the two ends of the study segment in Montreal, dwell time distributions (limited to 15 minutes) are shown in Figure 4-13 below. The two distributions are found to be significantly different, with a p-value of .001 using the Kolmogorov-Smirnov test. This can be attributed to a higher number of very low dwell times that are most likely detected in vehicles passing by. In Figure 4-13, it can be seen that the 1 minute bin at the intersection (A) has roughly 60% more MACs than the same bin in the alley (B). This is 15% higher than the average increase of MACs seen at the intersection, implying that this increase is not simply a result of higher device populations seen at the intersection.

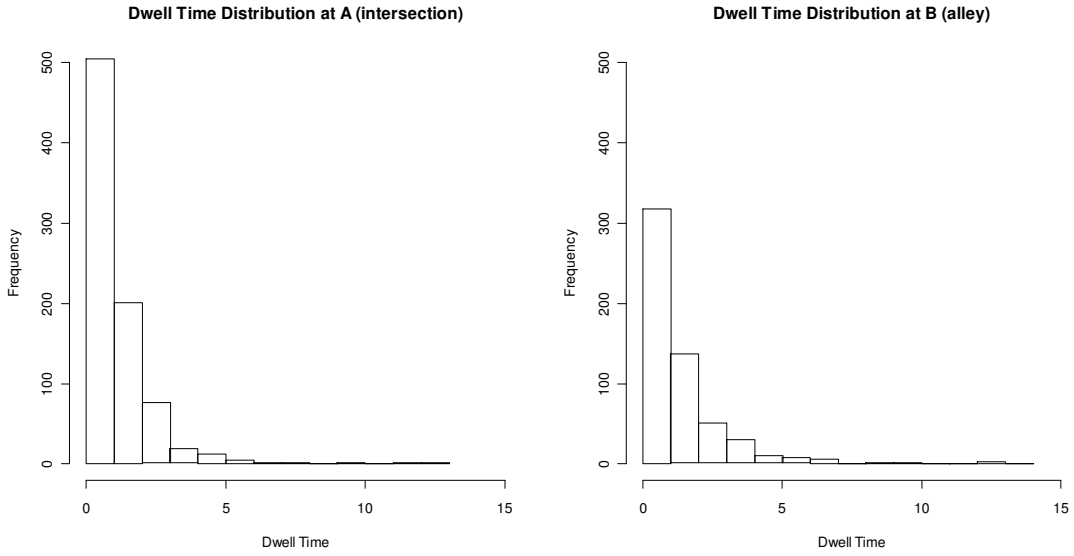
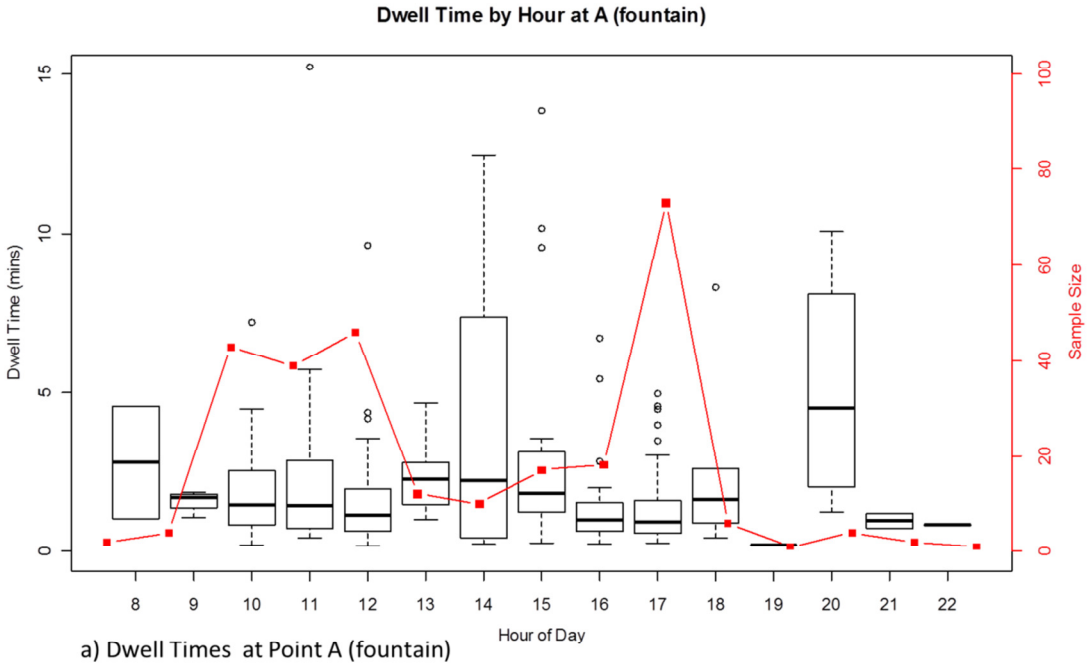


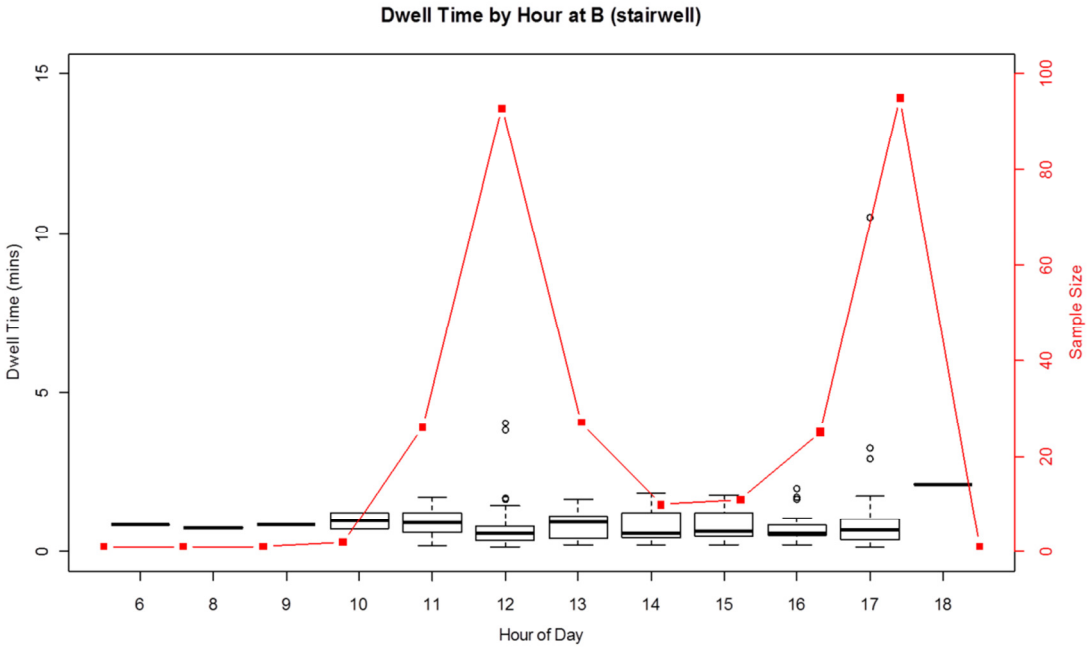

**Figure 4-13: Dwell time distributions at the Montreal sensor locations**

The Seattle location exhibits different trends due to the event-based nature of the pedestrian flows. Furthermore, the locations differ in characteristics – location “A” is next to the

Drumheller fountain, where many stop to take pictures as well as meet others, while location “B” is on an uneventful stairway leading to the UW Husky stadium. Thus, the dwell times at location “A” are notably higher – 2.14 vs. 0.81 minutes at location “B”. More devices (422) were also detected at the fountain than at the stairwell (353). The dwell times are also obtained primarily during the beginning and end of the Graduation Ceremony, with very few detections at other times. Figure 4-14 shows the boxplots by hour of the dwell times seen at both sensor locations. Two peaks of detected device population samples can also be seen, as more people collect to make the Ceremony and later depart. The first population peak at the stairwell is sharper, taller and occurs about an hour later than the first peak at the fountain, suggesting that people spent about an hour on average strolling around campus prior to going down the stairwell to the stadium.




a) Dwell Times at Point A (fountain)




b) Dwell Times at Point B (stairwell)

**Figure 4-14: Dwell times at the University of Washington, Seattle location**

Figure 4-15 shows the dwell time distributions (limited to 15 minutes) at both ends of the study segment. The distributions vary significantly from one another (Kolmogorov-Smirnov test p-value less than 0.001) – as expected, the dwell times at the fountain are generally longer; as it itself is an attraction, unlike the stairwell. There are almost no dwell times over 2 minutes at the stairwell.

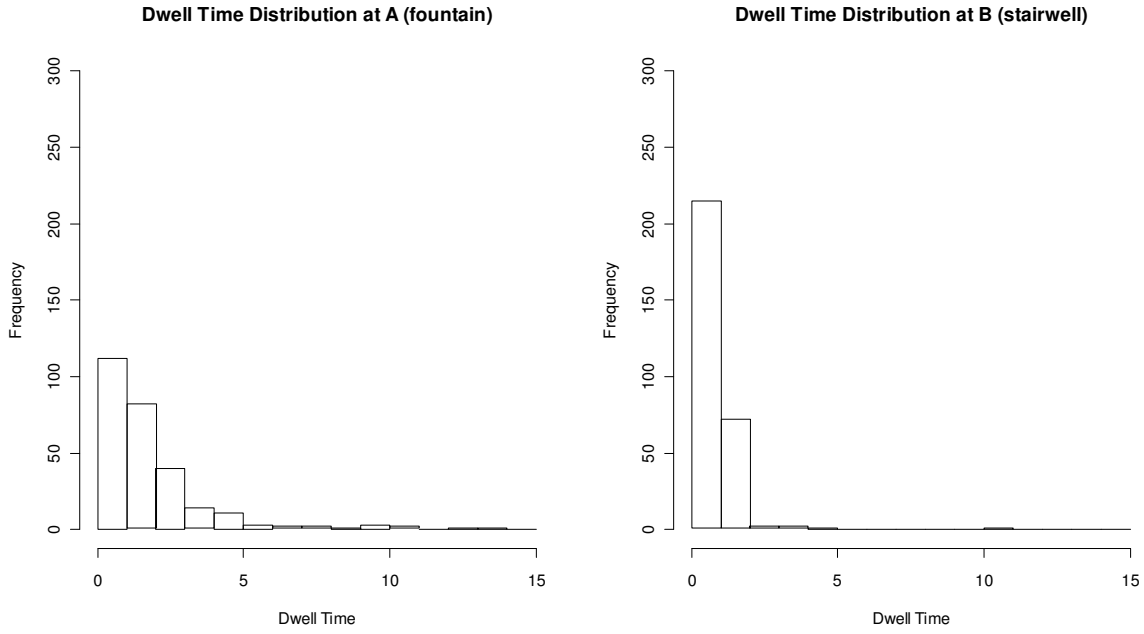



**Figure 4-15: Dwell time distributions at the Seattle sensor locations**

## 4.2.4 Single Sensor Paradigm Issues

Bias is of concern with the proposed methodology. It may be expected that more affluent, younger populations would carry more Bluetooth-enables devices, especially highly visible devices such as headsets. However, there are additional considerations, such as device type

(some less expensive cellular phones are capable of continuous broadcast only) and technological prowess (forgot/unable to turn Bluetooth visibility off) that also play a role. Preliminary testing on bus stops in various Seattle neighborhoods suggest that there may actually be more visible devices present in less affluent neighborhoods, but additional work is necessary to determine the extent and direction of bias.

Overall, Bluetooth appears to be capable of providing automatically and continuously basic pedestrian movement trend information which can be used for planning purposes. With more personal device data becoming available correctly interpreting and filtering the available data is becoming the most pressing issue in travel analysis. The locations chosen in this study relied on selective sensor location to filter pedestrians from other devices present in the area.

## 4.3 Corridor Data Collection Paradigm Applications

### 4.3.1 Data collection approach

A dual-node data collection effort involves collecting data at two (or more) sensor locations and co-relating the data between them – re-identifying the individual entities seen before. The key to re-identification is to match the MAC addresses seen at the respective sensors. Although the main concept is straightforward, the realities of the protocol make the matching step a bit more involved – for example, each device is likely to be detected multiple times at each location and choosing which record to consider representative can affect results. Two specific applications are discussed and tested in the following sections: (1) a pedestrian travel behavior application and (2) a highway segment application.

### 4.3.2 Pedestrian Travel Behavior Application

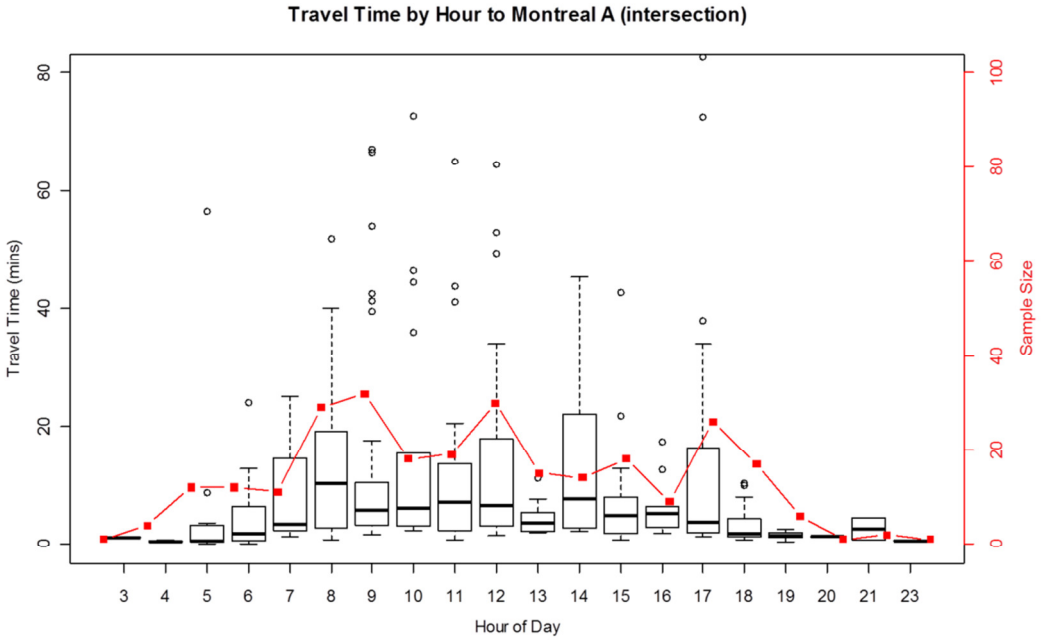
#### *4.3.2.1 Introduction*

Extending the experimental setup (shown in Figure 4-10) discussed in the previous section to a corridor evaluation by cross-correlating the MAC addresses seen at either end of the corridor provides an insight into the travel behavior of the pedestrians traversing the corridor segment. Dwell time can only capture the behavior of individuals in the sensor's proximity, while analyzing the sensors as a corridor can also give an indication of what time was spent in

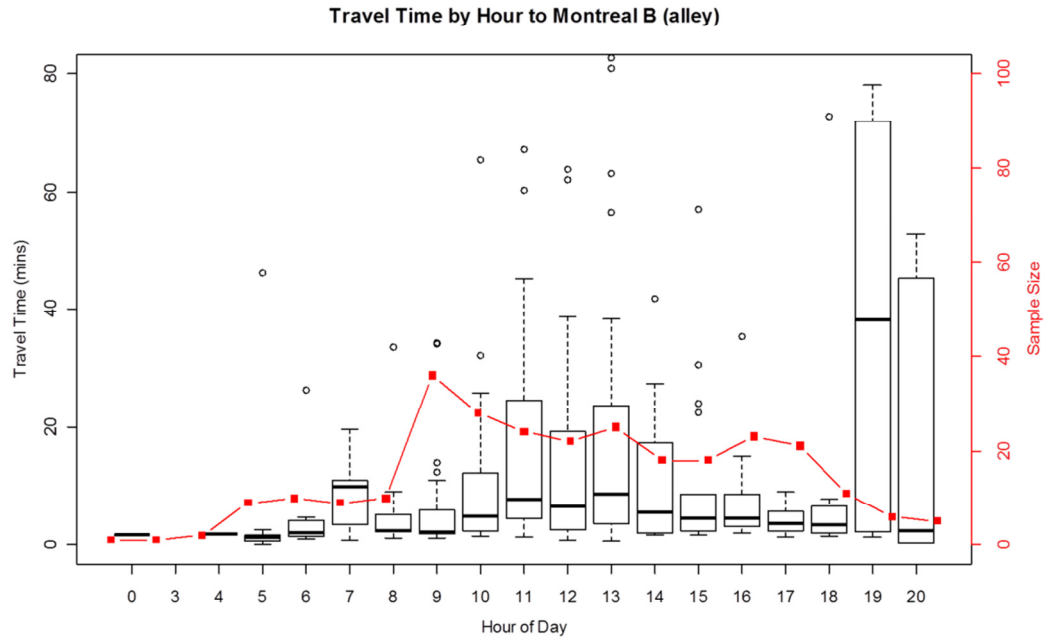
between the two locations. The time spent can be interpreted as travel time, provided a consistent approach to calculating it is adopted and the travel time between points is understood to also include any signal delay or other "dwell" time. Once again, the ambiguity of the detection time of Bluetooth devices leaves us with some freedom to define convention – for example travel times can be considered from the first sighting of the MAC address at point "A" to the first sighting of the MAC address at point "B", or first-first. Or, the matching can be done as last-last or median-median. Within the following sections, a first-first convention is used, and was chosen arbitrarily as the difference in error between conventions was small.

#### *4.3.2.2 Experiment Results*

Hourly summaries of travel times in each direction recorded in the Montreal corridor are shown in Figure 4-16. As mentioned before, travel times were calculated using the first-first paradigm, where the first sighting at each sensor is used to determine the travel time interval. Most travel times collected were under 10 minutes (median 4.27, average 11.48 minutes), with no apparent bias towards either direction (277 devices travelled from A->B and 279 travelled in the reverse direction). Since the section examined has multiple cafes, restaurants and shops, average travel times higher than the necessary time to walk the whole path distance are to be expected. It can be seen in Figure 4-16 that the travel times to the alley become very long in the evening – this could be a result of individuals spending more time in pubs in restaurants.




a) Travel Times to Point A (intersection)




b) Travel Times to Point B (alley)

**Figure 4-16: Travel times between the two sensor locations in downtown Montreal**

Figure 4-17 shows the distribution of the measured travel times, with a clear exponential trend. The most common travel times are under 5 minutes, which indicates that most people do not stop for too long on their way through, although the average speeds are still lower than what is considered to be standard walking pace (3-4ft/s).

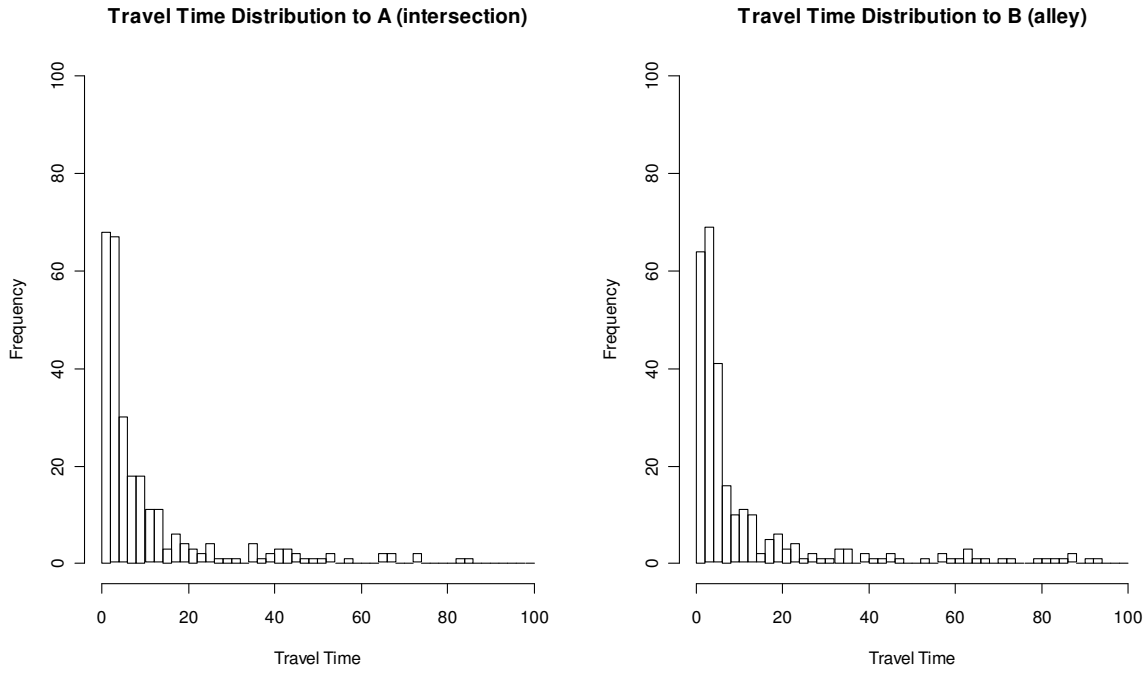
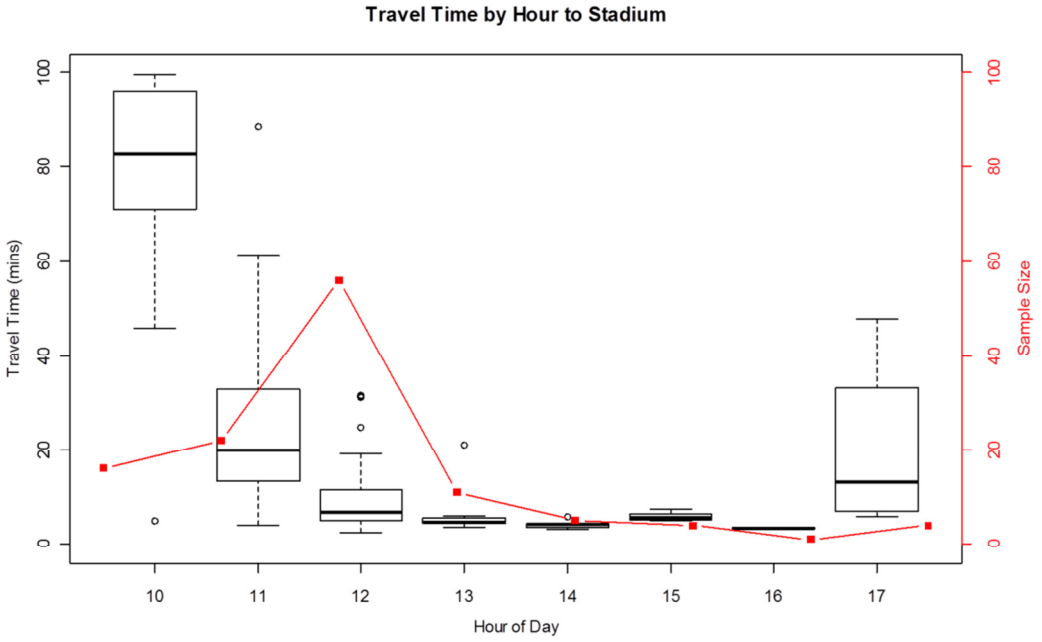


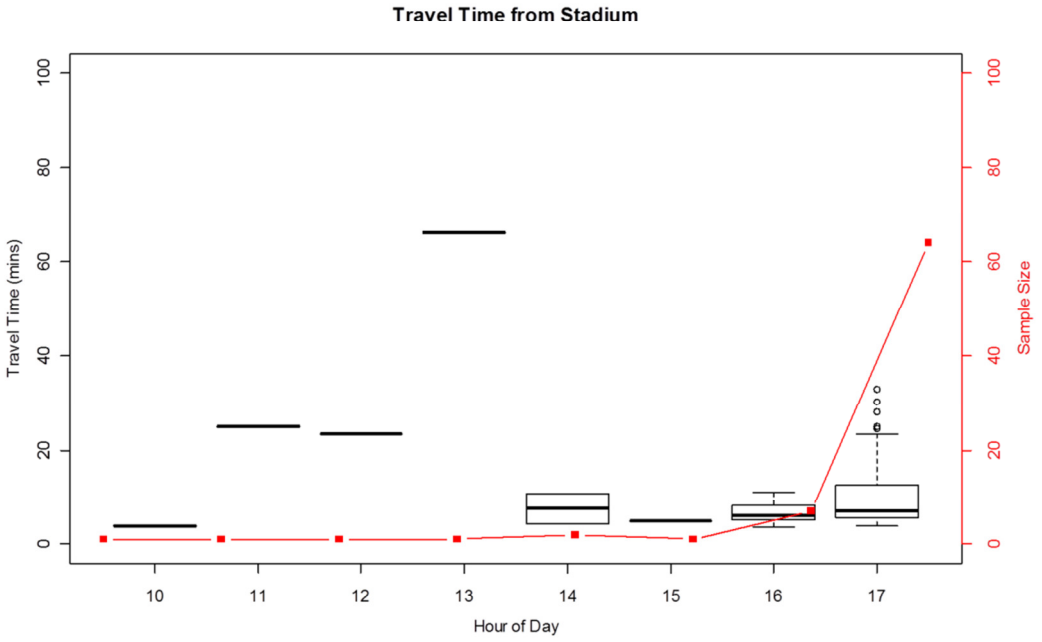

**Figure 4-17: Montreal travel time distributions in both directions**

Travel times in Seattle are higher and display strong directional trends, as can be seen in Figure 4-18. At first, visitors began arriving at the Ceremony over an extended period of time and their travel to the stadium takes a long time, likely due to them stopping to take pictures or

meet others and tour the campus. This is indicated by the higher travel time averages between the hours of 10:00 - 12:00. Once the Ceremony begins, travel times decrease; as people simply rush to the stadium directly (they're late!). There are also less devices present, suggesting that most people made it to the ceremony on time. The lower sample sizes also mean that the data collected during these intervals is less representative. Finally, a second peak occurs, now in the opposite direction, as people leave the stadium almost concurrently, this peak is narrower, lasting from 16:00 to 17:00. It can also be seen that no travel times were collected outside of the Ceremony due to low pedestrian volumes - data is in effect available only for the hours of 10:00 – 17:00.




a) Travel Times to Husky Stadium




b) Travel Times from Husky Stadium

**Figure 4-18: Travel times to and from Husky Stadium during Graduation**

Figure 4-19 shows the travel time histogram for the Seattle location – the most common travel time is around 10 minutes, which equates to a speed of roughly 0.58 m/s (1.91 ft/s), indicating that most people took time to do other activities besides simply walking from the fountain to the stadium.

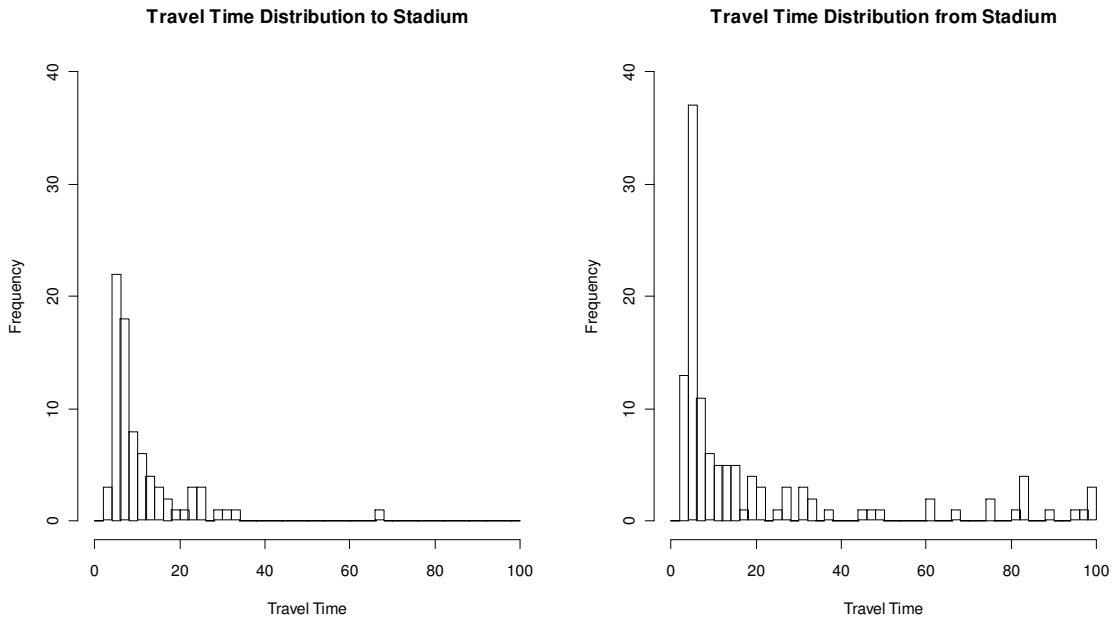
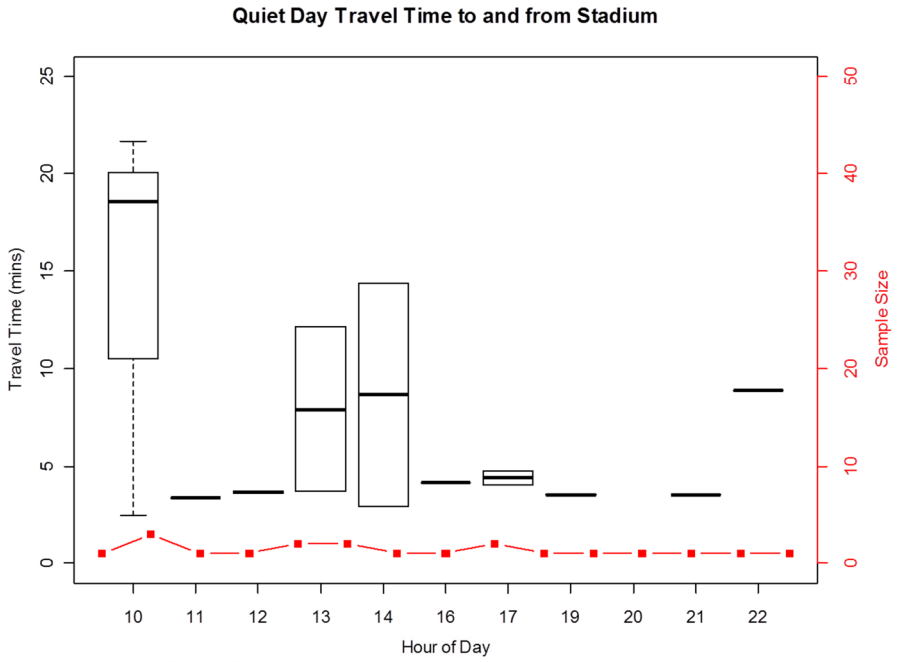


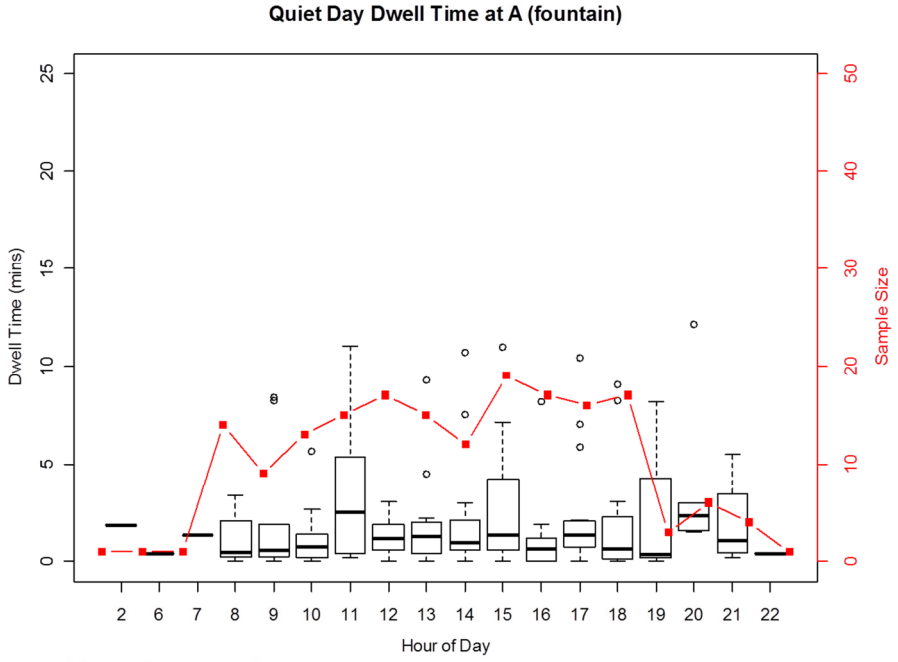

**Figure 4-19: Distribution of travel times to and from Husky Stadium in Seattle**

The graduation ceremony can be compared with the previous day 6/10/2011, which was the last official day of the quarter, during which no classes were held. No large events were held at Husky Stadium as well, a "quiet" day. In fact, only 17 devices were detected on the stairwell

over 24hrs (starting at 9:30am), compared to 181 detected at the fountain. Still, even the fountain device number is less than half of what was seen during Graduation. Figure 4-20 shows the travel times found between the locations and the dwell times experienced at the fountain. The travel time data is very sparse, often with only one or no data points per hourly period. The dwell times at the fountain are well represented, but are shorter than those experienced during Graduation, averaging 1.91 minutes vs. 2.14 minutes, perhaps implying less tourism activity.




a) Travel times the day before graduation




b) Dwell times at fountain

**Figure 4-20: Travel time and dwell time on a day before Graduation on the UW Campus.**

### 4.3.3 Highway Segment Application

#### *4.3.3.1 Introduction*

Since its inception, the primary focus of MAC address data collection in transportation has been vehicle corridor travel times. Most work on Bluetooth in transportation focuses on this area. One of the primary concerns with Bluetooth detection has been the ability to capture fast moving vehicles. Since the Bluetooth protocol requires up to 10.24 seconds to detect a vehicle, it is imperative that the detection range of the MAC address collection device is sufficient to work at high speeds, for example is a vehicle is moving at 60 mph, the detection zone needs to be about 900 ft (275 m) in diameter to guarantee that the vehicle is in range for 10.24 seconds.

#### *4.3.3.2 High-Speed Travel Time Collection Capability*

A freeway test was done on February 22$^{nd}$, 2009, early in the development cycle, to ensure that sufficient data could be collected when fast moving vehicles were present. The chosen corridor was a 3-mile long section along the SR-520 floating bridge in Seattle, WA at 24$^{th}$ Ave and 76$^{th}$ Ave overpasses. The speed limit on the bridge is 55 mph. Average speeds in free-flow conditions tend to be around 60 mph. A portable Advanced License Plate Reader (ALPR) system was loaned from WSDOT to check the accuracy of the obtained data. Figure 4-21 shows: a) the locations chosen for testing (the east side locations is at 76$^{th}$ AVE and the west side

location is at 24th AVE) and b) the testing setup at the 24th Ave location. The MAC address readers were equipped with 7 dBi antennae.

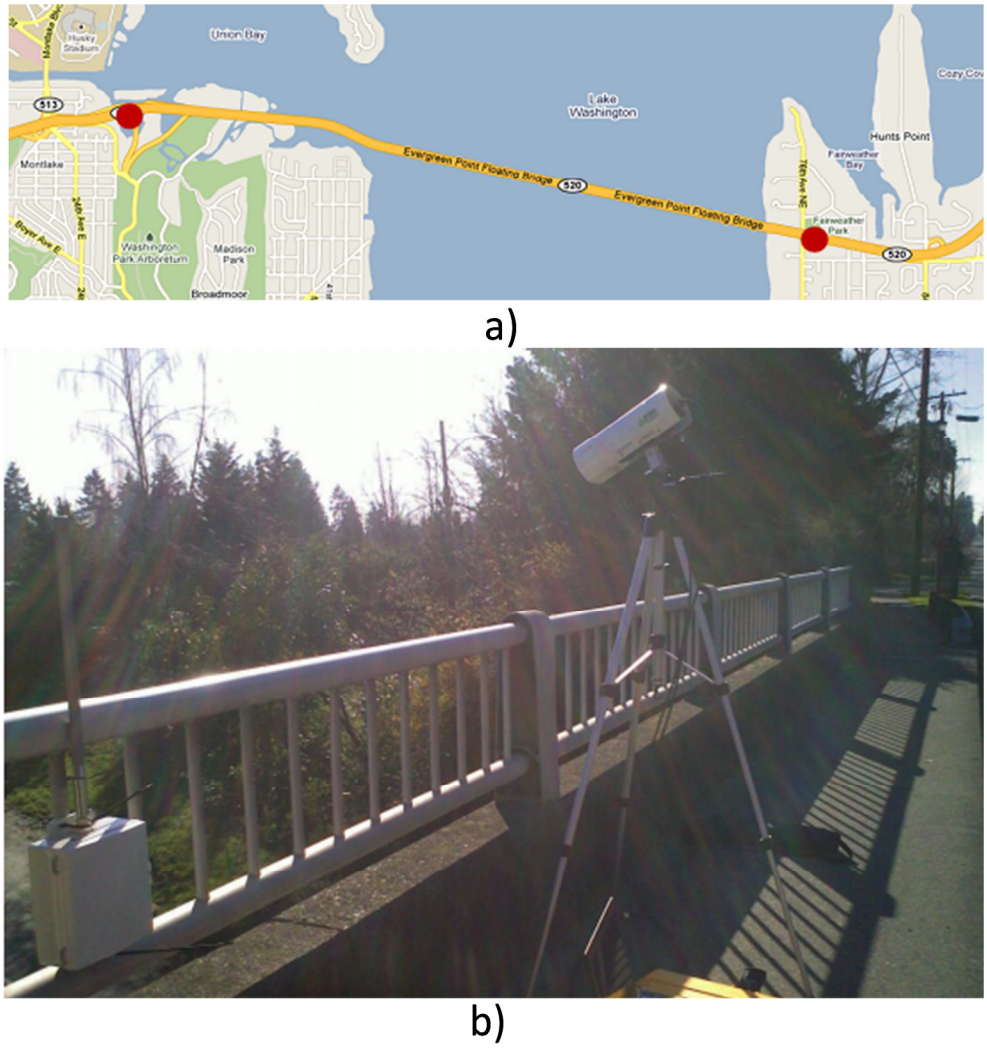



**Figure 4-21: a) Selected freeway test corridor on SR-520. b) Bluetooth sensor (left) and portable ALPR (right) used to collect travel time data at the 24th Ave location.**

The results confirmed the device's ability to collect data on freeways, with the system

collecting a sample that was consistent with what can be expected on arterials, that being around 10%. During the hour long test, from 8:00 am to 9:00 am the ALPR devices captured 1957 vehicles at the 24th Ave location and 1368 vehicles at the 76th Ave location. It is important to note that the ALPR sensors were capturing just one of the two lanes, and only one direction – westbound. The number of unique MAC addresses obtained at the two locations were 432 and 190, respectively. A shielding effect of one of the concrete barriers on 76th Ave overpass is thought to be responsible for the lower detection rate. The matching rate, or the ability to find the same MAC address in both the 76$^{th}$ St. and 24$^{th}$ Ave MAC address datasets, was 61% for the corridor, 116 matches (of a maximum possible 190), compared to the ALPR system's 39% or 533 matches (of a maximum possible 1368). Although the ALPR system was able to obtain more samples from a given direction, the MAC address method was capable of covering all lanes and both directions while providing a higher matching rate.

Figure 4-22a shows the comparison between ALPR and Bluetooth travel times on SR-520 in the westbound direction (the only direction measure with ALPRs). The travel times were aggregated into an hourly average, and the travel times in each lane were assumed to be identical (both lanes are standard general purpose lanes). The average error, or difference in Bluetooth-derived average travel time compared ALPR-derived travel time, for the hour-long test was 9.6%, ranging from 6% to nearly 20%. One of the most noticeable trends is the fact that all the error obtained was positive. In other words, Bluetooth based travel time estimates were consistently above the "ground truth" ALPR measurements. However, in this test the exact

location of the centerlines and detection zones of the Bluetooth and ALPR sensors was not known, thus a compensating adjustment had to be made. The two data sources were adjusted to a common mean. After a mean shift of .293 minutes (the mean difference between the travel times in the two datasets), the error rates reduced to a maximum of 9.4% and a minimum of -3.95%, within the FHWA recommended values for travel-time reporting. Figure 4-22b shows the resulting error and Bluetooth travel times after adjustment.

Although the SR-520 test site would have been ideal for longer testing using a number of configurations, the use of a portable ALPR unit required in-person data collection at both ends of the corridor. Further restrictions were encountered due to WSDOT security concerns on freeway overpasses, therefore allowing only an hour of testing to be performed. SR-522 is equipped with permanently deployed ALPR units, making data collection there easier.

It should be noted that the Bluetooth readers were mounted at a height of about 30 feet above the roadway in this scenario. This results in a larger detection zone compared to what is experienced when the sensors are mounted near ground level (about 5 to 7 feet). The antennae used in the experiment have downward tilt of about five degrees, so the range of the antenna increases with height above ground plane. With the sensors mounted at a height of 30 feet, the detection range theoretically grows to about 400 feet (radius), giving an 800 foot detection zone, or the capacity to detect about 80% of the “detectable” traffic, which is consistent with the 60% matching rate observed.

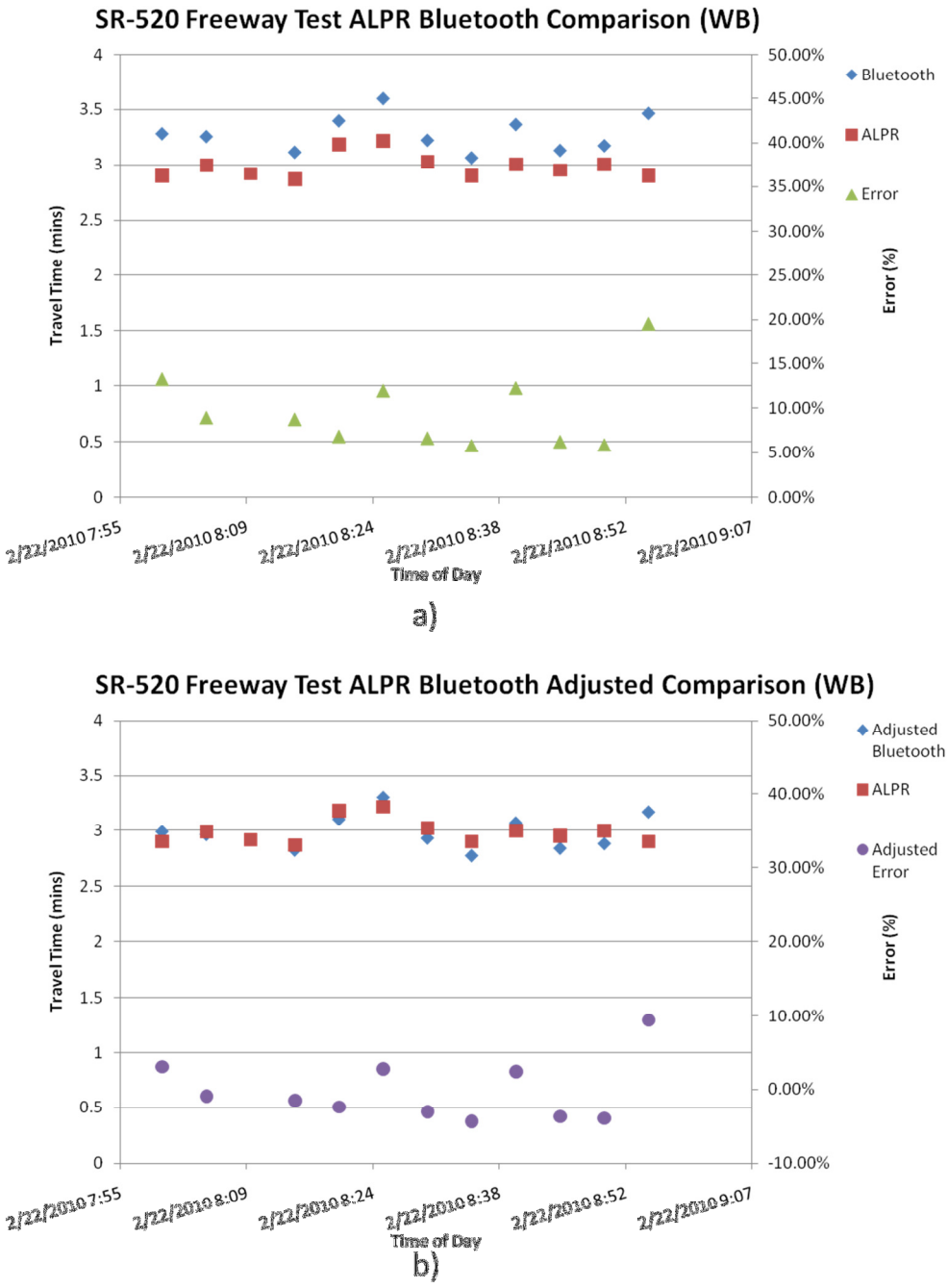



**Figure 4-22: SR-520 freeway test**

### 4.3.4 Corridor Paradigm Issues

The tests described above address issues regarding the capability of the designed devices to accurately collect a sufficient sample size. MACAD has been demonstrated to work on high-speed corridors, where the detection time is limited, and in rural areas, where the prevalence of Bluetooth devices has previously been considered to be lower. As such, MACAD could be used for mobile device MAC sensing in a range of locations and provide sufficient data if Bluetooth devices are present. However, the issue of temporal uncertainty remains – the range of error (from 6 to 20 percent) gives a clear indication that there are additional investigations that are necessary to relate detection range, vehicle speed and travel time error.

## 4.4 Mobile-node Data Collection Paradigm Applications

### 4.4.1 Introduction

Let us imagine a dense urban core network, which is typically grid-based. Figure 4-23 shows three images of the network in consecutive order by time. Three individuals populate this network. The large (blue) circle in the upper left represents an “observer”, or an individual that constantly records GPS coordinates and scans for the MAC addresses of surrounding discoverable devices as they progress through the network, using an app that is installed on their smartphone. The smaller (red) circle, surrounded by the larger (blue) one is an individual that is not recording anything, but is broadcasting either Bluetooth or Wi-Fi signals, so that their MAC address is visible. Since the red circle is in range, their unique identifier has been revealed, and their location at t=0 is known to be the upper leftmost intersection. The other large (yellow) circle at the bottom right represents yet another “observer” moving along the grid. As time progresses, the individuals continue about their business and move in separate directions. It should be noted that at t=1, the position of the red circle is not known and can only be imputed from data collected later at time t=2. At that time, shown in the last frame, the yellow circle overlaps the red, once again noting the unique identifier. At the end of the sequence, we have measured travel time across six segments, encountered one additional device and were capable of determining the spatial and temporal characteristics of the three individuals, all while providing software to just two.

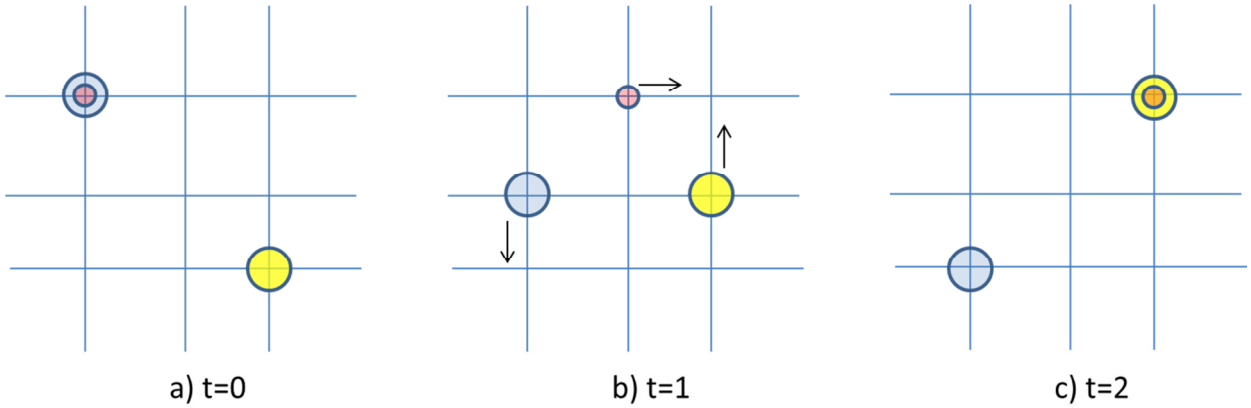



**Figure 4-23: Hypothetical time lapse scenario of two agents and a detected user**

As a result of the above paradigm, a number of system parameters can be obtained with relatively little effort. Specifically, variables that can be directly measured; Number of External Devices Encountered, Number of Measured Network Segments (assuming entities travel only on network); and system variables that act as constraints: Bluetooth and Wi-Fi Activity Rates, Device Range, Percentage of Population as Probes. With the given variables and constraints, it becomes possible to express the following given a network topology: Spatial Distribution of Encounters, Temporal Distribution of Encounters, Travel Times along Segments and Segment Lengths.

### 4.4.2 Mobile Node Approach Simulation

To further illustrate the concept and the variables at play, a simulation was created using VISSIM (PTV, 2012), a popular traffic modeling tool, and the Processing programming language

(Reas et al., 2007). The overall architecture of the simulation is shown in Figure 4-24. In VISSIM, a network is constructed and populated with entities that follow prescribed routing patterns. The simulation then outputs second-by-second updates on the location of each entity within the simulated lifespan of the model. These updates are interpreted as object trajectories in a custom Processing module. Within the module, some of the simulated entities are considered to have Bluetooth devices visible, and some are considered to be acting as mobile Bluetooth sensors. As the entities progress through the network, those acting as sensors have a chance to detect each other, as well as those entities that are currently Bluetooth-visible. The percentage of the population carrying devices, as well as the percentage acting as mobile sensors is user-adjustable, as are the specifics behind the detection mechanism (currently set to a uniform distribution to simulate Bluetooth).

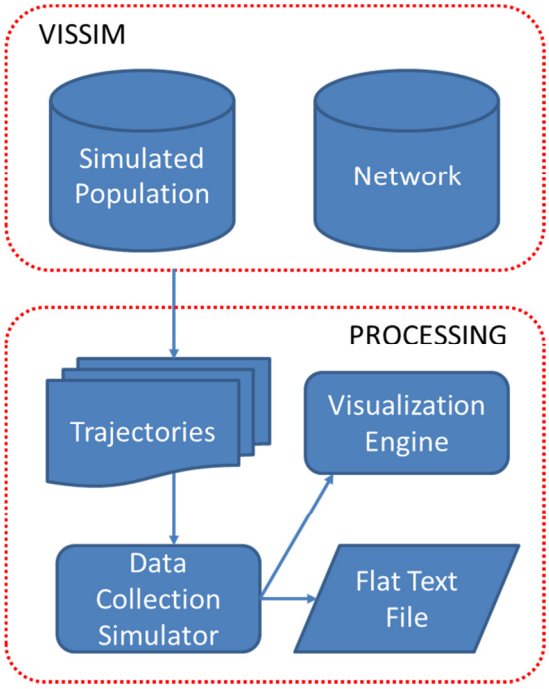



**Figure 4-24: Mobile sensing simulation architecture**

Figure 4-25 shows the simulation tool in action – Figure 4-25a shows the simulated entities as they progress through a network – in this case, it is a simple grid network on which typical VISSIM pedestrian entities are traversing from one corner to the furthest other corner. The lower left corner to upper right corner movement is double the volume of the others. The radius of detection is set at 10 meters (standard Class II Bluetooth device, which is most common) and 892 entities are simulated in this example. Figure 4-25b shows the distribution of detection events on the network, which 4-25c show the raw Bluetooth-based trajectories


**Figure 4-25: Simulation of 892 entities on a simple grid network**

Using this tool, some basic characteristics of the mobile sensing paradigm can be gleaned. Perhaps of greatest interest is the relationship between the number of mobile sensors, or "observers" and the percentage of the total visible detected devices to be detected. Using the simulation tool, the following exponential relationship was obtained for the grid network scenario:

$$O = 1.8038e^{0.0415*D} \qquad \textbf{(Equation 4-1)}$$

Where $O$ is the percentage of the population acting as observers and $D$ is the percentage of the total Bluetooth-visible population detected. However, it should be noted that this relationship is highly dependent on the network, and re-running the same 892 individuals on a single link produces a different, albeit still exponential relationship:

$$\boldsymbol{O = 0.6772e^{0.034*D}} \qquad \textbf{(Equation 4-2)}$$

The data used to estimate these relationships are shown in Figure 4-26. Although the relationships themselves are likely to change for different populations, networks, and MAC broadcast protocols, the exponential nature of encounter-based detection is likely to remain, as the nature of the problem is similar to those studied in disease transmission, which base infection rates on exponential models.

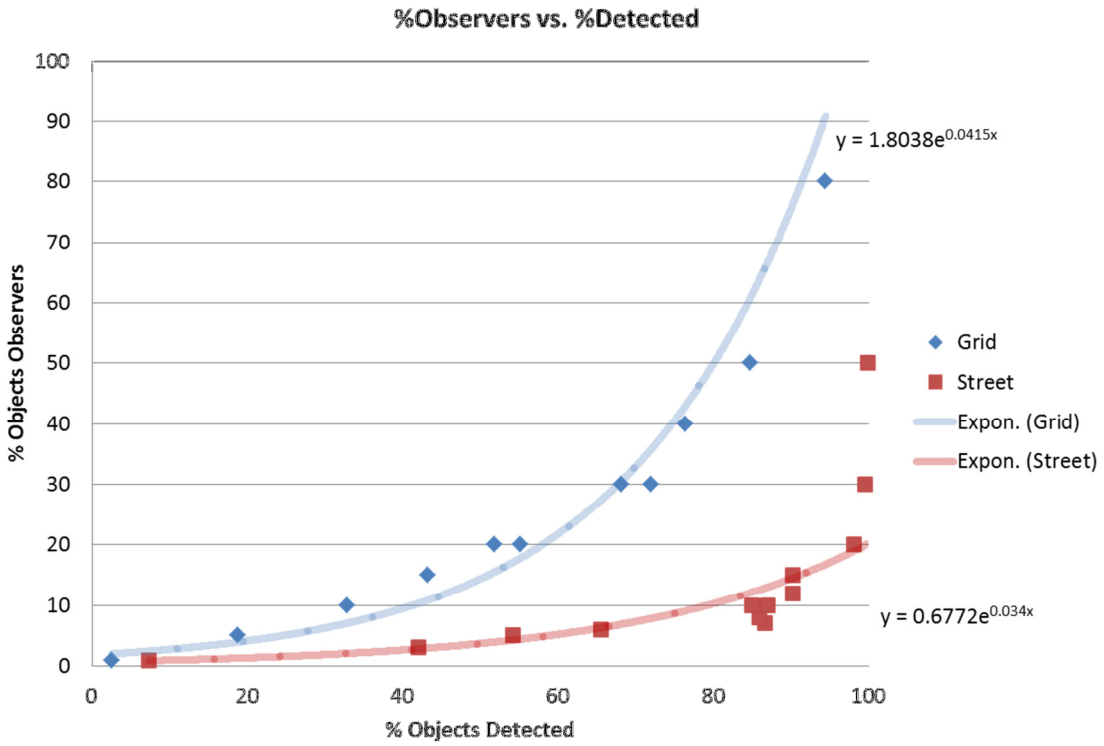



**Figure 4-26: Relationship between the number of observers and the proportion of visible population detected**

The exponential nature of relationship between the number of detectors and the observable population at that level implies that even at low levels of observer population percentage, a sizeable amount of the observable population can be detected. This is encouraging, as the portion of the population that is likely to participate by downloading such an app is likely to be small.

### 4.4.3 Pedestrian Route Estimation Application

Most smartphones on the market have Bluetooth and GPS functionality, making them perfect platforms for the mobile monitoring approach. Google's Android operating system is quickly becoming one of the most popular mobile device platforms, in part due to the open source nature of the development environment, which allows end users to create apps with minimal inconvenience and effort. In light of this information, a small app was written for the Android operating system that continuously scanned for surrounding Bluetooth devices and recorded the current GPS coordinates of the device. WiFi-based location services were turned off to ensure that there would be no errors that could result in case of hand-offs and switches between GPS and WiFi. Figure 4-27 shows a Motorola Droid phone running the software, displaying a detected MAC (of a device belonging to the author), while still finding its current location. This particular device is equipped with a Class 2 Bluetooth chipset, granting a range of around 10m for detection of surrounding Bluetooth devices.

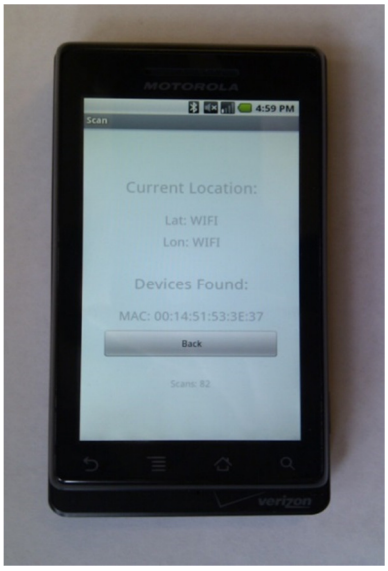


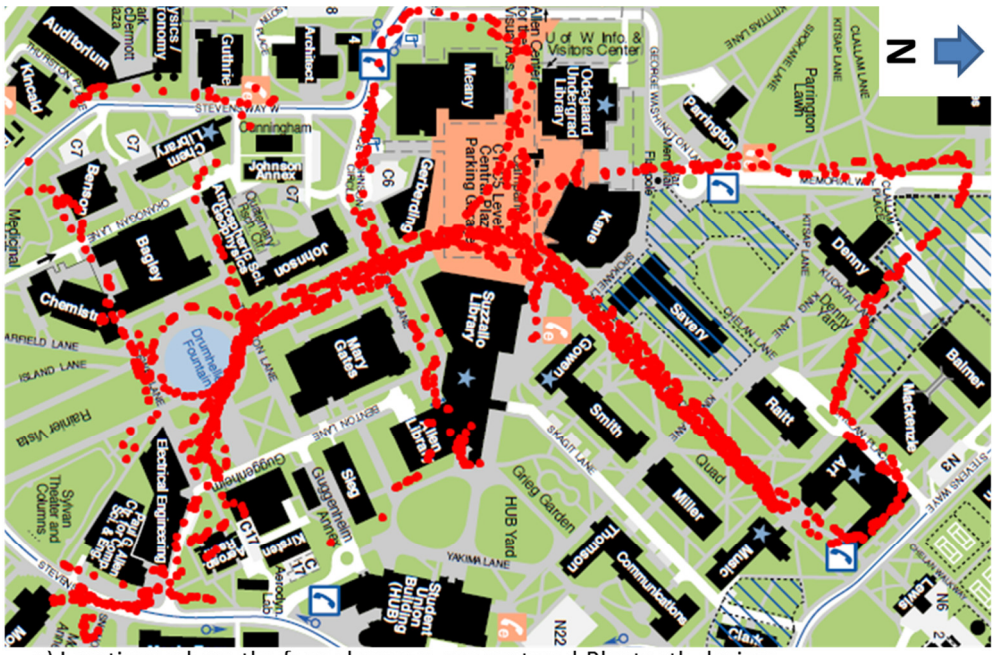

**Figure 4-27: A Motorola Droid handset running the Mobile Monitor application**

(Phones used in study courtesy of Dr. Alan Borning)

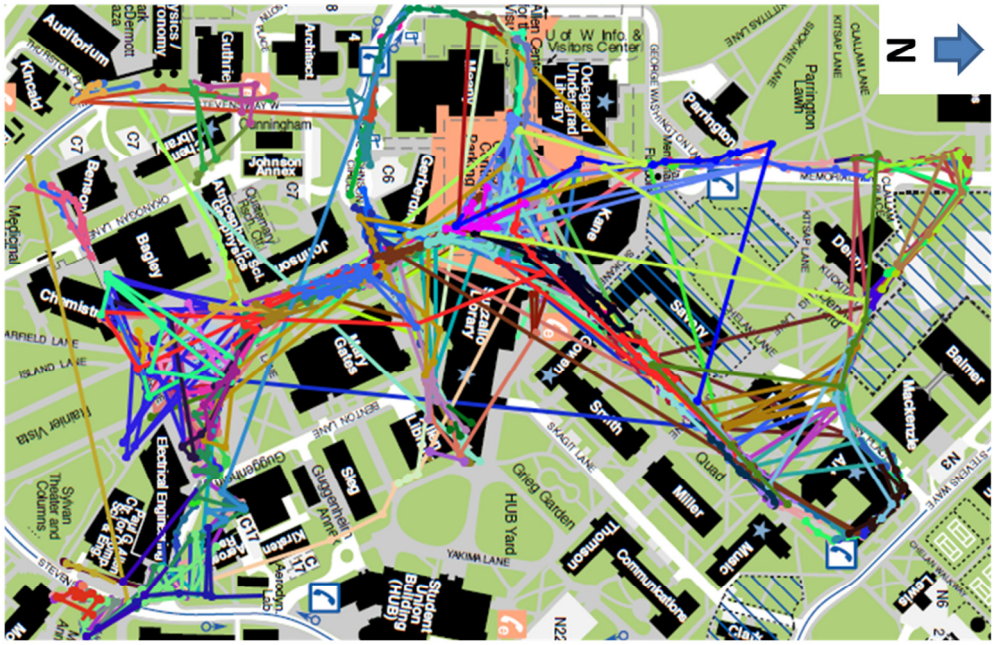
#### *4.4.3.1 Study Site*

Four Motorola Droid phones were used in the experiment. Four volunteers (hereafter called "observers") walked for fifty minutes from 1:10pm to 2:00pm on 4/20/2011 (sunny, warm) at the University of Washington central campus, encountering Bluetooth devices along the way. The locations of Bluetooth encounters (which are representative of the paths) are shown in Figure 4-28a. During the short 50-minute experiment, 546 unique devices were discovered by all four observers. The collected sightings were then composed to create device trajectories, shown in Figure 4-28b. The trajectories were created by plotting the coordinates at which the MAC address has been seen. There are two types of trajectories that were observed – ones that resulted

from the observer following a particular device and walking alongside (shoaling) and ones where a device was seen momentarily by more than one observer (encounters). These encounters often occur at longer distance intervals, and can result in trajectories that are unrealistic if plotted without network knowledge, as can be seen in Figure 4-28.



a) Locations where the four observers encountered Bluetooth devices



b) Collected Bluetooth device trajectories

**Figure 4-28: Collected trajectories on UW campus on 4/20/2011 1:10pm to 2:00pm**

**using 4 observers**

#### *4.4.3.2 Experiment Results*

As observers progress through the environment, they may encounter both static and moving devices. One of the main concerns during this proof-of-concept experiment was that predominantly static devices would be encountered. A graph of distance vs. travel time of all discovered device trajectories is presented in Figure 4-29. Each point represents a single device trajectory, composed of discovered links (points at which the device was repeatedly detected). The distance is defined as the sum of the discovered links, while the travel time is the sum of the time intervals for each link. Looking at Figure 4-29, it can be seen that there are several clusters of devices. One cluster, for points below ~500 seconds (8.3 minutes) of travel time, shows a clear linear trend. These are matches that were made quickly, thus the crow's-flight distance remain a valid approximation of the taken path. Other clusters tend to have very short distances, but longer "travel times" – these are likely to be static devices encountered by different observers. For example, observer "A" may have walked past Suzzalo library, where many static devices are present. These were recorded and when observer "B" walked by 20 minutes later the MACs were matched, creating a trajectory. Because of the possibility of GPS error and temporal uncertainty of Bluetooth detection (discussed in the previous section), the devices appear to have moved a short distance during these 20 minutes. While static device discovery may be of interest for population estimation purposes, the primary purpose of this paper is to evaluate this approach for capturing pedestrian travel. Figure 4-30 shows data that excludes trajectories lasting over 500

seconds (sample size = 311/546) and a linear model with a slope of 0.54 m/s – equating to a speed of about 1.77 ft/s – a fairly low value for walking speed, usually considered to be around 1.2 m/s (4 ft/s) on average.

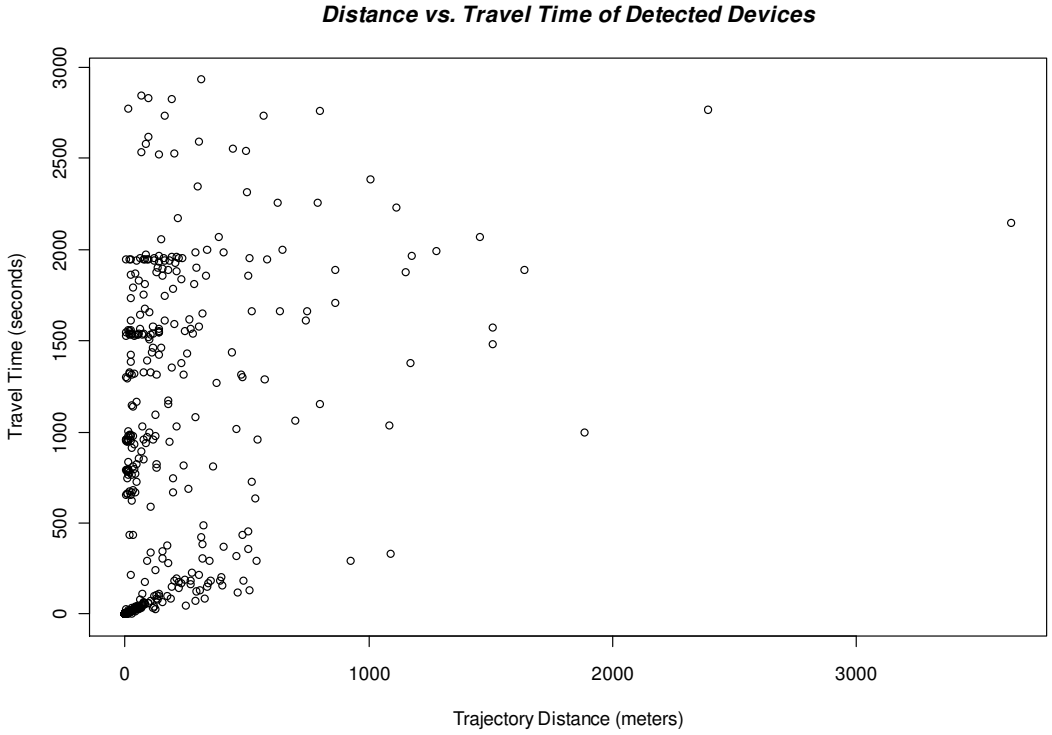



**Figure 4-29: Unfiltered Device Distance vs. Travel Times**

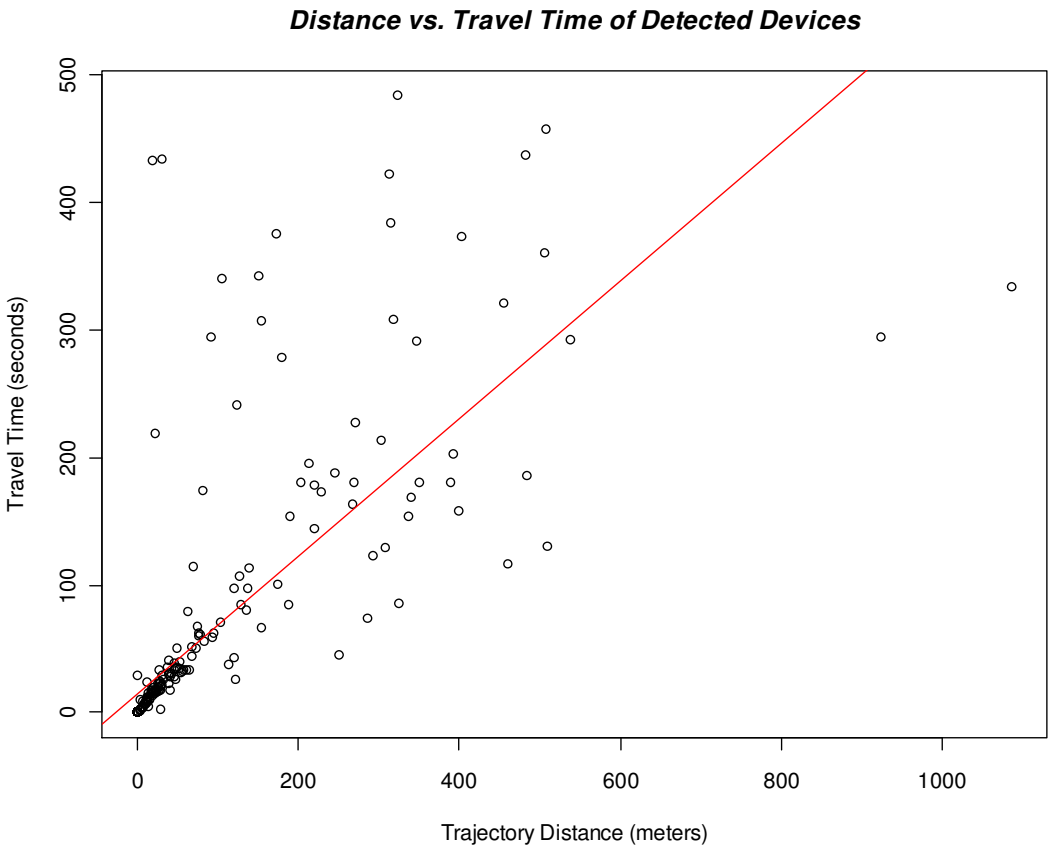



**Figure 4-30: Filtered Device Distance vs. Travel Times**

Figure 4-31 shows the speed distribution of the filtered (less than 500 second) travel times. Here, the peak occurs around 1.6 m/s or 5.2 ft/s, a much more likely walking pace for a college campus. The lower value seen in Figure 4-30 may be due to static devices that were discovered in consecutive scanning intervals, resulting in matches with very low distances and travel times. Besides these suspect lower values, there is little other noise, as the experiment was confined to a college campus, where walking is the dominant mode.

To further explore the collected data, a timeline plot is shown in Figure 4-32, showing the number of devices detected every minute. Class break occurs from 1:20 p.m. to 1:30 p.m. and this is apparent in the data – there is a large peak happening at around 23 minutes past the hour, as students shuffle between classes. The highest peak is around 50 devices/minute, and the average during the experiment is about 11 devices/minute (about 2.73 devices/minute for each observer). Once classes begin, fewer devices are discovered, as the campus becomes much less active.

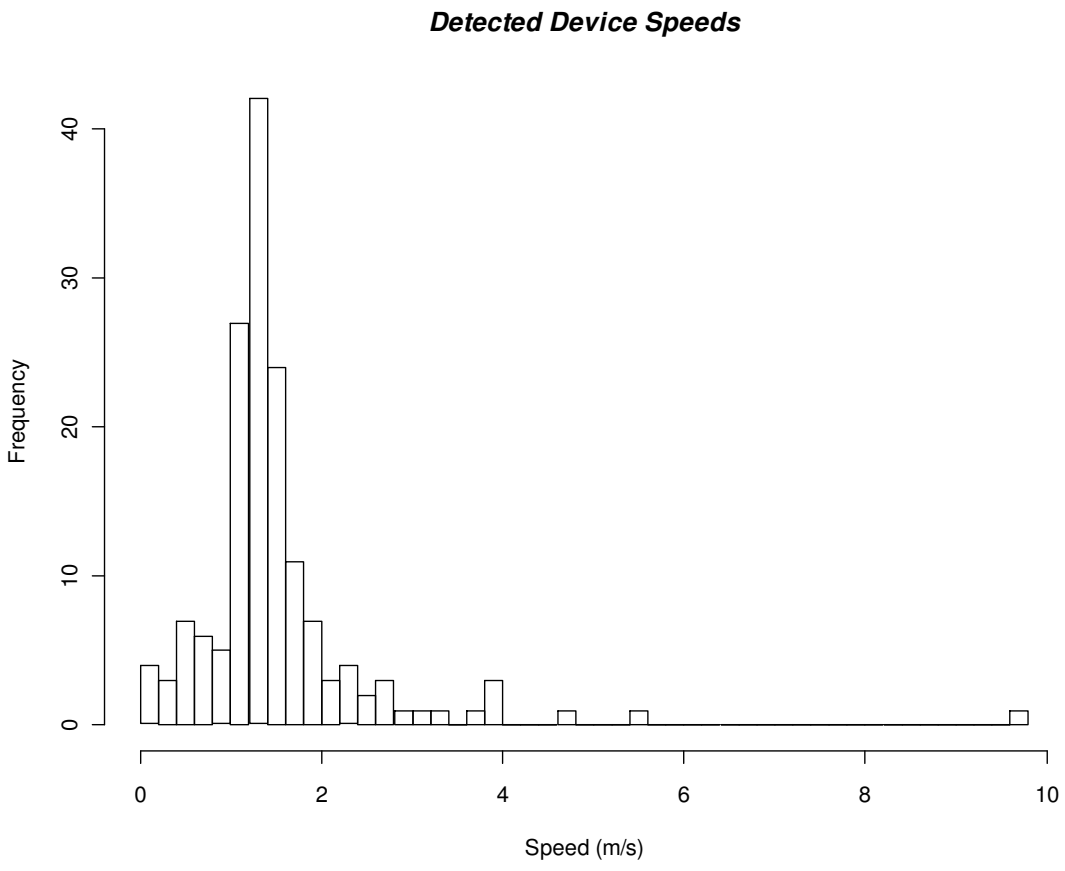


**Figure 4-31: Filtered Speed Distribution of Discovered Devices**

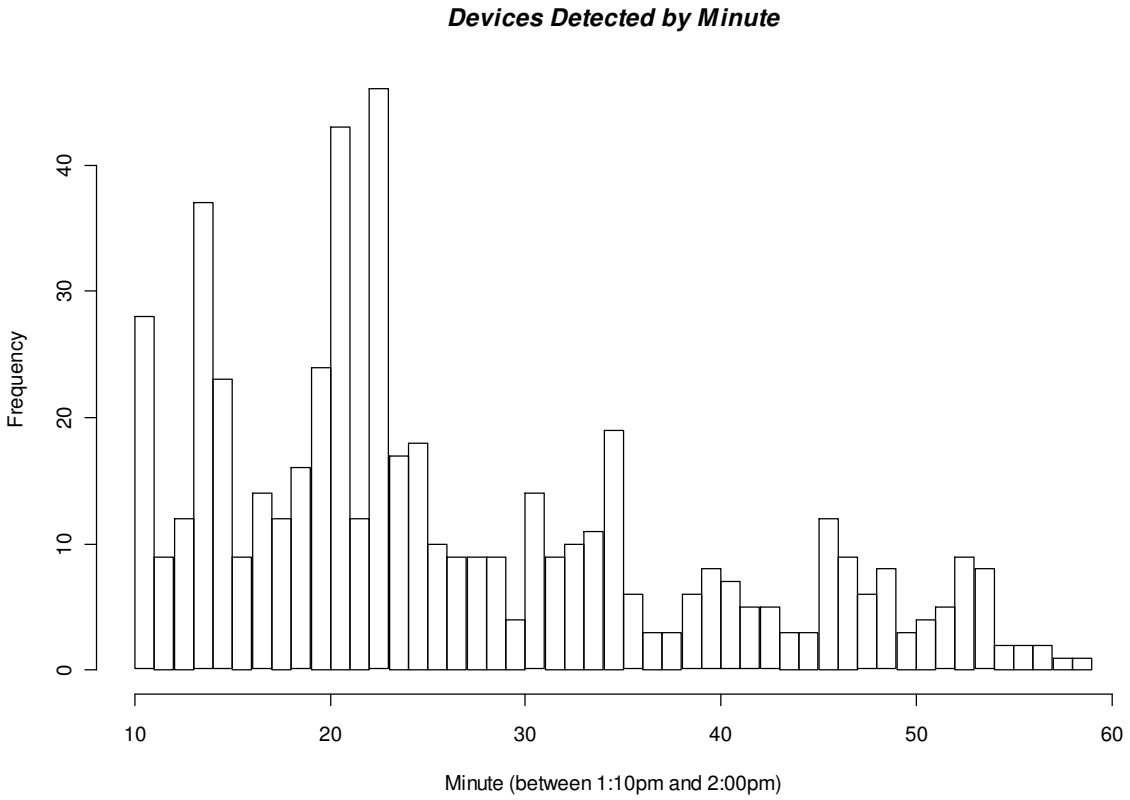



**Figure 4-32: Devices Detected by Minute of Experiment**

To get a better idea of the relative sample size of this approach a verification count was done on 5/2/2011 (also a Monday, sunny, warm) to compare the obtained results to a manual count and static Bluetooth detection conducted in the highest volume thoroughfare, between Mary Gates Hall and Johnson Hall. Two Droid phones were given to volunteers, who stood across from one another along the line shown in Figure 4-33. Pedestrian trip-line counts in 5-minute intervals and detected Bluetooth devices were recorded. Results obtained during the 30-

minute count are shown in Figure 4-33. During the test, 1817 persons walked between the two observers and 41 devices, or 2.25%, were detected (after filtering for static devices in the area – devices that were present the entire experiment duration).

| Time | BT Count | Filtered BT Count | Ground Truth | % Detected |
|---|---|---|---|---|
| 13:15 | 28 | 14 | 483 | 2.89 |
| 13:20 | 11 | 9 | 533 | 1.68 |
| 13:25 | 9 | 8 | 620 | 1.29 |
| 13:30 | 2 | 1 | 89 | 1.12 |
| 13:35 | 10 | 9 | 56 | 16.07 |
| 13:40 | 1 | 0 | 36 | 0 |


**Figure 4-33: Manual Data Collection Results**

### 4.4.4 Mobile Paradigm Issues

In addition to the previously discussed notions of population and temporal uncertainty, a new type of uncertainty arises under the mobile sensing paradigm – a spatial uncertainty with respect to the exact route chosen by the detected entity in between the opportunistic detections afforded by the mobile sensors. Because we no longer can control sensor placement, it is difficult to assert which corridors, or links or routes were used by the owner of the re-identified MAC device. A way to obtain the most plausible path between two detections is necessary in order to resolve the spatial uncertainty issue. This, along with the population and temporal components of the overall uncertainty are discussed in the following chapter.

# Chapter 5 Strategies for Systematic Reduction of Population, Temporal and Spatial Data Uncertainty

## 5.1 Data Uncertainty in MAC sensing

The applications and corresponding experiments discussed in the previous chapter suffer from a common set of limitations, caused primarily by data sparseness. Due to the opportunistic nature of the protocols involved in MAC sensing, combined with the incomplete proliferation of Bluetooth devices into the general population (and limited Bluetooth-visibility) it can be argued that the data that is collected in the applications discussed in Chapter 4 are sparse. The sample size was shown to range from two to ten percent of the population in the experiments above. This uncertainty in sample size and composition can create volume estimate errors in applications that rely on counting and population bias errors in applications that necessitate a representative sample of the population, e.g. bus stop wait time. As for temporal uncertainty, the travel time error on SR-520 was roughly ten percent, indicative of the temporal uncertainty component.

However, the temporal component can also cause dwell time errors, which may be important in applications such as signal delay estimation. Finally, the sheer number of possible routes that can be assigned under the mobile detector paradigm gives rise to trajectory estimation error. A diagram of the uncertainties, the primary theme of this dissertation, is shown in Figure 5-1.

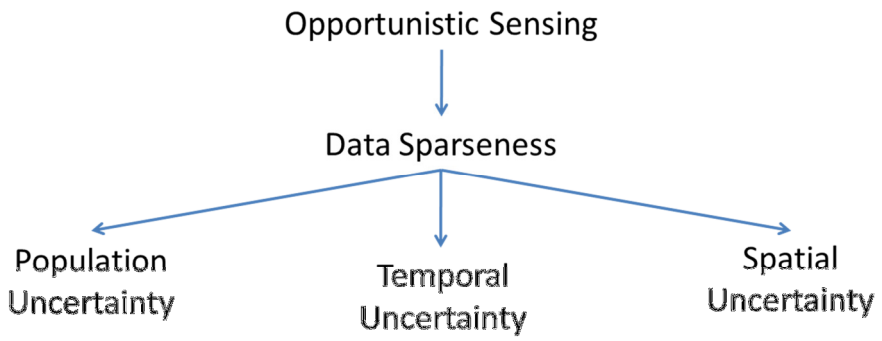



**Figure 5-1: Inherent issues in opportunistic MAC sensing paradigms**

It should be noted that each of these uncertainties can propagate in to a variety of errors, from volume count error due to under-sampling and/or population bias, travel time error due temporal uncertainty (and potentially population and spatial uncertainty as well) to trajectory error which could also be affected by all three uncertainties. Strategies for reducing the errors associated with these uncertainties are discussed in the following sections. A number of different approaches are examined, from basic filtering and thresholding, to hardware-based approaches and data-mining techniques.

## 5.2 Analysis Tools

To evaluate the various strategies for uncertainty reduction, a series of tools were created for the DRIVE Net platform, accessible at http://www.uwdrive.net. The platform allows for visualization of transportation data on an interactive map. The user can request a number of pre-determined analysis types, or can choose to export data to perform custom ones. The visualization and routing tools were implemented using Open Street Map (Open Street Map Foundation, 2012), PostgreSQL (The PostgreSQL Global Development Group, 2012) and PGRouting (PGRouting, 2012). Analysis tools are implemented in C# and R (R Development Core Team, 2008), and connect via a MySQL database and through the Rserve package (Urbanek, 2003).

### 5.2.1 MAC Matching and Filtering Engine

To facilitate MAC address data analysis, a MAC matching and filtering engine was built using the C# programming language. The primary purpose of this engine is to generate dwell time, travel time and trajectory data from a set of MAC addresses, depending on experiment type. The system has been designed to ingest MACs coming in either via a MAC database or a flat text file containing MAC address data. Figure 5-2 shows the engine schema and the corresponding interface. The obtained MACs are first sorted by individual, MAC address, thus creating a sequence of sightings for each address encountered. Then, a gap-time analysis (looking at sufficiently long gaps within the sequence) is conducted to determine the number of

trips/dwells the sighting sequence contains. The gap time is set by the user. Once the individual trips/dwells are obtained, three types of analyses can be conducted on the data - one to obtain dwell time, one to obtain travel time and another to output the trajectory of a given device. The latter is used in mobile device data analysis only, while the first two analyses are used primarily for static sensors. Basic filtering, discussed in sections 5.3, 5.4 and 5.5 is also performed prior to output. The dwell time analysis ensures that a particular device is continuously checking in, and outputs the number of check-ins as well as the check in and check out times. The travel time analysis provides info regarding the first and last sightings of a device at each sensor, arranged in a O/D matrix. Trajectories are output as information regarding a particular sequence of sightings, e.g., the latitude and longitude of the detected device and the corresponding time stamps. The output can be piped directly into the DriveNET database, or an automatically generated Excel sheet.

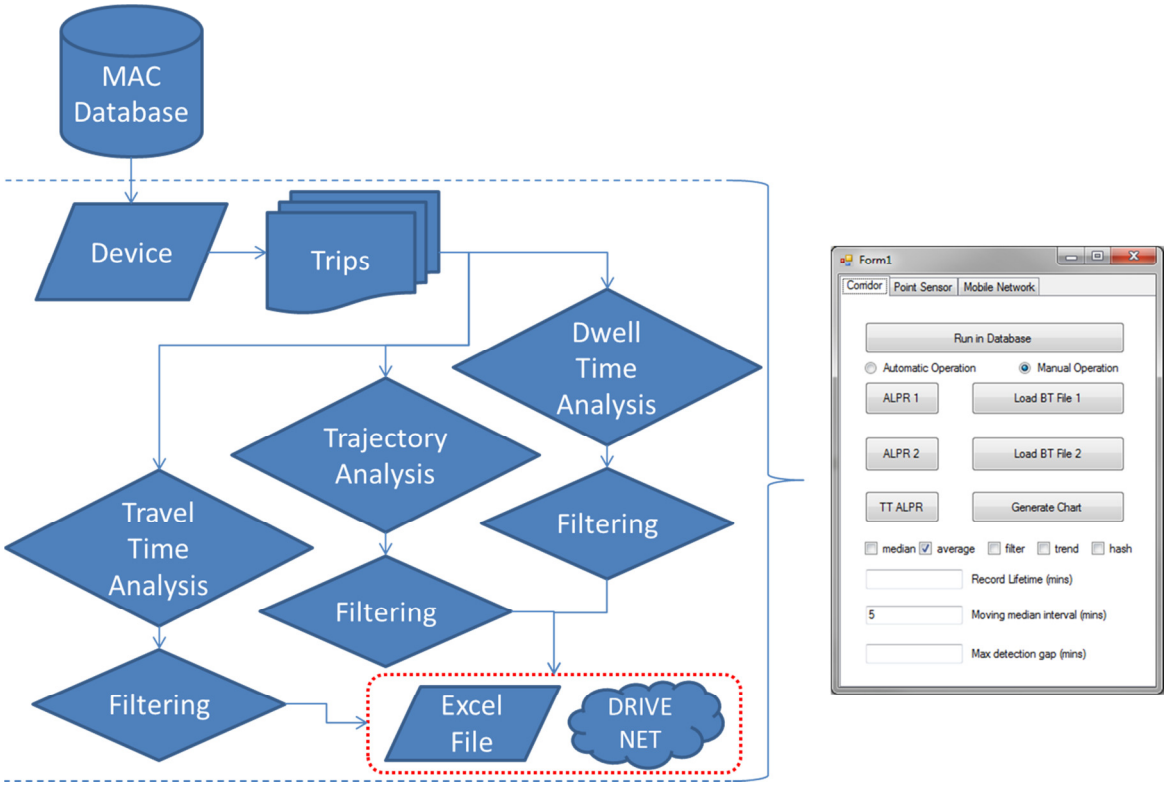



**Figure 5-2: MAC Matching and filtering engine schema and interface.**

The implemented tools were meant to be sufficiently general for re-use in other data analysis applications. Using the MAC matching and filtering engine, two tools were implemented as tabs on the main webpage, were created: (1) a "Corridor Sensor Comparison" tab and (2) a "Mobile Device Routing" tab. Each is explained in greater detail below.

### 5.2.2 Corridor sensor comparison tab

A screenshot of the Corridor Sensor Comparison tab can be seen in Figure 5-3. This tab allows for quick and seamless data integration and comparisons, making it an ideal candidate for

data quality evaluation. Figure 5-3 below demonstrates the system in action – the user has selected two instrumented corridors for a week during December 2012 and is able to run an analysis on the selection by pressing the "Compare Corridor Sensors" button. This opens up two sub-windows containing R-generated images and tables regarding the comparison made, as shown in the figure. A scatterplot, a cumulative density plot, a q-q plot and histograms for the entire selected dataset are generated. Additional analyses including sample summaries, Kolmogorov Smirnoff tests, paired t-tests and error rates (MAPE, MPE, RMSE and MAD) are shown as well (taking ALPR as ground truth). The number of sub-windows generated is dependent on the number of corridors selected. The comparisons contained are dependent on the sensor groups selected, but can include up to all of the following: ALPR sensors, Bluetoad Bluetooth sensors, Sensys Systems sensors, Blip Bluetooth sensors, Inrix probe vehicle data and UW Bluetooth sensors.

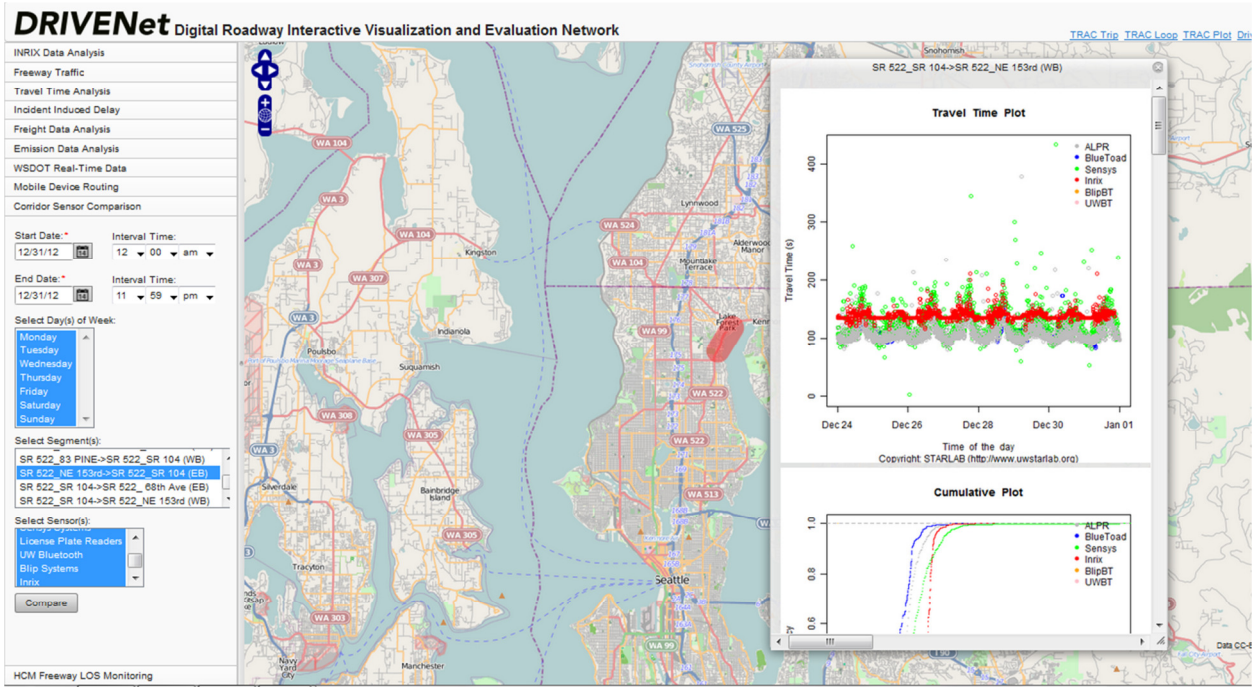



**Figure 5-3: Corridor Sensor Comparison tab in DriveNET**

### 5.2.3 Mobile device routing tab

The Mobile Device Routing tab was created to visualize and analyze data created by mobile sensors, but is also capable of showing results from static sensors for comparison purposes. Figure 5-4 shows the tab displaying mobile and static sensor data during the 4/20/2011 campus test. This tab functions as the main analysis tool for mobile device data and is the primary tool for the mobile device experiment analysis described in Section 5.5.

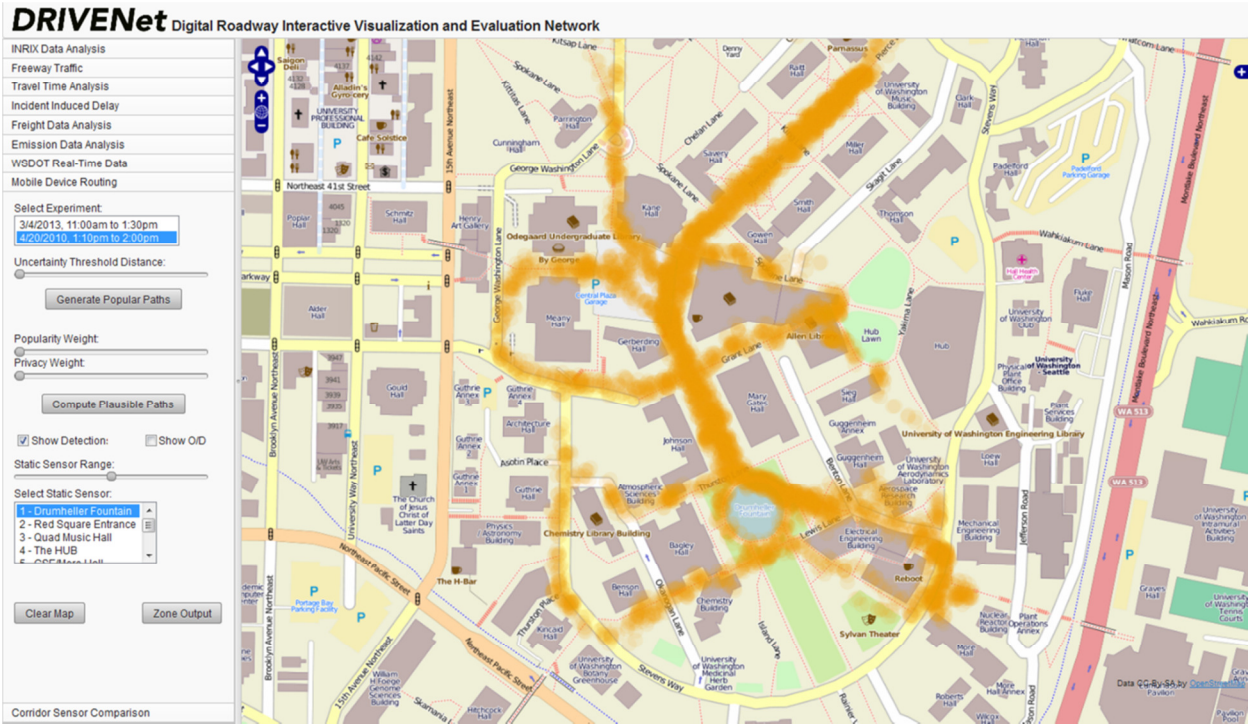



**Figure 5-4: Mobile Device Routing tab in DriveNET**

## 5.3 Reducing Population Uncertainty

Population uncertainty is a major issue when utilizing MAC address data. One of the most challenging aspects of MAC address data usage is low sample size and varying sample size – as mobile device become more prolific, more individuals carry them on their person. However, concurrent efforts to limit the exposure and misuse of MAC address identifiers often limit the Bluetooth-visibility of individuals. For example, many manufacturers now limit Bluetooth visibility to 120 seconds to reduce the chance of personal identifier exposure. One of the

potential solutions to sample size limitations has been the use of Wi-Fi MAC identifiers in tandem with Bluetooth ones. To capitalize on this data gain, the latest version of the UW MACAD device (v 3.0) was designed to capture Wi-Fi devices as well. Sample size, however, is not the only issue in MAC address data collection. Another issue has been the potential bias that can be introduced into the obtained dwell time, travel time or routing data from extraneous Bluetooth or Wi-Fi devices that belong to persons using modes that are not of interest. Because the technology is essentially blind (a device is a device, not a car, a pedestrian or a bus), special care must be taken to ensure that the sample obtained is representative of the population of interest – some MAC trips must be discarded from the data. Sections 5.3.1 and 5.3.2 describe a set of strategies that have been evaluated in mitigating this issue.

### 5.3.1 Site selection and MAC-brand based filtering

One of the easiest means of ensuring a particular population sample when using static sensors is choosing a location that can only have the population of interest in range. For example, during the Montreal test described in Section 4.2.3, one of the sensors was mounted in a pedestrian-only alley. This is useful not only for dwell time analysis of pedestrians at that location, but also at other locations – the known-to-be-pedestrian MACs from the alley can used be to calibrate travel-time thresholds at the other locations. This same filtering concept can be applied to the information contained in the MAC addresses themselves – as described in Section 4.2.2, the first three hex digits of a given MAC are indicative of the device manufacturer. Because some manufacturers, such as Parrot, only manufacture in-vehicle systems, the mode of

some of the MAC addresses detected can be known, assuming sufficient sample sizes. This knowledge can then be used to set appropriate threshold values for gap time and dwell time parameters used in the outlier filtering techniques discussed next.

### 5.3.2 Outliers and Filtering

#### *5.3.2.1 Outlier Sources*

There are numerous potential sources for outliers within the travel time data. Perhaps the most apparent cause on a given corridor is drivers that stop on their way through the corridor, or choose a route that is much longer than most users. This creates a delay that is not experienced by other users, thus resulting in an outlier. Since the additional delay is unlikely a factor of the network design or any other transportation considerations, it is often not of interest in basic corridor-level evaluations. This type of outlier is often easy to recognize and is present in both license plate reader and MAC address matched travel time data. These outliers can cause travel times that are not representative of the general pattern (assuming one is known or expected). Multiple modes present on the same corridor can also be a cause for outliers when one is looking at auto-only travel times. The discrimination step occurs during the filtering of the MAC data. Procedures used to screen and filter data obtained from MAC address readers are described in the following section.

#### *5.3.2.2 Data filtering*

Travel time filtering is accomplished by the MAC matching and filtering engine described briefly in Section 5.2.1. The most basic filtering technique is thresholding, and threshold parameters for both travel time and dwell time have been incorporated into the analysis tool. Some MAC address data collected appears to have a definitive threshold – for example, Figure 5-5 shows data collected during the campus experiment described in Section 4.4.3. There is a gap between the data that have travel times of less than 500 seconds and those having more. Upon closer examination, those data that have travel times less than 500 seconds follow a clear linear trend, with a slope roughly equivalent to walking speed. Looking at the histogram of the speeds of the objects, it can be seen that it is roughly normally distributed with a mean of 1.6 m/s. Thus, there is evidence that 500 seconds is a reasonable threshold. The values above this threshold represent static devices that are not of interest in travel time analysis (however, may be of sole interest in dwell time analysis).

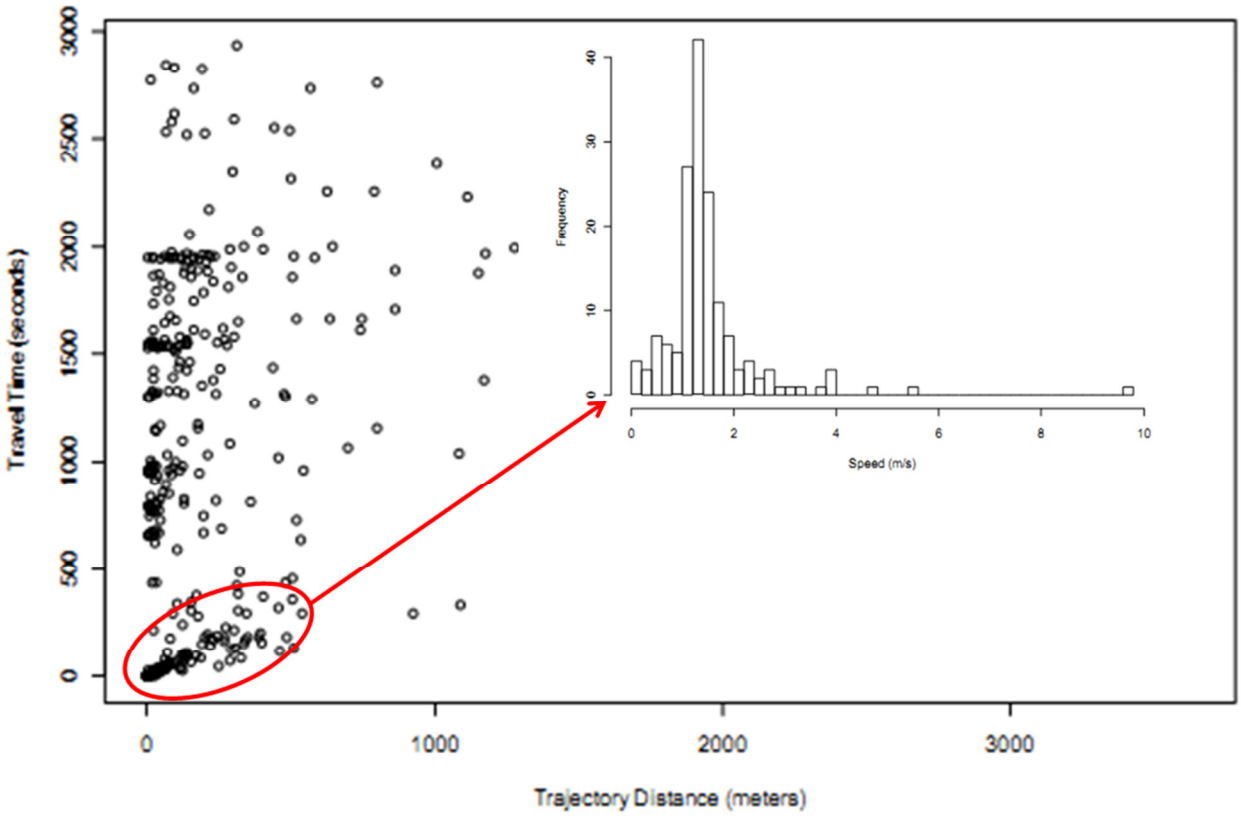



**Figure 5-5: Isolating pedestrian mode from MAC data**

However, threshold is not always that apparent. In addition to hard thresholds on travel and dwell time, the software allows for a moving median analysis to be used on the data. A moving median filter, based on the one used by Quayle et al. (2010) is used. A standard deviation calculation based on a sliding time window is used to filter the results:

$$\sigma = \sqrt{\frac{1}{t}\sum_{i=x-\frac{t}{2}}^{t}(p_x - \mu)^2} \qquad \textbf{(Equation 5-1)}$$

where t is time window, $p_x$is the travel time at timestamp x and $\mu$ is the mean calculated for the time window **t**. If a particular travel time measurement was within one standard deviation above the localized mean, it was accepted as a valid data point. This allows for effective filtering of some of the more predictable (corridor) MAC data without prior behavior knowledge or extensive data exploration.

## 5.4 Reducing Temporal Uncertainty

Temporal uncertainty in MAC address data application occurs due to the characteristics of the Bluetooth protocol (up to 10.24 seconds connection time), the signal noise that is present within the detection area and the range captured by the sensor's antenna(e). The following sections seek to establish a relationship between these factors and better predict their effect on travel time: a vital transportation metric. A single corridor, instrumented with license plate readers that acts as ground truth data sources for travel time is considered in the evaluation.

### 5.4.1 Study Site

A 0.98 mile section of SR-522 (Bothell Way NE), shown in Figure 5-6, was selected for this study. The section is located on the northwest section of Lake Washington in Washington State. This corridor is ideal due to the availability of ALPR data along the corridor, minimal pedestrian and cyclist presence and a high volume of over 50,000 vehicles per day (Mizuta,

2007). The section starts at NE 170th Street in the City of Lake Forest Park and ends at 61st Ave NE in the City of Kenmore. The short length of the corridor emphasizes the need for error analysis and mitigation – the Bluetooth device range, especially for stronger antennae can contribute to the travel time error encountered, as most travel times within the corridor are less than 2 minutes.


**Figure 5-6: Study route on SR-522 [Image from maps.google.com]**

Spectrum data was collected for this experiment to ensure that there was not a significant source of background noise that would severely impact detection quality. Since the Bluetooth protocol uses spread-spectrum frequency hopping, the device skips from frequency to frequency, thus largely not impacted by local sources that may be operating within a narrow band of the

2.399 MHz to 2.483 MHz spectrum, used by both protocols. However, additional Wireless Local Area Networks (WLAN) located at the same location could impact the detection performance by occupying large portions of the spectrum and rendering it unusable. Since WLAN networks have only 11 different channels, each of which occupies 22 of the 79 available Bluetooth channels (Hewlett Packard, 2002), the presence of multiple WLANs in the area could reduce performance if the signal strengths of those networks is sufficient. It is important to ensure that the test sites chosen do not contain significant contamination of the 2.4 GHz spectrum.

Figure 5-7 below shows the spectrum characteristics at the 170th ST NE site. Each point on the graph represents a one-hour average along a 327 KHz strip of the spectrum, for a total of 256 strips. The location does appear to have several active networks that occupy some bands, but the signatures are narrow, thus creating little competition for Bluetooth devices. More importantly, the magnitude of the detected networks is very small, with the highest peaks reaching well under -100 dBm. Signals below -100 dBm are considered to be out of range for the directional and omni-directional antennae, thus having little impact on the detection speed.

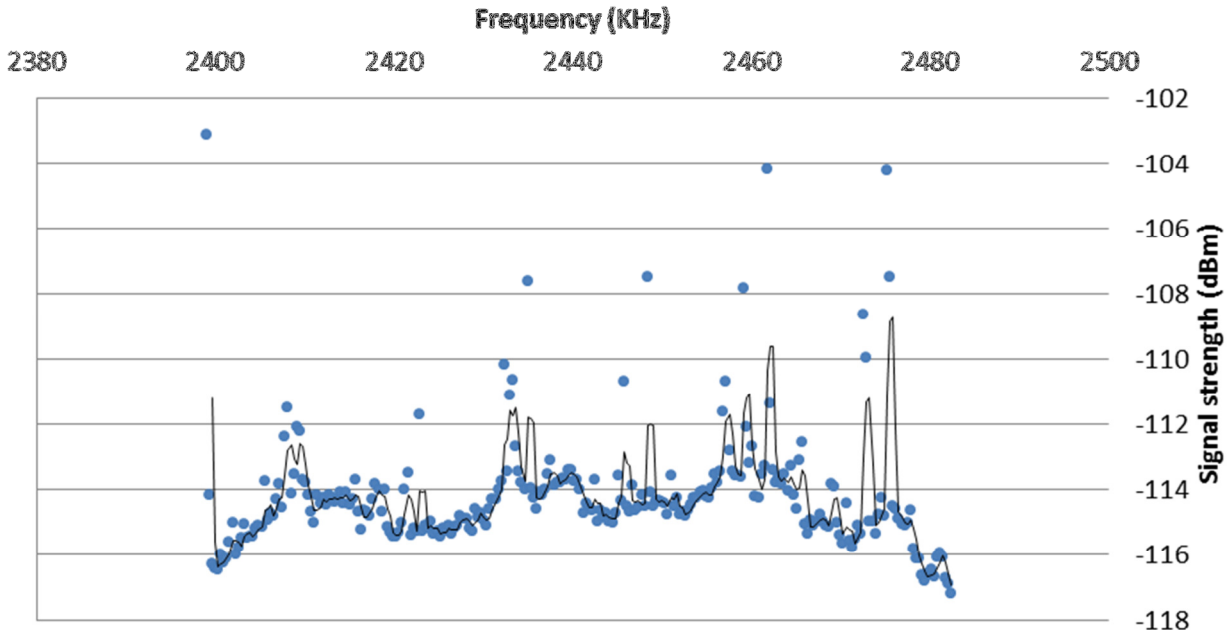



**Figure 5-7: Spectrum average for 170th St NE.**

Figure 5-8 shows a similar diagram for the NE 61st Ave site. The signature at this location

is slightly different, as there appears to be two WLANs present, show on the right side as the wide peaks. However, the signal strengths are still too weak to cause any significant interference to the Bluetooth detectors.

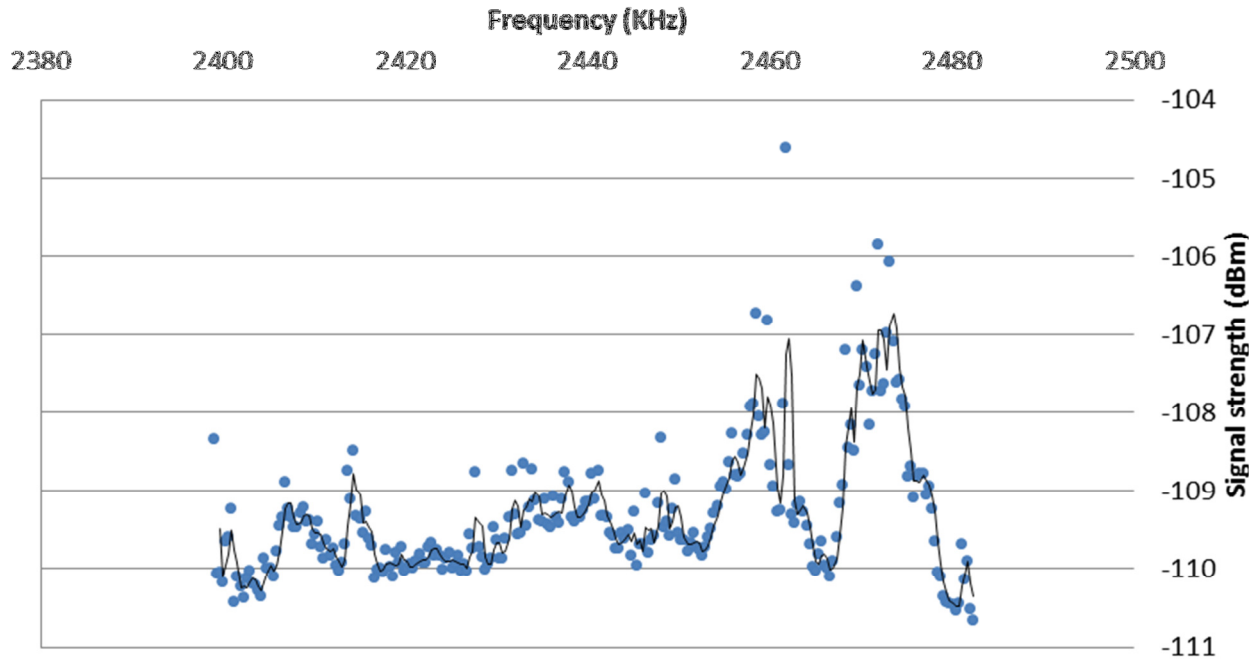



**Figure 5-8: Spectrum average for NE 61st Ave.**

To determine the effects of mounting two MACAD devices adjacently and operating them concurrently a short test was done to see the number of "collisions" that the devices would experience. Figure 5-9 below demonstrates the overall noise levels that are present when one device is scanning vs. when two devices are scanning. The graph shows the full 2.399 to 2.483 spectrum on the x-axis and time on the y-axis. Green areas represent "clear" sections of the

spectrum where signal was strong. Yellow represent sections with some interference and the red sections represent moments when there is strong interference, indicating that another device was also using the spectrum. The testing was done at the 170th St site. Based on the resulting images, it is difficult to say that an additional Bluetooth device has a significant effect on the number of collision experienced by one device. The amount of red and yellow areas remained roughly the same, indicating that the additional device was unnoticeable among the noise. Both a) and b) of Figure 5-9 below contain about 68% red and yellow sections.

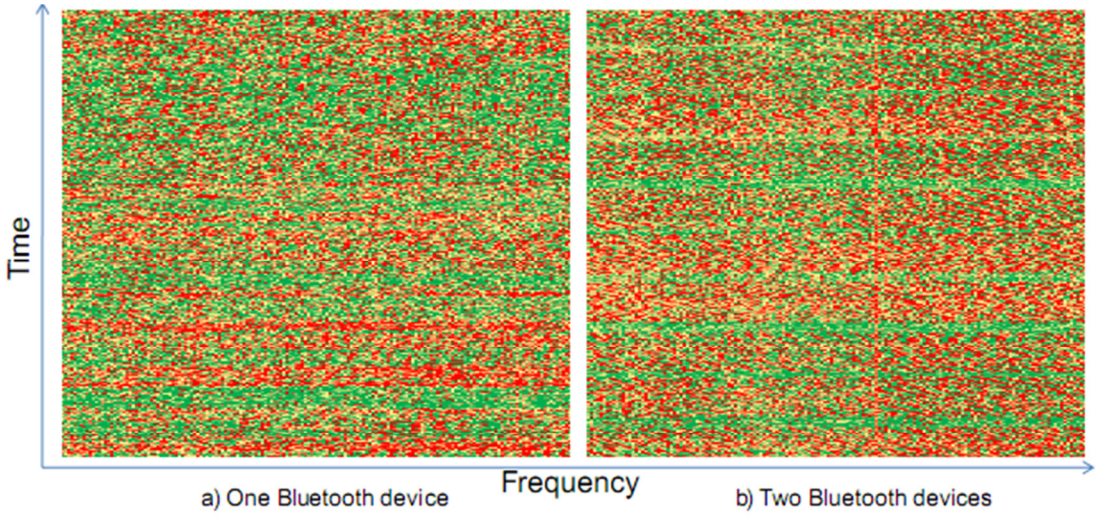



**Figure 5-9: Spectrum noise image**

### 5.4.2 MAC Address Data Acquisition

Up to four MACAD devices were used to collect travel time data, using a combination of antennae types and strengths and on-site placement positions. Table 5-1 shows the variables considered in this study. Three types of antennae were used in testing, a 7 dBi weatherproof

omni-directional antenna, a 9 dBi weatherproof omni-directional antenna and a 12 dBi directional, 35 degree vertical and horizontal spread antenna mounted in the lid of an MACAD device. These are denoted as “O7”, “O9” and “D12” in Table 5-1. The number of detectors at each location, up to two, is also considered as a variable. Finally, when two detectors were mounted at the same end of the corridor, they were either mounted one across from another (opposite), denoted as “O” or at the exact same location, as denoted by “S”. If only one sensor was mounted, “S” is used to indicate no overlap. All of the eleven different configurations were tested and are summarized in Table 5-1.

The lane-ft covered variable represents the cumulative linear feet covered by the sensor configuration. These values are estimations based on manufacturer specifications and empty-field range testing. The values were computed by overlaying the approximate sensor ranges over the test site and measuring the lengths of the through lanes covered by the sensors. Figure 5-10 shows the lane-ft covered by the 12 dBi directional sensor at the NE 170$^{th}$ St location. The clover-like shape represents the 12 dBi directional antenna bloom as specified by the manufacturer.

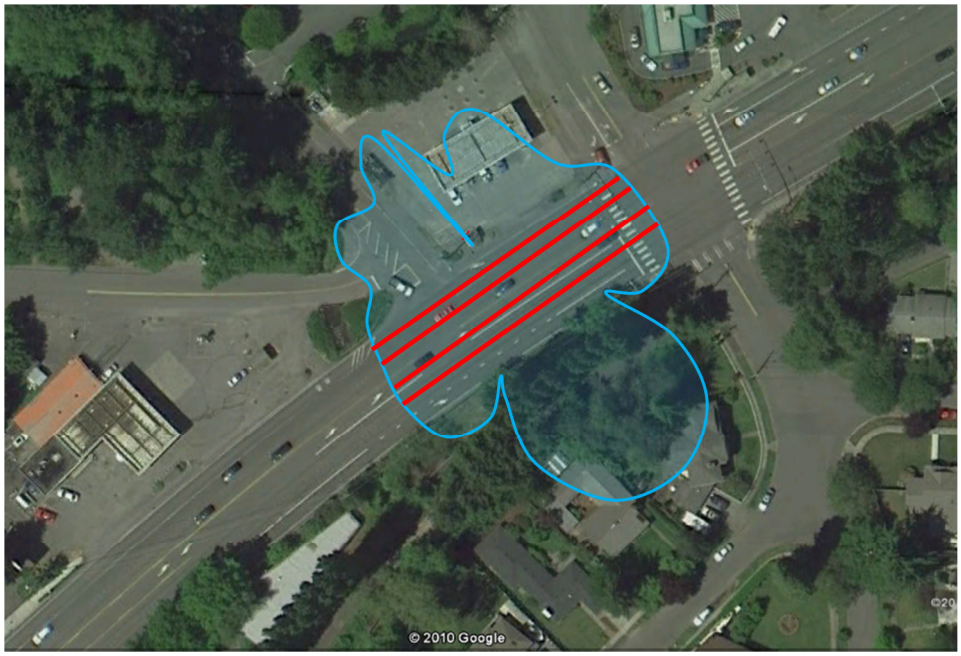



**Figure 5-10: Lane-ft coverage of a 12 dBi directional sensor at NE 170th St**

**Table 5-1: Bluetooth device mounting and antenna configurations**

| Configuration (1-11) | Mounting Location Opposite (O) Same(S) | Number of Detectors (1) (2) | Site Location 170th & SR522 (A) 61st & SR522 (B) | Antenna Type Omni 7dBi (O7) Omni 9 dBi (O9) Direc. 12 dBi (D12) | Lane-ft Covered Total (all sensors) |
|---|---|---|---|---|---|
| 1 | S | 1 | A | O7 | 445 |
| | S | 1 | B | O9 | 917.5 |
| 2 | S | 1 | A | O9 | 917.5 |
| | S | 1 | B | O7 | 445 |
| 3 | S | 1 | A | O9 | 917.5 |
| | S | 1 | B | O9 | 917.5 |
| 4 | S | 2 | A | O7, O9 | 1365.5 |
| | S | 2 | B | O7, O9 | 1365.5 |
| 5 | S | 2 | A | O7, D12 | 855 |
| | S | 2 | B | O7, D12 | 855 |
| 6 | S | 2 | A | O9, D12 | 1327.5 |
| | S | 2 | B | O9, D12 | 1327.5 |
| 7 | O | 2 | A | O9, D12 | 1367.5 |
| | O | 2 | B | O9, D12 | 1367.5 |
| 8 | O | 2 | A | O7, D12 | 852.5 |
| | O | 2 | B | O7, D12 | 852.5 |
| 9 | O | 2 | A | O7, O9 | 1402.5 |
| | O | 2 | B | O7, O9 | 1402.5 |
| 10 | S | 1 | A | D12 | 410 |
| | S | 1 | B | D12 | 410 |
| 11 | S | 1 | A | O7 | 445 |
| | S | 1 | B | O7 | 445 |

### 5.4.3 Test Configurations

Detectors were conveniently mounted at a height of about 1.5 meters (5 ft) above the

ground on roadside signage poles. Directional sensors were pointed across the roadway, near the westbound side of the route, as close as possible to the westbound ALPR detection zones. Figure 5-11 shows all of the possible sensor footprints that were tested in this study and their approximate detection zones. Directional footprints are shown with solid lines and omni-directional footprints are shown with dashed lines. Bluetooth sensor locations are marked with an "x" and ALPR detection zones are shown as rectangles. These footprints were permutated through 11 different configurations that represent the potential variability of setups, bearing in mind the locations of the ALPR sensors. The directional antennae, for example, were only mounted near the ALPR detection zones as other placements were unlikely to produce better results. The westbound side provided convenient mounting locations for numerous sensors and was thus chosen as the primary focus of this study. The estimated ranges for the 7dBi, 9dBi omni-directional and 12dBi directional antennae are 40 meters (131 ft), 70 meters (230 ft) and 40 meters (131 ft), respectively. These sensors were configured to try and match the westbound ALPR detection zones as closely as possible. Eastbound travel times picked by these sensors are likely to be more different from their ALPR counterparts as they are separated by an intersection. This is clearly shown in the collected data and the results are presented separately.

Permutations with identical setups at each of the two locations (NE 170$^{th}$ St and 61$^{st}$ AVE NE) were primarily tested, but two configurations (1 and 2) with disparate antenna strengths were tested as well. Each antenna type was tested standalone as well as in tandem with another antenna type. During tandem tests for configurations 5-9, data for configurations 10 and 11 was

extracted by looking at only one sensor set (while ignoring data from the other two). Since the interference between two devices was measured to be minimal, the impact of doing the two tests at once was considered negligible, while providing useful insights into the additional accuracy afforded by the extra devices.

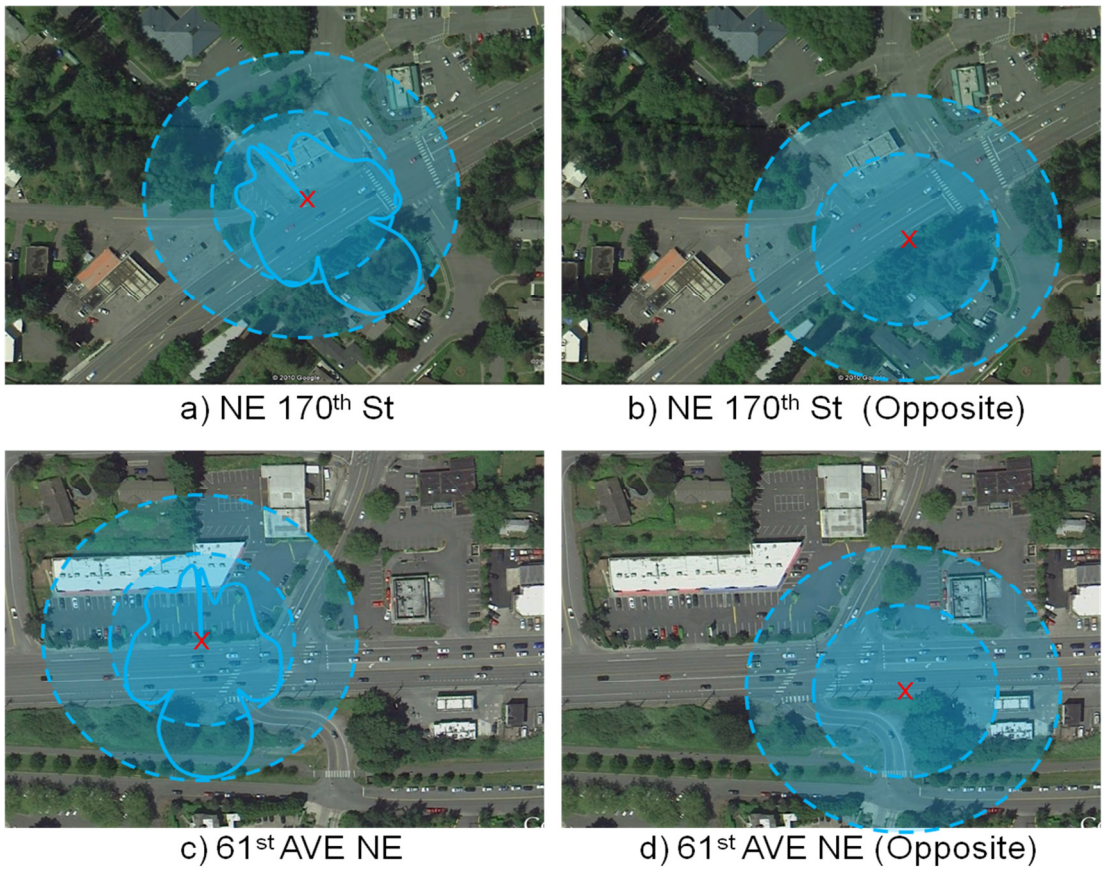



**Figure 5-11: Sensor configurations [Background images from maps.google.com]**

### 5.4.4 License Plate Data Acquisition

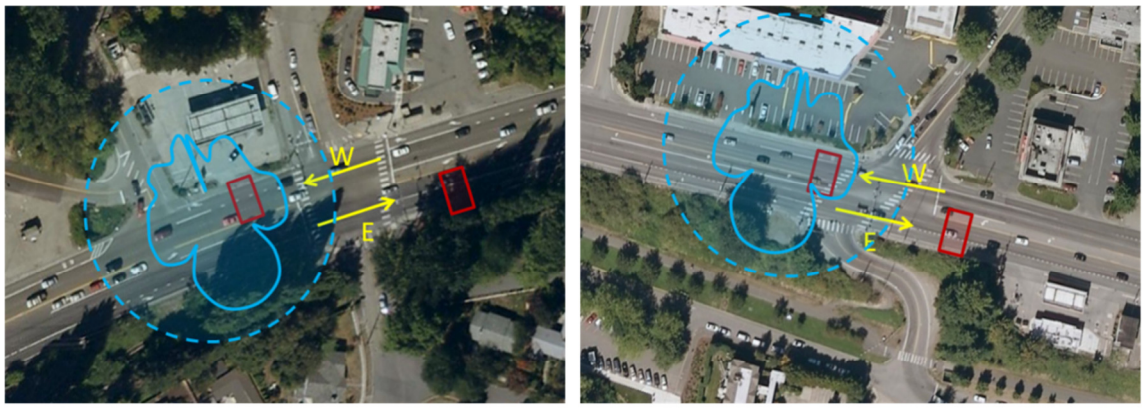
The examined section of SR522 has a speed limit of 45 mph and is a six-lane arterial with four inside general purpose lanes and two transit-only outside lanes. ALPR readers are installed on the arms of the intersection signal heads to read license plates from the rear of passing vehicles. All the westbound ALPR readers were designed to read the vehicles traveling in the inside lane (closest to the median). All the eastbound readers were designed to read the vehicles traveling in outside general purpose lane (Mizuta, 2007). ALPR data is reported in aggregated 5-minute averages in the eastbound and westbound directions. ALPR capture rates are also reported upstream and downstream and are used as surrogates for volume data. Details of the installed systems can be found in (PIPS, 2009).

### 5.4.5 Experiment Results

Due to the misalignment between the eastbound ALPR detection zones and the MACAD detection areas, the results for each direction are presented separately. As will be shown in sections 5.4.6 and 5.4.7, the westbound measurements were more accurate than the eastbound ones. This is due to the eastbound ALPR detection zones not correlating well with the antenna footprints. Figure 5-12 shows the approximate relative position of the detection zones and footprints. Last-to-last matching, or using the last available timestamp for each bypassing MAC for matching, was used to obtain the travel times on SR-522. This was done in order to minimize the effect of intersection delay on the results, as the timestamp is taken after the vehicles leave

the intersection, regardless of direction of travel. This is also closest to how the ALPR sensors are collecting data – the license plate is read right after the vehicle passes the intersection. Although this approach demonstrated better results than first-first or median matching, it was still insufficient to completely circumvent the problem, as the last timestamp may still occur within the intersection area due to noise and signal blockage issues.

The combinations of mountings, antennae strengths and sensor quantities were tested during the week of July 19th-27th, 2010. The tests were stopped for a break on the afternoon of July 20th to the evening of the 21st, when the ALPR units were switched off for maintenance.




a) NE 170th St b) 61st AVE NE

**Figure 5-12: Sensor detection zones**

### 5.4.6 Error Analysis Westbound

#### *5.4.6.1 Descriptive Analysis Westbound Direction*

Figure 5-13 shows the 1-hr average travel time results in the westbound direction. Red points and lines are Bluetooth (BT) travel times while blue ones are ALPR travel times. The testing intervals for each configuration are labeled – configurations 10 and 11 run in parallel with 5-9. To differentiate them from other configurations their results are shown in orange. Trend lines are generates using a 5-point moving average window. Overall, the sensors follow the travel time trends recorded by the ALPRs. It can be seen that tandem sensor configurations do a better job of following the trends.

Figure 5-14 demonstrates the 1-hr averages of travel time error rates compared to the volumes encountered during testing in the westbound direction. Total volume in both directions is shown in blue and error in red. The graph is once again segmented into the testing configurations and error rates for configurations 10 and 11 are shows separately in orange. Trend lines were generated using a 5-point rolling average. The results imply that although there is some correlation of travel time error with volume, some configurations are less affected by this than others.

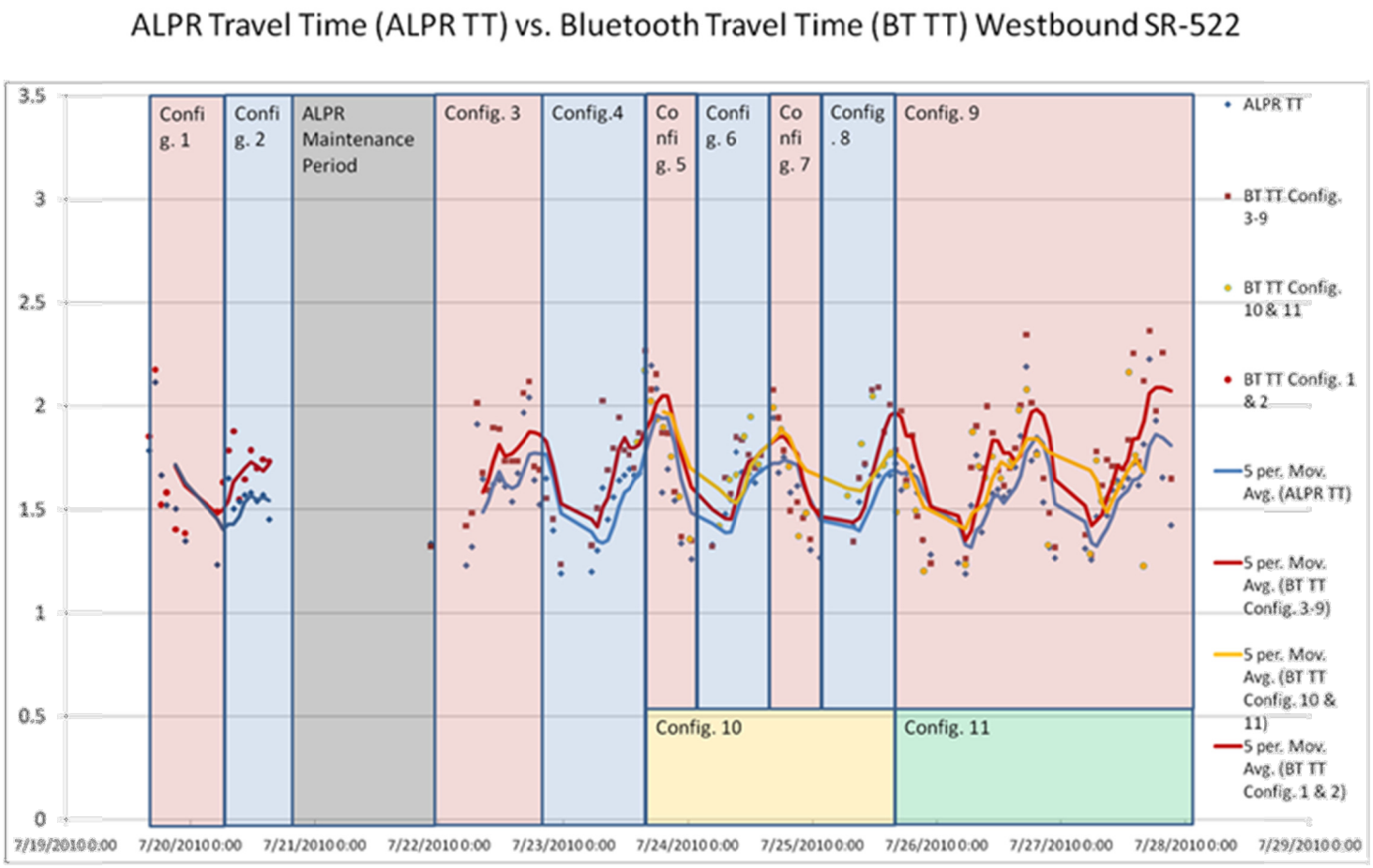

ALPR Travel Time (ALPR TT) vs. Bluetooth Travel Time (BT TT) Westbound SR-522
Confi g. 1
Confi g. 2
ALPR Maintenance Period
Config. 3
Config.4
Co nfi g. 5
Confi g. 6
Co nfi g. 7
Config . 8
Config. 9
Config. 10
Config. 11
3.5
3
2.5
2
1.5
1
0.5
0
7/19/2010 0:00
7/20/2010 0:00
7/21/2010 0:00
7/22/2010 0:00
7/23/2010 0:00
7/24/2010 0:00
7/25/2010 0:00
7/26/2010 0:00
7/27/2010 0:00
7/28/2010 0:00
7/29/2010 0:00
ALPR TT
BT TT Config. 3-9
BT TT Config. 10 & 11
BT TT Config. 1 & 2
5 per. Mov. Avg. (ALPR TT)
5 per. Mov. Avg. (BT TT Config. 3-9)
5 per. Mov. Avg. (BT TT Config. 10 & 11)
5 per. Mov. Avg. (BT TT Config. 1 & 2)

**Figure 5-13: Travel time comparison westbound SR-522 (ALPR – blue, BT – red + orange) (1hr averages)**

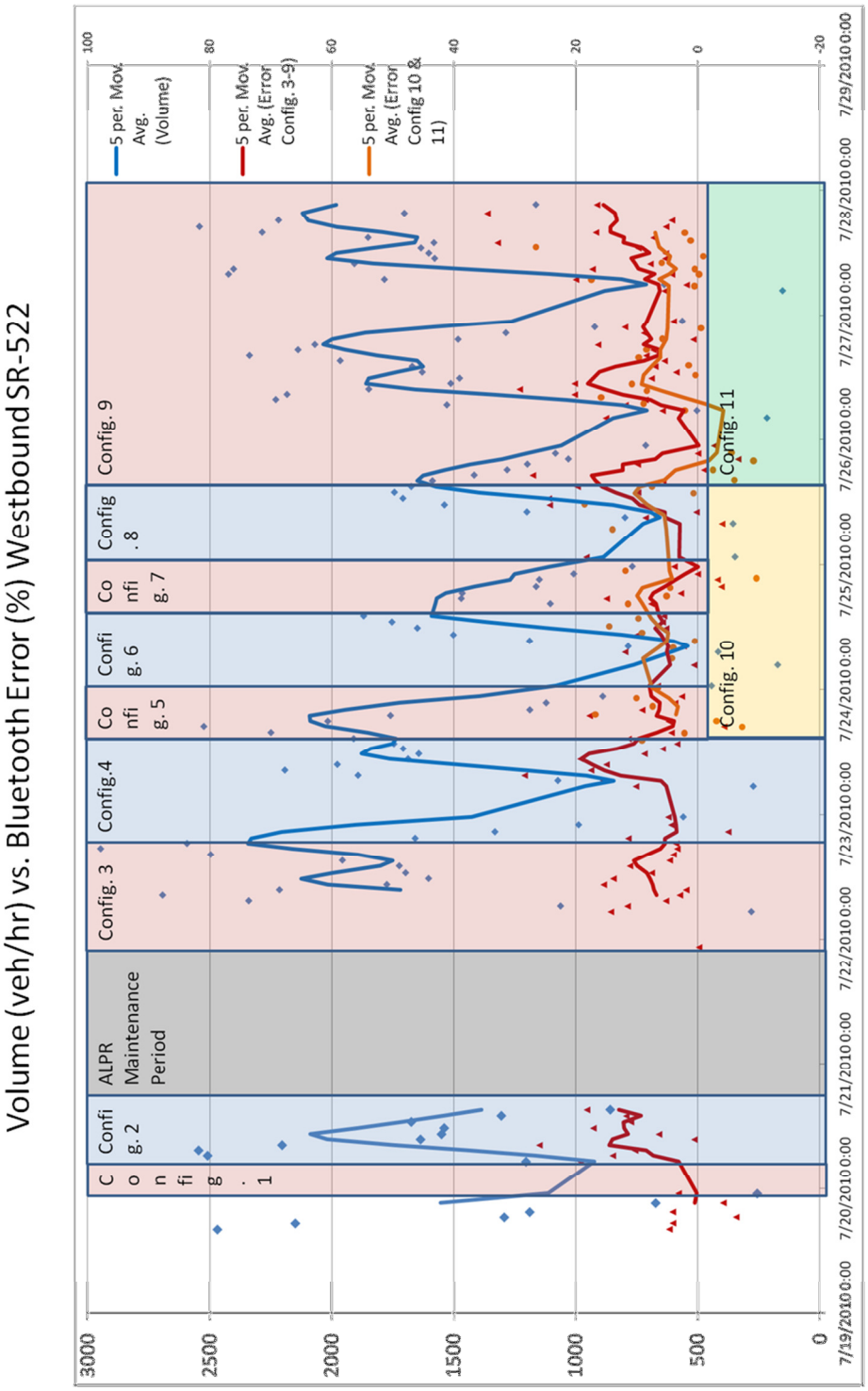

Volume (veh/hr) vs. Bluetooth Error (%) Westbound SR-522
5 per. Mov. Avg. (Volume)
5 per. Mov. Avg. (Error Config. 3-9)
5 per. Mov. Avg. (Error Config 10 & 11)
Config. 1
Config. 2
ALPR Maintenance Period
Config. 3
Config.4
Config. 5
Config. 6
Config. 7
Config. 8
Config. 9
Config. 10
Config. 11
3000
2500
2000
1500
1000
500
0
100
80
60
40
20
0
-20
7/19/2010 0:00
7/20/2010 0:00
7/21/2010 0:00
7/22/2010 0:00
7/23/2010 0:00
7/24/2010 0:00
7/25/2010 0:00
7/26/2010 0:00
7/27/2010 0:00
7/28/2010 0:00
7/29/2010 0:00

**Figure 5-14: Westbound SR-522 error and volume (1hr averages)**

Taking a closer look at the westbound data, it can be seen that configuration 5, 6, 7 and 8 appear to be more accurate overall. These configurations contain a directional antenna that successfully discriminates the vehicles waiting at the intersection approach, outside its narrower range. Single sensor layouts also appear to have a lower error. This is expected, as the smaller overall footprint reduces error, which is especially true in the westbound direction, since the MACAD directional detection beam is focused right over the ALPR detection point. This smaller footprint however reduces the total available matches, thus reducing the accuracy of more precise 15-min intervals examined in the next section.

#### *5.4.6.2 Error Modeling Westbound*

Initial efforts in interpreting the data focused on modeling the detection rate and relating that to the accuracy of the acquired travel times. A generic approach to error modeling was taken, considering all possible variables and their relationship with accuracy. A multivariate regression model was developed for each direction to determine which variables are significant. A 15 minute time window was chosen as an analysis element to show variation in traffic patterns while minimizing the effect of contamination by signal delay. All variables were aggregated to 15 minute intervals. Ten variables were considered in all:

(1) Volume (Categorical: <500[LOW], <1000[MED], >1000[HIGH])

(2) Detection Rate (Percentage of Volume)

(3) Matching Rate (Percentage of Volume)

(4) Lane-ft Covered by All Sensors in Configuration

(5) Directional Antenna (Categorical: 0 [no],1 [yes])

(6) Opposite Side Tandem Sensors (Categorical: 0,1)

(7) Sensor 1 Antenna Strength (Categorical [dBi]: 7,9,12)

(8) Sensor 2 Antenna Strength (Categorical [dBi]: 7,9)

(9) Sensor 3 Antenna Strength (Categorical[dBi]: 7,9)

(10) Sensor 4 Antenna Strength (Categorical[dBi]: 7,9,12)

A generic model was first attempted using all variables:

$$E_k = \beta_0 + \beta_1 V + \beta_2 D + \beta_3 M + \beta_4 L + \beta_5 R + \beta_6 O + \beta_7 S_1 + \beta_8 S_2 + \beta_9 S_3 + \beta_{10} S_4 + \epsilon_k$$

(**Equation 5-2)**

where $E_k$ is the absolute error in fractional minutes, $V$ is the volume in veh/hr, $D$ is the detection rate in percent, $M$ is the matching rate in percent, $L$ is the sensor lane-ft coverage, $R$ is the directional variable, $O$ is the opposite side variable and $S_{1\text{-}4}$ are antenna strengths of sensors in dBi 1-4. $\epsilon_k$ is the regression error term. The resulting model for the westbound direction and their variables, with relative significance levels is presented in Table 5-2 (only significant variables shown).

**Table 5-2: Westbound error regression model results**

WEST

Coefficients:

| | Coefficient | Std. Error | t value | Pr(>\|t\|) | Significance Level |
|---|---|---|---|---|---|
| (Intercept) | 0.2902 | 0.0128 | 22.6430 | 0.0000 | .001 |
| Volume LOW | -0.0598 | 0.0134 | -4.4660 | 0.0000 | .001 |
| Volume MED | 0.0382 | 0.0091 | 4.1960 | 0.0000 | .001 |
| Detection Rate | 0.0050 | 0.0013 | 3.8620 | 0.0001 | .001 |
| Match Rate | -0.0098 | 0.0019 | -5.2190 | 0.0000 | .001 |
| Linear Coverage | 0.0000 | 0.0000 | -2.3500 | 0.0191 | .05 |
| Opposite | 0.0330 | 0.0112 | 2.9380 | 0.0034 | .01 |

Adj. $R^2 = .2101$

The resulting model confirms some of the anticipated concerns regarding volume, with lower volumes resulting in more accurate travel times. This can be attributed to the fact that the lower volumes accumulate less signal delay, as vehicles do not back up or wait as long on approaches. Medium volumes increased error in the westbound direction, implying that volumes over 500 veh/hr resulted in additional intersection delays that were passed on to the MACAD system. Higher detection rates were shown to increase the error. This is also expected because, under the same volume level, a higher detection rate is typically associated with a larger detection zone and a larger detection zone will lead to a larger spatial error. Matching rates had a negative correlation, implying that improving matching rates will reduce error by providing a

larger sample size. Linear coverage played a similar role to detection, larger zones contributed to the error. Opposite-side tandem mounting was found to have an increasing effect on error in the westbound direction. This may be caused by the fact that the opposing side sensor at 61st St NE was mounted close to the eastbound ALPR detection zone, which allowed it to capture westbound vehicles waiting at the light. The NE 170th location was configured to avoid this issue.

### 5.4.7 Error Analysis Eastbound

#### *5.4.7.1 Descriptive Analysis Eastbound Direction*

The eastbound side of the test bed shows greater variations and errors. In Figure 5-15, the single sensor configuration (shown in orange) is notably farther from the ALPR trend than the tandem configuration data obtained concurrently.

As can be seen in Figures 5-16, there is a greater effect of volume on the accuracy of the Bluetooth MAC address readers due to the signal delay. Eastbound travel times are affected much more than westbound ones, as most of the configuration's mountings have the detection zone centered near the eastbound signal approaches. This results in more reads near the approach areas and progressively less as the vehicle leaves the detection zone. This skews the results towards reflecting the intersection delay.

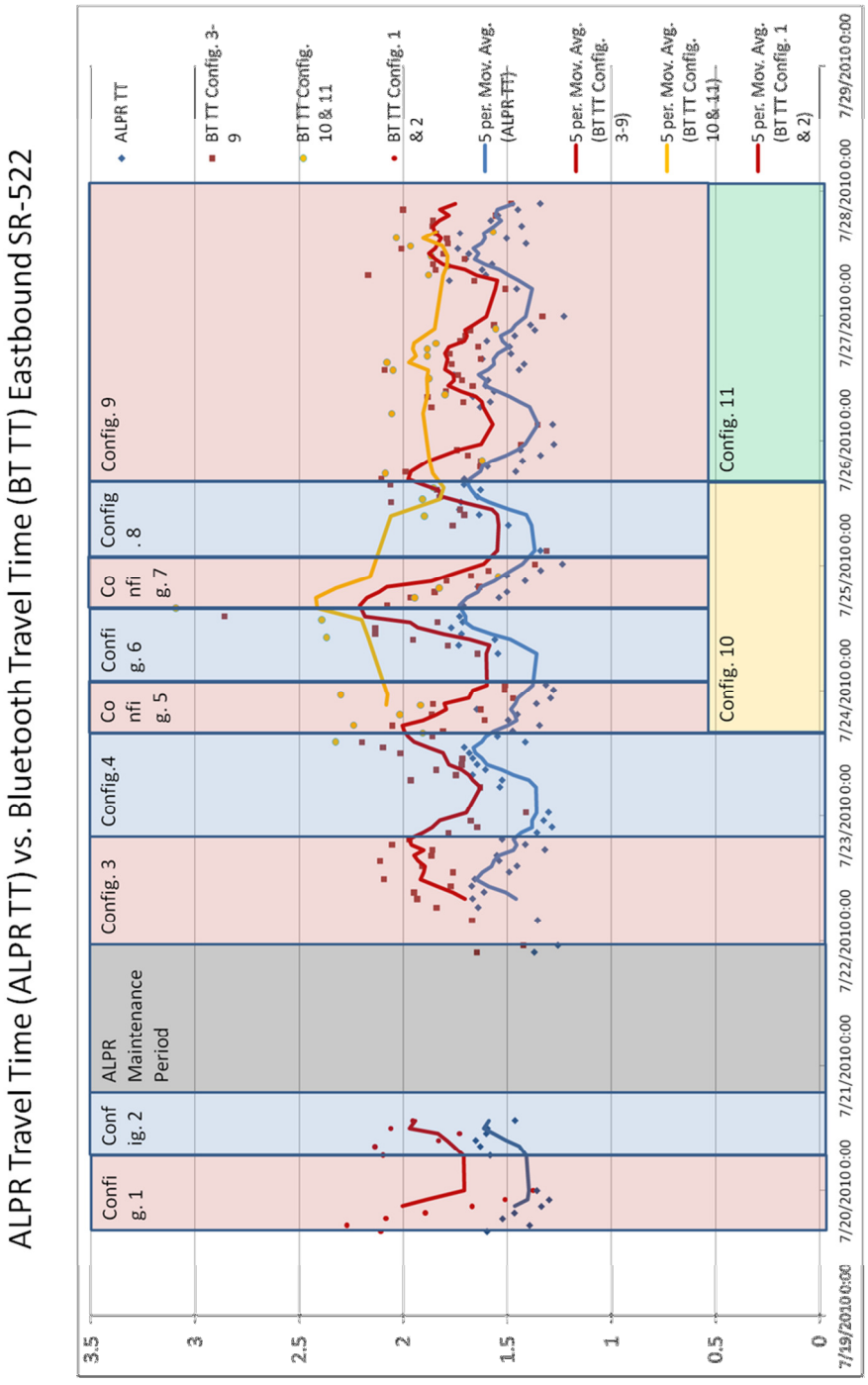



**Figure 5-15: Travel time comparison eastbound SR-522 (ALPR – blue, BT – red + orange) (1hr averages)**

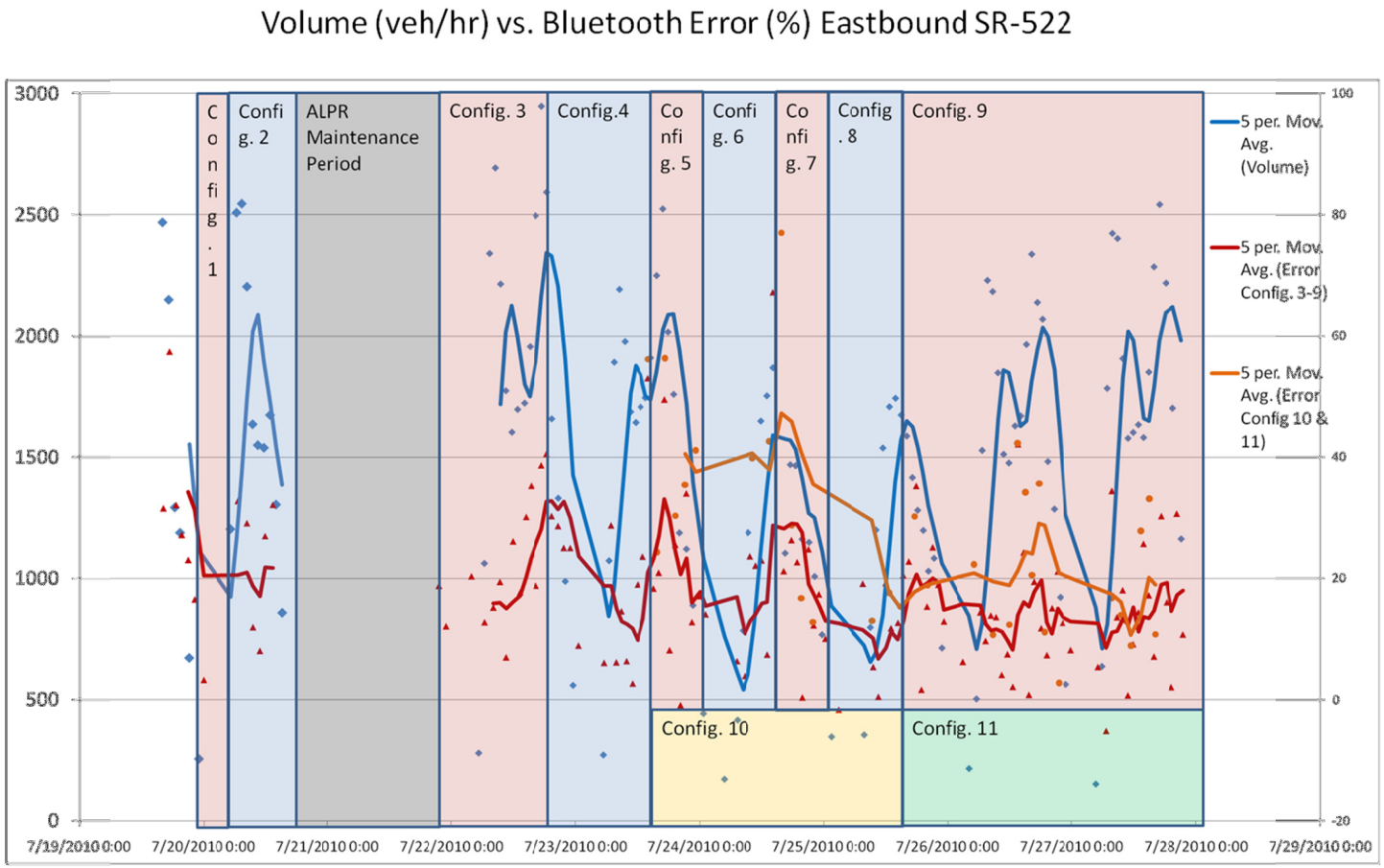

Volume (veh/hr) vs. Bluetooth Error (%) Eastbound SR-522
3000
2500
2000
1500
1000
500
0
100
80
60
40
20
0
-20
Config. 1
Config. 2
ALPR Maintenance Period
Config. 3
Config.4
Config. 5
Config. 6
Config. 7
Config. 8
Config. 9
Config. 10
Config. 11
5 per. Mov. Avg. (Volume)
5 per. Mov. Avg. (Error Config. 3-9)
5 per. Mov. Avg. (Error Config 10 & 11)
7/19/2010 0:00
7/20/2010 0:00
7/21/2010 0:00
7/22/2010 0:00
7/23/2010 0:00
7/24/2010 0:00
7/25/2010 0:00
7/26/2010 0:00
7/27/2010 0:00
7/28/2010 0:00
7/29/2010 0:00

**Figure 5-16: Eastbound SR-522 error and volume (1hr averages)**

### *5.4.7.2 Error Modeling Eastbound*

An eastbound model was developed using the same approach and the same initial set of variables as the westbound direction. However, the resulting set of significant variables turned out to be slightly different, with more variables playing a role. Since the relationship between the ALPR zones and MACAD zones was more complex, this is to be expected. Volume, detection rate, match rate and linear coverage still play a significant role however. Table 5-3 shows the regression model for the eastbound direction (once again, only significant variables shown).

**Table 5-3: Eastbound error regression model results**

EAST

Coefficients:

| | Coefficient | Std. Error | t value | Pr(>\|t\|) | Significance Level |
|---|---|---|---|---|---|
| (Intercept) | 0.3495 | 0.0452 | 7.7360 | 0.0000 | .001 |
| Volume LOW | -0.2328 | 0.0288 | -8.0980 | 0.0000 | .001 |
| Volume MED | -0.0844 | 0.0235 | -3.5870 | 0.0004 | .001 |
| Detection Rate | 0.0229 | 0.0034 | 6.8300 | 0.0000 | .001 |
| Match Rate | -0.0100 | 0.0026 | -3.8920 | 0.0001 | .001 |
| Linear Coverage | 0.0001 | 0.0000 | 2.2840 | 0.0227 | .05 |
| Directional | 0.1663 | 0.0270 | 6.1710 | 0.0000 | .001 |
| Antenna 2 Strength 7 dBi | -0.2823 | 0.0390 | -7.2460 | 0.0000 | .001 |
| Antenna 2 Strength 9 dBi | -0.3454 | 0.0612 | -5.6450 | 0.0000 | .001 |

Adj. R^2 = .2669

For the eastbound direction, directional antennae and antenna strength was found to have an increasing effect on error. Since the directional antennae were focused on the westbound side ALPRs, an increase in error is to be expected due to misalignment. Reduced error due to antennae strength (the stronger the lower the error – 7 dBi has less of a decreasing effect than 9 dBi) at sensor 2 (NE 170$^{th}$ St) can interpreted as creating a larger sample size. The eastbound direction was further from the mounted sensors away for most configurations – in such cases, antenna strength makes more of a difference, as smaller antennae have a harder time collecting samples.

It is worth noting that detection rate was not shown to be significant in either direction. This was somewhat unexpected, and discouraged the use of the initial detection-based model outlined in the proposal. There may be a couple explanations for this occurrence. First, there may have been too much noise from non-vehicular sources that increased the detection rate without providing subsequent matches. Second, the diversion rates for the corridor may have been too high, once again resulting in detections without matches. Discussion of detection and match rates for each configuration is presented in the following section.

### 5.4.8 Configuration Comparison

Further insights into the performance of the MACAD devices can be gleamed from

comparing the different configurations tested. In doing so, one can determine the most successful setup that was capable of providing the most accurate results, despite of the additional issues caused by the signal delay. A discussion of the performance of each configuration is given in the following section, once again separated by direction. While examining the data, it is imperative to recall that the tested corridor is less than 1-mile long, which results in the largest footprints taking up nearly 20% of the corridor.

#### *5.4.8.1 Westbound*

Table 5-4 presents a basic comparison of the tested configurations based on error statistics – average error, standard deviation of error, and min and max error in terms of minutes. The statistics are computed on 15-min intervals. Of the configurations tested, configuration 6 (9 dBi omni and 12 dBi directional) appears to have some of the best results, with a low average error and the lowest deviation in both the westbound and eastbound directions. Configuration 1 (a mix of 7 and 9 dBi antennae as singles) also fares well with the lowest absolute error, low standard deviation and a low maximum error. It can be seen from Table 5-4 that the absolute value of the max error is higher than the absolute value of the minimum error, supporting a case for positive bias.

**Table 5-4: Westbound 15-minute aggregate error statistics by configuration**

| Westbound Config. | Abs. Error (sec) | Std. Dev (sec) | Max Error (sec) | Min Error (sec) |
|---|---|---|---|---|
| 1 | 2.56 | 5.73 | 12.17 | -8.58 |
| 2 | 10.94 | 7.35 | 25.25 | -2.33 |
| 3 | 7.58 | 6.09 | 20.92 | -6.92 |
| 4 | 8.95 | 7.33 | 25.25 | -5.75 |
| 5 | 6.10 | 9.72 | 33.42 | -13.42 |
| 6 | 6.13 | 4.38 | 16.67 | -0.42 |
| 7 | 3.64 | 8.19 | 19.33 | -8.17 |
| 8 | 11.31 | 10.83 | 39.00 | -4.58 |
| 9 | 9.67 | 8.02 | 36.25 | -6.83 |
| 10 | 6.08 | 7.78 | 22.25 | -14.58 |
| 11 | 3.82 | 8.94 | 37.50 | -11.00 |
| Average TT: 91.8 sec | | | | |

Figure 5-17 shows the detection and matching rates for each configuration in the westbound directions. The matching and detection rates proved to be consistent with earlier studies (e.g. Malinovskiy, 2010), although certain configurations, notably tandem ones, had higher detection and matching rates. It can be seen that higher detection rates (Fig. 5-17a) correspond with higher matching rates (Fig. 5-17b), particularly for combinations 4 through 9. The rates were obtained by counting the number of detections or matches happening within a particular 15-minute time window and normalizing the value by the sum of ALPR volumes. As ALPR data was available for only one lane, the values were doubled in an attempt to reflect the total volume in all four general purpose lanes. Transit volume was ignored in this study. The westbound direction captured an average of 10.8% of the total estimated volume with 4.1% of

the estimated volume matched.

It is worth noting that both matching and detection rates can be over 100% theoretically, as contamination from non-vehicle sources may occur and vehicles can contain more than one device, resulting in an over-estimation.

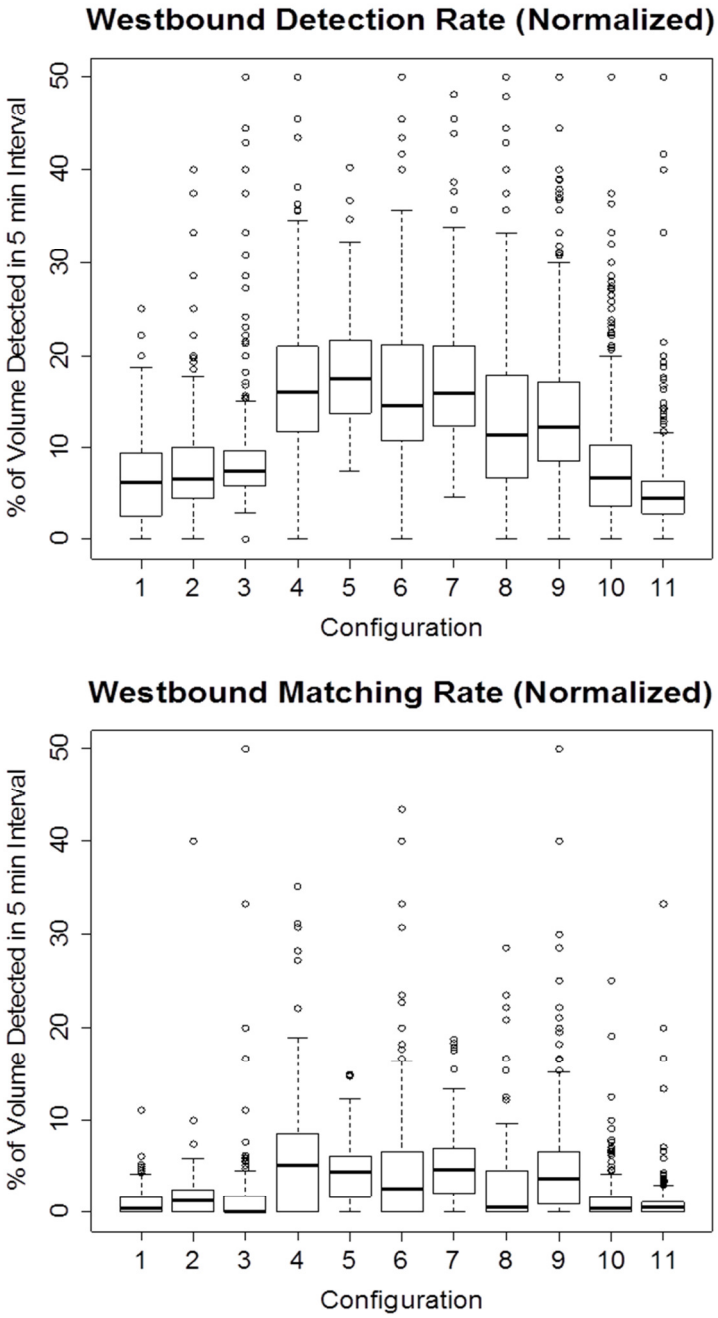



**Figure 5-17: a) Westbound detection rates normalized by ALPR volume b)**

**Westbound matching rates normalized by ALPR volume**

#### *5.4.8.2 Eastbound*

Table 5-5 presents the basic configuration comparison for the eastbound direction. As expected, the results are different. The average error jumps from 7.2 seconds to 19.8 seconds, reflecting the additional error from the intersection delay. However, it should be noted that configuration 6 still manages to demonstrate a relatively low error of 13.6 seconds, although this is still higher than any westbound configuration.

**Table 5-5: Eastbound 15-minute aggregate error statistics by configuration**

| Eastbound Config. | Abs. Error (sec) | Std. Dev (sec) | Max Error (sec) | Min Error (sec) |
|---|---|---|---|---|
| 1 | 28.20 | 17.34 | 61.92 | 1.08 |
| 2 | 20.79 | 10.95 | 40.33 | 0.68 |
| 3 | 19.36 | 10.11 | 52.52 | -5.32 |
| 4 | 17.41 | 11.12 | 45.73 | 0.38 |
| 5 | 21.72 | 12.62 | 47.02 | -1.23 |
| 6 | 13.57 | 7.97 | 31.22 | -2.88 |
| 7 | 23.53 | 23.02 | 97.12 | 1.08 |
| 8 | 8.40 | 6.95 | 20.13 | -6.28 |
| 9 | 13.80 | 9.93 | 41.18 | -13.03 |
| 10 | 33.16 | 22.98 | 114.52 | -1.35 |
| 11 | 19.34 | 9.23 | 39.25 | -1.98 |
| Average TT: 96.0 sec | | | | |

For this direction, the sensors captured an average of 11.4% of the estimated volume. The detections resulted in travel time matches for 5.2% of the total estimate volume. Figure 5-18 shows the detection and matching rates of the 11 configurations for the eastbound direction.

Similar trends as the westbound direction persists, with tandem configurations having higher detection and matching rates.

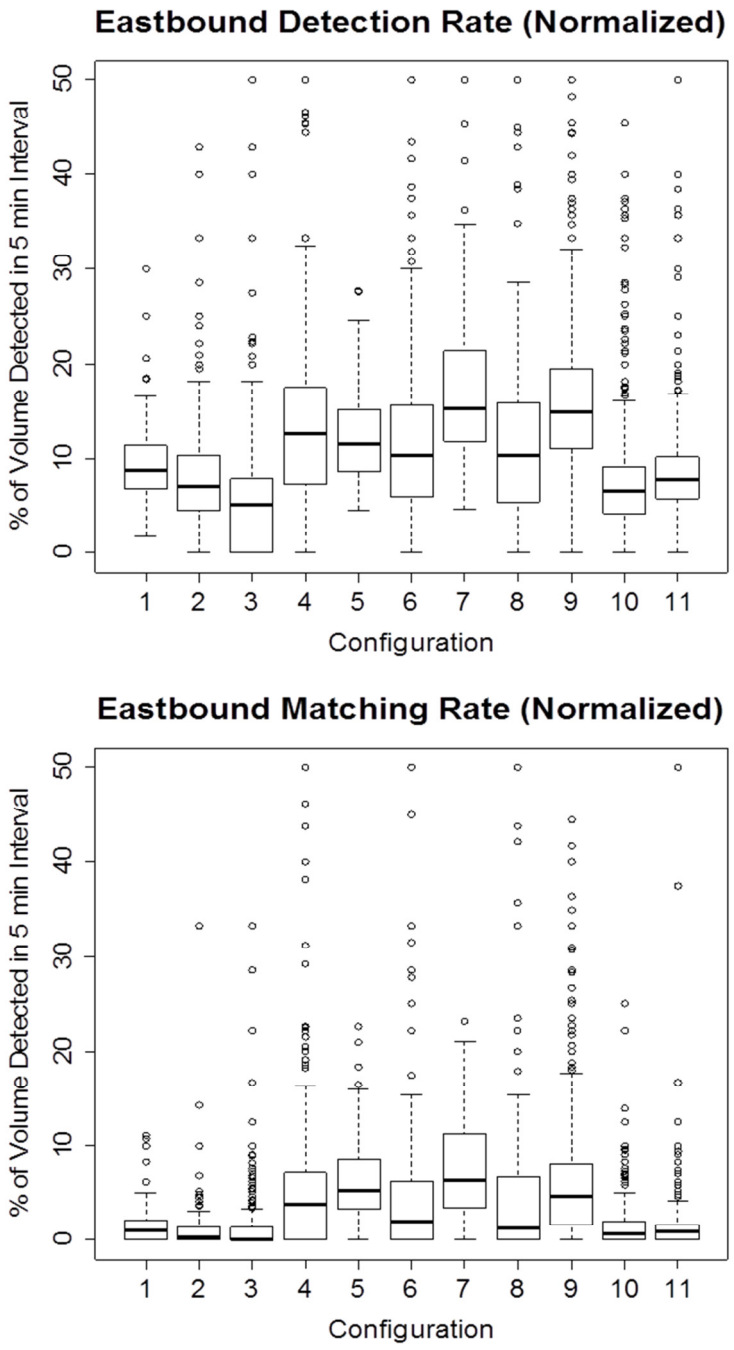



**Figure 5-18: a) Eastbound detection rates normalized by ALPR volume b) Eastbound**

**matching rates normalized by ALPR volume**

### 5.4.9 Configuration Comparison Summary

In general, configurations with higher matching rates provided more accurate results, particularly in the better aligned westbound direction. An additional intersection (47$^{th}$ St) that allows for diversion from the westbound direction only is likely responsible for the lower matching rates in the westbound direction. Configurations 5 and 6, or combinations of 7 dBi and 9 dBi antennae with a 12 dBi directional antennae mounted in the same location did consistently well in both travel directions, obtaining some of the highest matching and detection rates. Configurations 5 and 6 were also among the most accurate, with 6 being the closest to ground truth in part due to its larger antennae which allowed it to obtain a lower error rate in the eastbound direction. Although there is a directional component to this which may increase error in the eastbound direction, the sensors are mounted at the same point in each location, improving the accuracy in the westbound direction. The linear coverage of the sensor footprints is also modest compared to fully omni-directional configurations. Therefore, the findings of the configuration analysis are fairly consistent with the modeling results.

### 5.4.10 Conclusions and Recommendations

Considering the properly aligned westbound results only, a few conclusions can be made about the use of Bluetooth sensors for travel time extraction. The overall error, detection and matching rates, suggest that a combination of sensors is worthwhile. Detection and matching

rates tend to increase with optimal MACAD configurations, which results in an increase in accuracy. In all the experiments conducted, the MACAD Bluetooth methodology provided travel time estimates that were sufficiently accurate, with slight overestimates. The extent of the over-estimation highly depends on the configuration and antenna type and installation location, as shown by the results of this study. Errors ranged from 4% to 13%, but it should be noted that longer corridors (over one mile) would not experience such drastic differences, as the 10.24 sec protocol window plays a smaller role. In this study additional error sources also contributed to such a wide range of possible errors. The short SR-520 experiment, described in Section 4.1, serves as a good example of how precise Bluetooth sensors can be on longer corridors without intersection delay and other potential contaminants.

When using Bluetooth or other MAC-address readers, one has to be very careful of data contamination by intersection delay. Ideally, the sensors would be mounted mid-block to prevent such contamination, but the location of the ALPR sensors dictated Bluetooth device locations in this study. Another potential contamination factor was the proximity of bus stops near the detector locations – if a passenger's Bluetooth device was detected at the first location whilst they were on the bus, after which they have disembarked and walked past the first location, the travel times would be close to the vehicle travel times, yet contain an additional source of error. This problem is exacerbated in areas with high-volume bus stops.

The use of Bluetooth readers to measure travel time provides a comparable alternative to

ALPR technology and can be used with less effort and lower costs. Shorter corridors however, do pose challenges for the Bluetooth detection scheme due to the inherent "zone to zone" detection nature offered by these sensors. In such cases it may be tempting to reduce the detection area in order to decrease the size of the detection zones and thus reduce the error. However, when the zones are reduced, the matching rate drops dramatically. In the experiments described above, configurations that used just one detector per site (thus reducing the detection zone size) had less than half the matching rate of configurations that used two detectors per site, regardless of antenna choice. Of all the configurations attempted, combinations of omni-directional antennae with large detection zones provided the best results, with low absolute error and high matching rates. Combination configurations (4, 5, 6, 7, 8, and 9) had average matching and detection rates of 7.92% and 15.35%, respectively; while single-sensor (at each location) configurations had rates of 3.43% and 9.37% respectively. The higher detection rates may also increase due to extraneous sources, but the matching rates were shown to be statistically significant in reducing error.

Across all configurations, the reported Bluetooth travel time was 8.0% higher than the actual travel times reported by the ALPR sensors. All error rates encountered were well within FHWA's recommendation levels. Although reducing the overall error was a concern, the main goal of this study focused on determining which configuration will provide the lowest relative error, not minimizing the overall error. Lower overall errors can be accomplished using a more discerning filtering algorithm. The least error prone configurations (1,5,6 and 11) reported

travel times that were, on average, 4-7% above the ALPR travel time.

For the eastbound direction, additional intersection delay not considered by ALPR sensors is likely to have played a role that contributed to additional error, severely degrading the results. However, about half of the configurations tested were still able to produce results well under the FHWA threshold.

Errors encountered during this study were almost always positive. This implies that there is still a bias towards slower vehicles within this corridor study. As alluded to in our prior studies (Malinovskiy et al., 2010), this is likely the result of the inherent nature of Bluetooth technology – there is bias towards slower vehicles that have a higher chance of being detected due to longer residence times within the detection zone.

## 5.5 Reducing Spatial Uncertainty

Spatial uncertainty occurs when the exact location of a detected MAC device is unknown. Specifically, the exact location of the device is never truly known due to the nature of the protocol (related to the temporal uncertainty discussed above), however, the largest uncertainty lies not in where in the given detection zone a device is currently located, but which routes a given device owner has taken in between a set of sensors. This issue is of higher interest within mobile sensing, as there are no pre-defined travel corridors. Thus, an innovative means of

asserting the most likely path taken by the detected device must be developed.

### 5.5.1 Inference of plausible paths

The inference of plausible paths can be done in a number of ways. The simplest approach is to assume that the shortest path is the path always taken. Under such a construct the MAC sightings data obtained can be assigned to a known network of available links and the shortest paths between each consecutive sighting of the device can be found. These shortest paths can then be stitched together to provide a complete plausible trajectory for the individual. This approach is illustrated in Figure 5-19 a-c. The green circle represents the first sighting, the red dot represents the last sighting and the blue dots are intermediate ones. However, as can be seen in the figure, there may be a number of different possible paths to choose from, particularly on a network such as an urban grid system. Furthermore, the longer the distance interval between sightings, the more options exist. In Figure 5-19c, a completely plausible path is shown (in red and orange) between all the points. Without additional information, there is nothing to suggest that this path is any less likely than another. However, using the mobile sensing approach, we do have additional information about other travelers. This information can help us determine whether some routes are preferred over others - “popular paths”. This information can be leveraged to create better guesses regarding the plausible paths between MAC sightings, thus reducing overall spatial uncertainty.

Figure 5-19(a-d) shows the concept in action – the pink highlight color represents apriori

path popularity (darker – more popular). Additional popularity information can be gleaned from sections of the trajectory shown in Figure 5-19(a-b). Because mobile sensors often move with the very entities they are trying to detect, two types of interactions are common: (1) following behavior, where the mobile sensor and the sensed entity are moving along the same path and or direction; (2) encounters – where the mobile sensor briefly encounters the sensed device either passing in an opposite direction, at an intersection, or the like. Leveraging this duality, one can obtain path popularity values from the trajectories that have high resolution, or from the "following behavior" ones. The "encounter" segments of trajectories can then be reinforced using the information regarding popularity gleamed from those high resolution trajectories. Figure 5-19c shows the additional path popularity info that can be obtained from the "certain" segment of the trajectory. The remaining "uncertain" portion of the trajectory can then be estimated using a shortest path algorithm on a network where the links are weighed not only by distance, but also by the popularity of a given path. Figure 5-15d shows the final computed trajectory, which follows the most popular paths, while also being one of the shortest paths available.

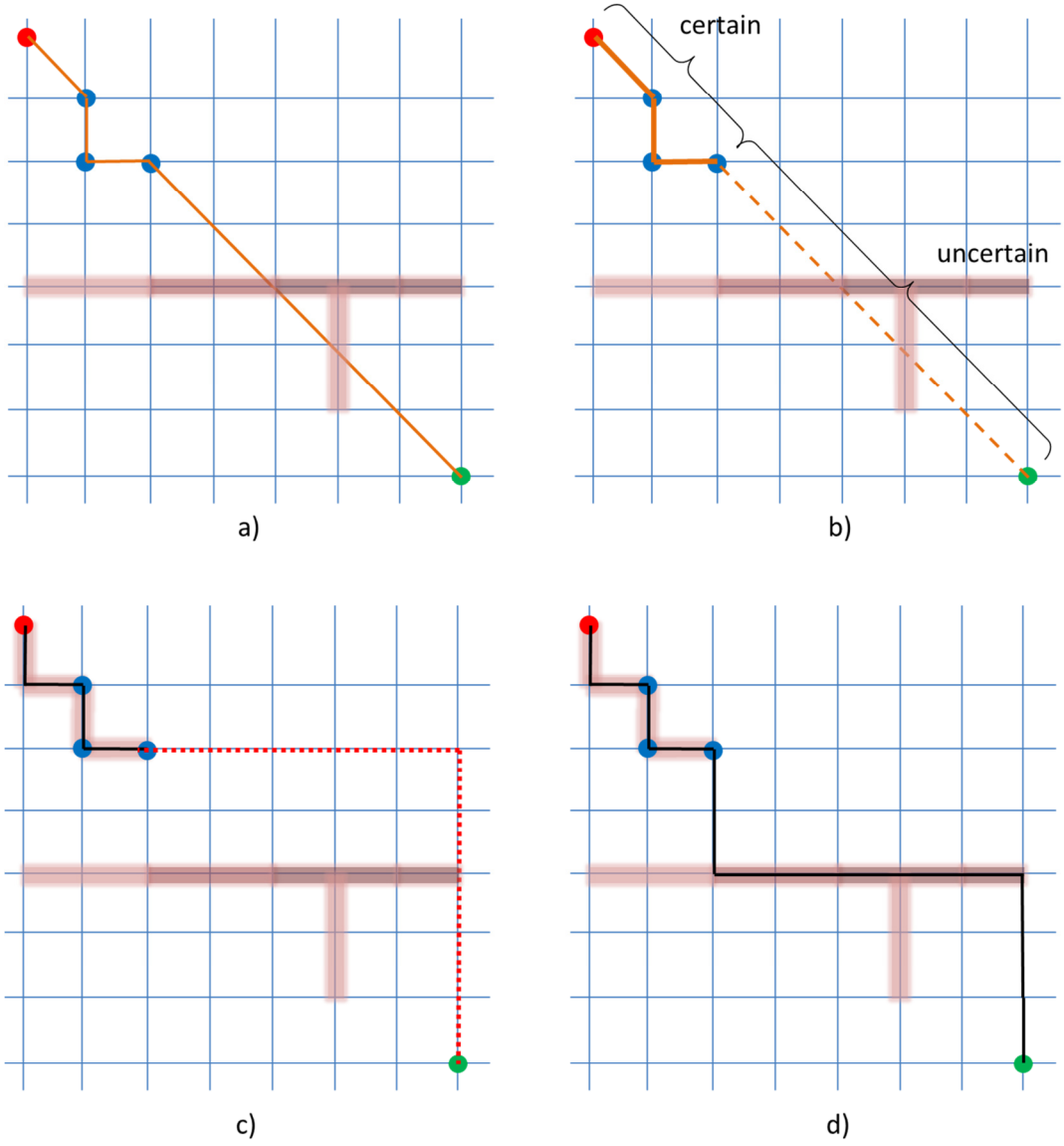



**Figure 5-19: Inference of plausible paths**

This concept has been implemented in the DriveNET system, using the tools mentioned in Section 5.2. A diagram of the implemented system is shown in Figure 5-20. The primary pre-

processing of data occurs within the MAC Matching and Filtering Engine, discussed in Section 5.2. The PG Routing Engine is an open-source routing library that is capable of running shortest path algorithms on PostgreSQL databases. The routable GIS network was obtained from King County and modified to limit the network to the University of Washington Seattle campus only. Additional links were also inserted to better represent the extent of the network. The GIS files were then loaded into the Routable Network contained in the primary DriveNET PostgreSQL database. When MAC trajectories become available, route popularities are calculated and the corresponding weights in the Routable Network are updated based on a cost function that is designed to consider distance, popularity and potentially other factors. Likely routes (plausible paths) are calculated for all MACs seen based on the Routable Network link weights. Additional details follow in the next sections.

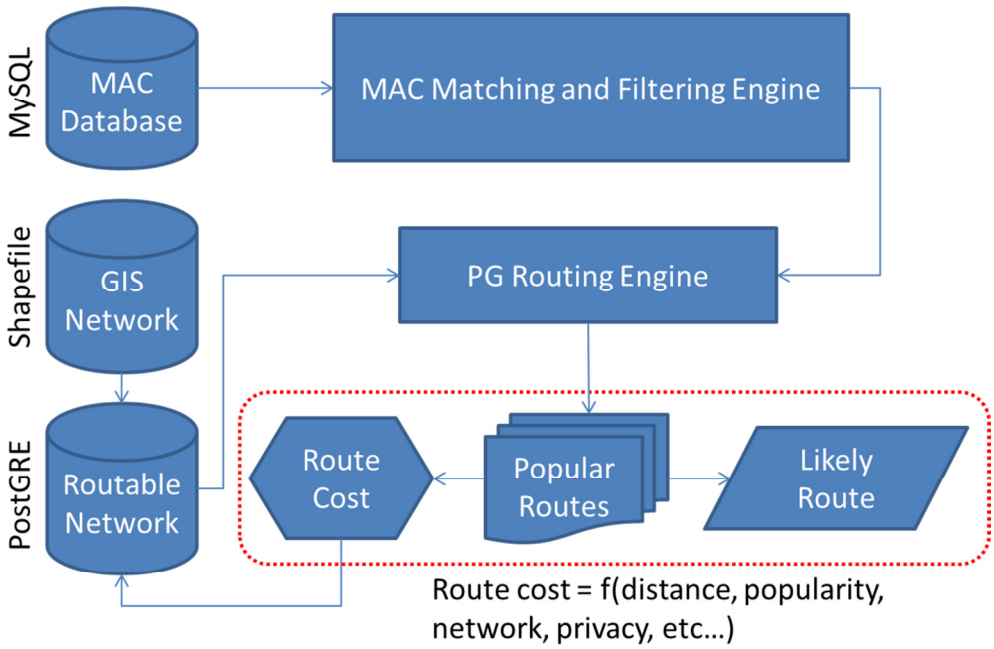



**Figure 5-20: Diagram of route imputation system**

#### *5.5.1.1 Popular Routes Estimation*

To estimate popular routes, the trajectories obtained must be split into "certain" and "uncertain" sections, whereby the "certain" sections are able to reinforce the "uncertain" ones. To do so, some mechanism for distinguishing between which trajectories act as reinforcement and which need to be reinforced must exist. A threshold based algorithm is the simplest means of accomplishing this task – if there is a gap of greater than a certain distance threshold between two consecutive sightings, then that portion of the trajectory is uncertain. Figure 5-21 shows the Routable Network (University of Washington Seattle campus) with popular routes highlighted in red – deeper red color means increasing popularity. This data was obtained from the experiment

described in Section 4.4. It can be seen that increasing the distance threshold increases the popularity of paths; this is reasonable, as more paths are deemed "acceptable" and are assigned to the network. It should also be noted that the relative popularity appears to be similar.

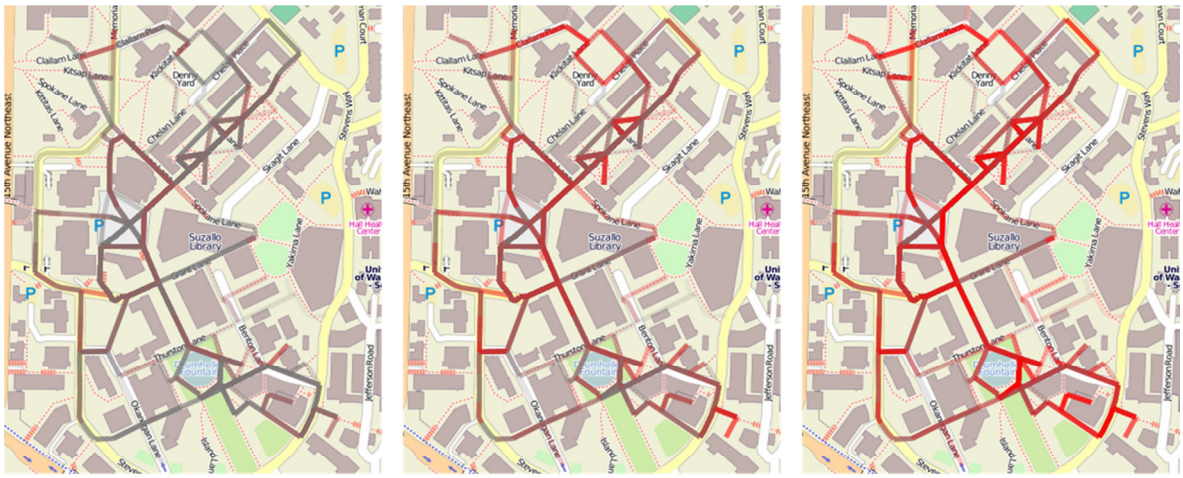



**Figure 5-21: Distance threshold (in meters) for certain/uncertain path discrimination**

Pseudocode for functions used in popular route estimation can be found in Appendix A.

#### *5.5.1.2 Routing Cost Function*

The routing cost determines the link weights within the Routable Network by adjusting the inherent distance of the link in accordance with other parameters deemed of importance, i.e. popularity. The basic form of the function is as follows:

$$if \quad \left(\alpha g(p) + \sum_{i=1}^{n} \beta_i h(\ldots) < 0\right)$$

$$then\ w = d + max\left\langle \alpha g(p) + \sum_{i=1}^{n} \beta_i h(\ldots) \middle| -\gamma d \right\rangle$$

$$else\ w = d + min\left\langle \alpha g(p) + \sum_{i=1}^{n} \beta_i h(\ldots) \middle| \gamma d \right\rangle$$

**(Equation 5-3)**

Where *w* is the new link weight, *d* is the link distance, *g(p)* is a function of the popularity *p* and *h(...)* is a function that incorporates other potential factors (privacy, link centrality, attractions, etc..). α is the route popularity weight, β is the weight of respective parameters included in *h(...)* and *γ* is the maximum allowable proportion of distance that can be affected by all factors. This ensures that the maximum allowable decrease in link weight is not more than *γd.* This formulation allows for extension of the function to incorporate a range of possible parameters that would affect route choice, as outlined in Hoogendoorn and Bovy (2004). In the current implementation, *g(p)* is assumed to be a simple quadratic function, first starting at zero, growing positive, then reducing and becoming negative. This is meant to represent the individual's desire to walk on populated paths, but not ones that are too crowded. The current definition of the g(p) is as follows:

$$g(p) = \frac{\gamma d p^2}{(LOS_E/2)^2} - \frac{\gamma d p}{LOS_E/4}$$

**(Equation 5-4)**

**(Equation 5-5)**

$$LOS_E = 15 * w * t$$

Where BT is the percentage of persons with Bluetooth-visible devices, t is the time interval length in minutes, w is the link width and $\gamma$ is the maximum allowable proportion of distance that can be affected by all factors, as before. The constant 15 comes from the Highway Capacity Manual Level of Service determination, where Level of Service E is defined as 15ppl/ft/min (15 persons per minute, per linear foot of facility width). Level of Service E is considered to be the turning point where pedestrian density becomes a detractor. Thus, g(p) is set up to intersect the x axis, at $LOS_E$, or 15*w*t*BT, or the total number of people needed to be present on a given link during the study interval to cause LOS E. The remaining constraint was the vertex, which was placed at $LOS_E$ /2 and $\gamma d$.

Using this cost function, it becomes possible to update the link costs within the graph to better represent the routing decisions made by the owners of detected MACs. The following section explains the final stage of route assignment.

#### *5.5.1.3 Plausible Route Calculation*

Leveraging the existing network, newly weighed link costs from the routing cost function and the PGRouting shortest path algorithm, it becomes possible to place the detected trajectories onto the network. Figure 5-22 shows the trajectories from the campus experiment described in Section 4.4 and their corresponding mappings to the network, complete with imputed

intermediary points. Each route is calculated based on querying the network route between each of the timepoints contained within a trajectory for a given MAC address' trip. The pseudocode for this operation is as follows:

```
ImputeUncertainTrajectory(Trip t) {

	foreach Trajectory trajectory in t:

		foreach (consecutive) TimePoint p1 and p2 in t:

			newPoints  = getUncertainTrajectory(a,b);
			trajectory.updatePoints(newPoints);

		end;
	end;
}
```

The **`getUncertainTrajectory(a,b)`** function obtains the shortest path on the weighed network using PGRouting, while **`updatePoints(newPoints)`** ensures that the obtained intermediate points fit into the trajectory according to their proper timestamp and location. Trajectories are stored in a Java TreeMap data structure, to ensure the timepoints are kept in a consecutive order.

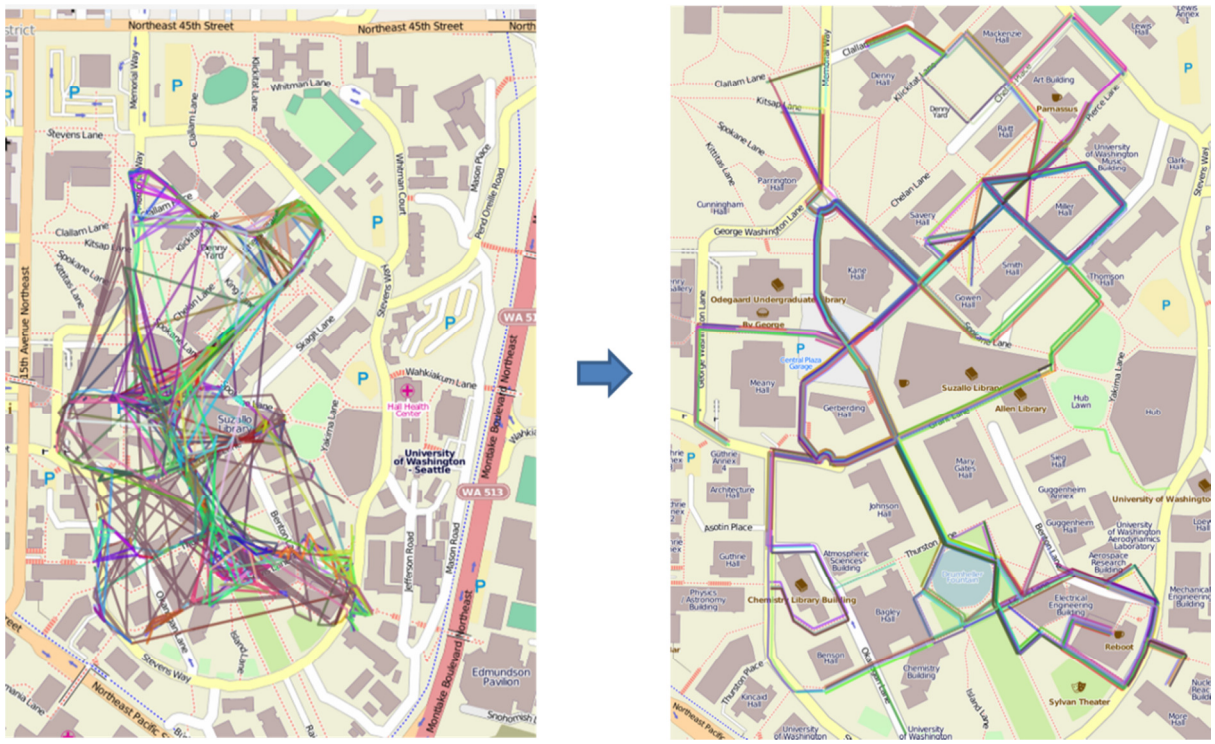



**Figure 5-22: Imputed plausible paths from campus experiment conducted on 04/20/2011**

Although the obtained paths are, at the very least, more plausible than the direct interpretation of MAC sightings, it is of interest to determine how much contribution the method provides. In addition, it is interesting to see how much additional explanatory power the popular route information contains. An additional campus test, described in the next section, was conducted to answer these questions.

### 5.5.2 Verification and Effectiveness

Verification of plausible path imputation is difficult, as the true paths of the entities in

question are not known and cannot be easily obtained. Although simulation is often resorted to in such cases, it was important to understand how the proposed methodology fares in collecting actual data. Thus, an experiment comparing static MAC readers and mobile ones was created. The main concept behind the verification test is matching MACs between static and mobile sensors. The set of MACs seen by each static sensor in an assumed range and the set of MACs seen by mobile devices restricted by GPS coordinates to the corresponding range are compared. First, the comparison is made without path imputation and then with path imputation. The difference in the total matches is considered to be the effectiveness of the algorithm in reducing spatial uncertainty.

#### *5.5.2.1 Experiment Description*

Based on the relative route popularity information obtained from the 4/20/2011 experiment described in Section 4.4, a set of eight static sensors was mounted on the University of Washington Seattle campus. Figure 5-23 shows the sensor locations. These locations were meant to cover the primary gates as well as destinations on campus. However, it should be noted that complete coverage was not necessary for verification. Four MACAD v3.0 devices (1 omni directional antenna, range up to 100m) and four Blip Track Bluetooth (2 directional and 1 omni directional antennae, range up to 100m) were used. BlipTrack sensors were used at locations 1,2,3 and 4 and the remainder were covered by UW MACAD v3.0 devices.

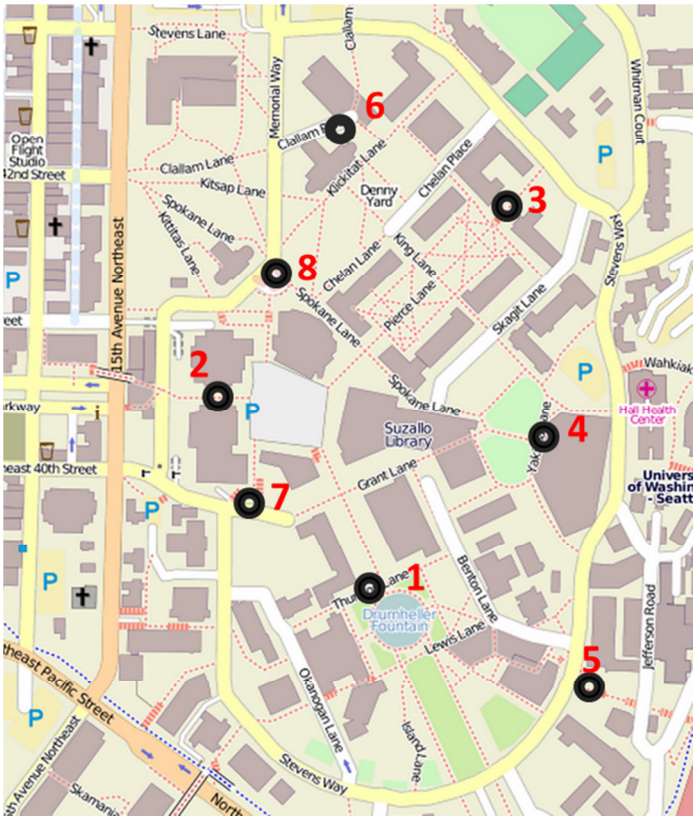



**Figure 5-23: Static sensor mounting locations on University of Washington campus**

Eight volunteers were asked to participate and were given equipment identical to that of the 4/20/2011 experiment, described in Section 4.4. However, instead of roaming freely, the volunteers (observers) were asked to complete two rounds of visits to each sensor, based on predetermined itineraries, shown in Table 5-6. These itineraries were meant to minimize the potential of multiple volunteers visiting the same sensor concurrently and had to be repeated twice. The experiment took place on 3/4/2013, from 11am to 1:30pm. Volunteers were also asked to roughly count pedestrians for 5 minutes at sensors corresponding to the parity of their

assigned id (i.e. observer #3 counts at sensors 1,3,5 and 7). These counts were used to roughly estimate the penetration of Bluetooth-visible devices within the population.

**Table 5-6**: **Observer sensor visit itineraries**

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Obs 1 | 2 | 6 | 5 | 3 | 4 | 8 | 7 | 1 |
| Obs 2 | 4 | 8 | 1 | 7 | 5 | 3 | 6 | 2 |
| Obs 3 | 5 | 2 | 7 | 6 | 8 | 4 | 1 | 3 |
| Obs 4 | 3 | 1 | 4 | 8 | 6 | 7 | 2 | 5 |
| Obs 5 | 1 | 4 | 8 | 2 | 7 | 5 | 3 | 6 |
| Obs 6 | 6 | 7 | 3 | 5 | 1 | 2 | 8 | 4 |
| Obs 7 | 8 | 3 | 6 | 4 | 2 | 1 | 5 | 7 |
| Obs 8 | 7 | 5 | 2 | 1 | 3 | 6 | 4 | 8 |

#### *5.5.2.2 Verification Results*

The data collected by the mobile observers is shown in Figure 5-24b. In comparison to the previously held experiment data (shown in Figure 5-24a), the coverage has expanded to multiple routes, as expected – the volunteers were free to choose their own routes between the eight sensors. However, there was also a drop in the total number of detected devices – 546 unique devices on 4/20/2011 vs. 450 unique devices on 3/4/2013. This may be explained by the

slightly different timing of the experiment (held later in the day), or due to the fact that half of the time spent by the observers was static, while counting pedestrians at sensor locations.


a) 4/20/2011 mobile device detections

b) 3/4/2013 mobile device detections

**Figure 5-24: Comparison of heatmaps of MAC devices detected on campus**

Overall, the mobile sensors picked up more unique devices than the static sensors, with static sensors picking up 343 unique devices during the same time interval (vs. 450 via mobile). 228 addresses were shared between the sensor types, with 565 MACs detected in total by both static and dynamic sensors. Flow between each of the eight static sensor locations was calculated by matching MACs seen at sensor pairs. The flows are displayed as ray charts, with thicker rays

depicting higher volumes, in Figure 5-25. Figure 5-25 also shows the percentage of the unique MAC addresses captured at each location – these do not sum up to one hundred percent, as many MACs were seen by multiple sensors.

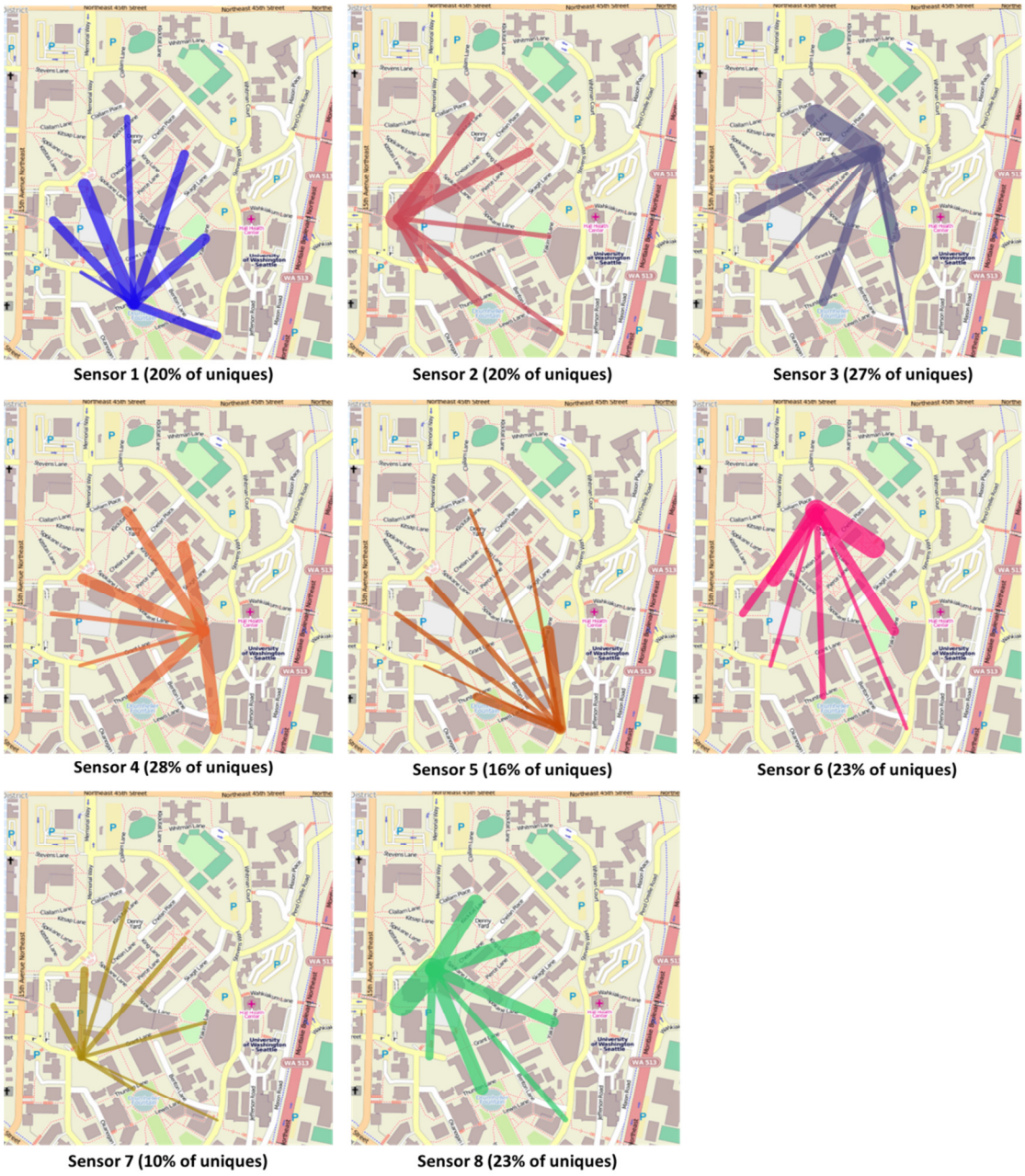



**Figure 5-25: Ray charts depicting pairwise flows for each static sensor location**

Similar analyses can be conducted for the mobile sensors, if an effective range is chosen as a

surrogate to the sensor's range. For example, we can choose 75m as the effective range (smaller than the actual range) to be the effective range, thus considering all MACs found by mobile devices within 75m of a sensor to belong to that particular sensor group. The width of the network links (w) was held constant at 10ft, roughly the average width of a path at the University of Washington. Using this zone threshold and average path width, we can achieve the same pairwise comparisons using dynamic sensor-collected data. Although the sample size for pairings is smaller due to the zone cut-off (90 mobile pairings vs. 409 static pairings), the general trends remain the same. Table 5-7 shows the results between a normalized comparison of the raw (un-routed) mobile sensor pairings and static sensor pairings. On average, the error is less than 5%, meaning that in general, the mobile sensors are able to capture the same pairwise travel trends as static sensors.

**Table 5-7: Relative errors in pairwise flows for mobile and static Bluetooth data**

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | Average Absolute Error % |
|---|---|---|---|---|---|---|---|---|---|
| 1 - Drumheller Fountain | | 7.8% | 0.1% | 9.7% | 1.1% | 2.5% | 17.0% | 13.7% | 7.4% |
| 2 - Red Square Entrance | 2.5% | | 1.4% | 3.6% | 7.5% | 0.3% | 13.7% | 1.8% | 4.4% |
| 3 - Quad Music Hall | 5.0% | 1.9% | | 1.4% | 2.1% | 1.8% | 7.4% | 6.0% | 3.7% |
| 4 - The HUB | 11.1% | 2.4% | 7.1% | | 2.5% | 5.7% | 4.8% | 1.5% | 5.0% |
| 5 - CSE/More Hall | 5.8% | 12.1% | 6.3% | 5.7% | | 6.3% | 9.3% | 1.7% | 6.8% |
| 6 - Paccar Hall | 5.0% | 0.9% | 2.0% | 1.2% | 2.4% | | 2.3% | 4.8% | 2.7% |
| 7 - Meany Hall Lower | 1.9% | 7.6% | 2.6% | 2.4% | 1.3% | 8.6% | | 2.3% | 3.8% |
| 8 - Memorial/Flagpole | 10.8% | 3.0% | 3.8% | 2.3% | 0.5% | 5.9% | 7.5% | | 4.8% |

In addition to comparing flows, the sets of seen MACs can also be compared to

determine if there is an overlap between the MACs seen by the static sensors and the MACs seen by mobile sensors in the same zones. Evaluation of path imputation is also possible, as the imputation technique places certain MACs in locations where they were not detected (as there was no mobile sensor available nearby at the time), but expected to have visited based on path reconstruction. Figure 5-26 shows the percentage of static MACs matched by the mobile approach with and without path imputation, with the popularity weight held constant at zero. The distance threshold of zero represents the baseline condition in which no path reconstruction is performed. It can be seen that path reconstruction, even without popularity imputation provides benefit in terms of the matched MACs (38% vs. 41.5% median matching rate - 3.5% more correct matches, or about a 10% improvement on average).

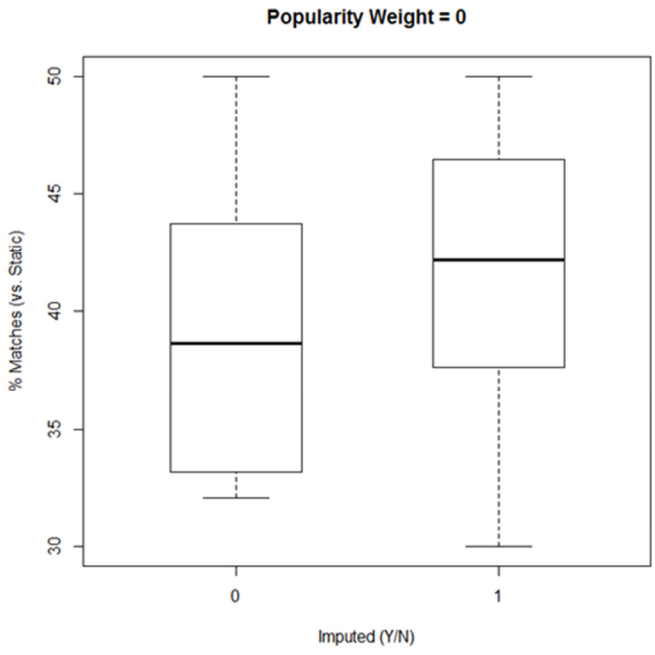


**Figure 5-26: Percent of correctly matched MACs without and with path reconstruction**

Examining the variations due to the popularity function, with weights (alpha) ranging from 1250 to 7500, shows that some additional benefit can be had at the higher alpha values, but appears to peak at alpha = 5000, with a median match rate that is 4% (at 44.5%) higher than the base condition of no poplarity imputation (path reconstruction only - shown in Figure 5-26). The value of alpha = 5000 is also where the maximum matching rate of 51.0% occurs, but at the 20 meter distance threshold. Figure 5-27 shows the results of this sensitivity analysis.

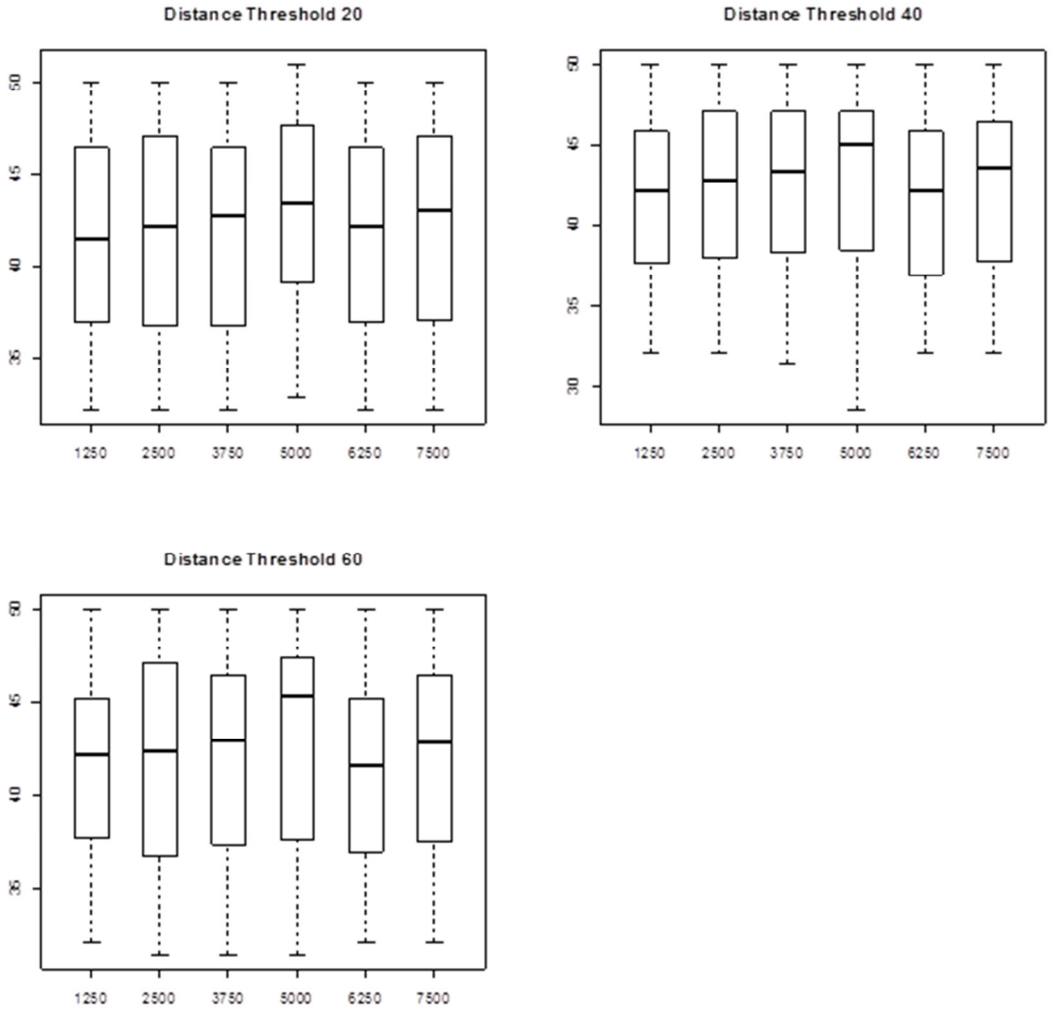



**Figure 5-27: a) Percent of correctly matched MACs by distance threshold with popularity weights of 1250 to 7500.**

Toghether, path reconstruction and popularity-based link weighing can improve the median match rates by 7.5%, from 38% to 44.5%. This may be a modest increase, but the aim of this dissertation is not to provide an optimal means of trajectory reconstruction, but rather to develop a framework for collecting and evaluating mobile MAC data. In this experiment, it was shown

that mobile MAC data is capable of capturing data that is representative of the movements detected via static sensors. Furthermore, it is possible to reconstruct trajectories of individuals traversing the network while concurrently increasing the accuracy of the mobile MAC data.

With the ability to reconstruct individual trajectories, privacy concerns heighten. However, as can be postulated from the above experiment, there is an evident tradeoff between data completeness (which reduces privacy/anonymity) and data quality. The key issue then, becomes preserving the maximum amount of anonymity to retain the necessary utility. The following chapter discusses some of the possible safguards that can be implemented to mitigate privacy concerns.

# Chapter 6 Privacy Concerns in Mobile Device Data

## 6.1 Overview of Inherent Privacy Issues

The primary privacy issue presented by MAC sensing is the lack of consent from the tracked party. Unlike GPS-based and survey methods, there is no agreement that links the data collector and subject, as there is no immediate means of knowing which subject is being tracked. The disclosure of private information (i.e. identity) can only occur if the data collected are intersected with another dataset that contains identifying information. While this maintains at

least an initial level of security for individuals, it also means that there is no easy way of informing the individuals whose data are collected about the effort. While organizations such as the Transportation Security Administration warned individuals about Bluetooth data collection happening in airports with a sign warning, it is very likely that by the time an individual saw the warning, their MAC address would already be in the database, considering the long range of Bluetooth sensors. Furthermore, such sign placement is not always practical, particularly in open areas such as squares and parks, where multiple entries exist. The dynamic MAC data collection approach discussed in this dissertation is even more difficult to control, as any individual or device carrying entity could potentially be recording MAC data. Thus, the open nature of inter-device communication protocols make it inherently difficult to establish any kind of consent framework. Furthermore, the cost of the efforts required to obtain consent would likely outweigh the benefits of using MAC data to begin with – if every individual whose data is captured must be contacted, it may as well have been a survey data collection effort.

Although the data collected is of device MAC addresses, which cannot be directly tied to individuals, sufficient data can produce patterns that can be used to isolate specific individuals with some precision. Thus, it is imperative to keep the MAC address data as secure as possible. For example, in the case of the proposed research, keeping MAC addresses for extensive periods is of little benefit, as shown via the distance threshold sensitivity analysis in Section 5.5.2. Longer unobserved periods result in more challenging trajectory reconstruction scenarios, the longer the gap in observations, the more potential routes and stops the observed entity could have

taken, making the data unreliable. Therefore, deleting the MAC address after 30 minutes and replacing it with the order in which it arrived (along with its device type) should provide sufficient privacy protection and quality data. Privacy is not an issue to be taken lightly and the best approach seems to be to directly convey to the users the type of data being collected and its eventual purpose. While there may be some debate about the use of ubiquitous data, it is unlikely that such a rich data set will continue to be ignored in the future, especially considering the tremendous growth that it may experience. A responsible framework for working with such datasets that protects privacy and maintains anonymity is the most plausible positive outcome. Further informing public of the data they produce contributes to transparency and allows for oversight.

## 6.2 Data sensitivity (identity disclosure) and indefinite storage

Identity disclosure is the primary individual risk that is conveyed by MAC sensing technology. As mentioned above, although there are no direct links between individuals and their device MAC numbers, it is becoming increasingly easy to construct that link either through extensive monitoring or additional datasets. For example, maintaining a list of all locations a particular MAC has been seen over the past year (and consumers often change devices every few years, so it is likely that this device remained with the same owner) can show where this device was found the majority of the time. Looking at the temporal dimension, it becomes possible to

isolate nighttime and weekday evening sightings to determine the home address and weekday daytime hours to determine the work address. This data can be further connected either through publically available listings (such as Yellow Pages, or even Census data) to the specific individual. This type of privacy attack is very difficult to guard against, as even encryption, deletion and modification of the identifier will not absolve the devices unique pattern of behavior. Thus, while manipulation of the stored MAC addresses can allow for some security in case the database is compromised, it still allows for identity disclosure if the observation period is sufficiently long. Therefore, it becomes imperative to not store the data for longer than the minimum period so that such behavior analysis can be conducted. For example; not storing MACs for longer than six hours would not allow the home-work pair to form for most persons employed outside the home. However, much of transportation analysis is focused on origin-destination pairings and this type of data is often found to have the most utility. Therefore, there is a direct conflict between utility and privacy. In such a case, aggregation of parameters is an acceptable solution in many fields. Thus, if data are to be retained for long periods of time, the exact sightings of the MAC address must be aggregated into TAZ, Zip Code or other large areas, sufficient to provide "k-anonymity," (Sweeney, 2002) or making the record indistinguishable from "k" other individuals. For example, guidelines regarding the values for acceptable values of "k" can mimic those of the United States Census Bureau.

## 6.3 Ease of collection

The other factor that makes this particular approach so privacy sensitive is the relative ease of collection the MAC address data. Other data sources, such as License Plate Reader data, may be more directly associated with individual identifiers, however, that data is relatively difficult to collect, as, at least for now, it requires some level of infrastructure investment of achieve meaningful results. However, MAC data collection is easily achievable by an individual and the sensor network can be easily scaled via something as inexpensive as a smartphone app, as demonstrated in this dissertation and by numerous other efforts, such as the MIT "Funf" Platform (Aharony et al., 2012). The open nature of the Bluetooth and WiFi protocols ensure that this is the case. While there are efforts by the manufacturers to curb Bluetooth discoverability (for example, certain versions of the Android platform limit device discoverability to 120 seconds), these tend to be limited to on-person devices. Many devices found in and on vehicles do not actively limit discoverability. Combined with the longer ownership time of vehicles compared to mobile devices, the likelihood of identity disclosure is much higher. Medical devices are yet another potential source of data – numerous implantable technologies are being developed for "real-time diagnosis" capabilities, or the ability to measure and report biometric data on demand. The visibility of such data identifiers remains to be seen, but the likelihood of identity disclosure is once again increased, as this data is likely to contain some level of unique information. In all, it seems that broadcasting of unique MAC identifiers is increasing. This can be a potentially dangerous trend from a privacy perspective. However, while the data collection

stage may be simple, analysis of such data has proven more difficult (as demonstrated in the dissertation). Thus, it is key to retain proper safeguards in any sufficiently sophisticated analysis methodologies. K-anonymity, or aggregation approaches, although not immune to external data linkage risk, can be an important component to these safeguards. A combination of MAC aggregation, for example, never recording the last digit of the MAC address, so that $k_{mac} = 256$ anonymity, combined with geographic aggregation $k_{geo}$ and temporal aggregation, $k_{time}$ would have multiplicative "k" values for anonymity, such that the total anonymity becomes $k_{total} = k_{mac} * k_{geo} * k_{time}$. This makes large "k" values of anonymity easier to attain.

# Chapter 7 Conclusions and Future Research

## 7.1 Summary of Research

Transportation data is changing rapidly. Traditional methods are constantly being replaced and supplemented by novel means of data collection, capable of collecting more information at a lower cost. Mobile device based data is one of the most promising data sources, as it offers scalability and equity. Not only are mobile devices relatively easy to detect (as in most cases they are trying to be detected, for example Bluetooth-visible devices), but also the data can be collected to represent a variety of modes, as demonstrated in this dissertation - transit, vehicle travel and pedestrian travel mobile device MAC data were collected. MAC address data is of particular interest, as it provides re-identification capabilities, is simple to collect and, as discussed, can be used to collect revealed preference data, as the subject is usually unaware that they are tracked. However, the nature of the communication protocols and the diversity of the device population, coupled with variations in sensor locations, results in data sparseness. This sparseness in turn manifests itself into uncertainties. The primary focus of this dissertation has been the exploration and mitigation of these uncertainties.

The work described in this dissertation spans a number of experiments with novel and established MAC address sensing techniques, from point sensor approaches using static sensors to more complicated examples using mobile devices as sensing platforms. The additional

complexity in sensor topology is mirrored by increasing uncertainty in population sample, temporal and spatial dimensions. Multiple strategies for dealing with the uncertainties are introduced, discussed and evaluated using an overarching analysis framework. It is important to recognize that the ideas discussed within this dissertation are applicable outside of the MAC address detection field. For example, one can easily envision a similar set of issues appearing in other opportunistic detection approaches, existing and future. Sparse GPS traces can benefit from the path imputation approaches discussed. Furthermore, a similar technique can be explored by using vehicle back up cameras (appearing on numerous new vehicles) as mobile license plate reader devices.

## 7.2 Research Contributions

Specifically, the research contained within this dissertation makes the following contributions to transportation engineering:

**1) Evaluation of novel MAC address data applications**

A number of novel MAC sensing applications were developed and tested as part of the research contained within this dissertation. To the author's knowledge, experiments regarding evaluating transit wait time with Bluetooth and estimation of pedestrian travel with Bluetooth sensors have been the first of their kind. Interest in non-motorized travel modes is growing and

the need for sustainable and consistent data sources is apparent. The approaches developed within this dissertation are likely to be re-used in a number of studies. Already, additional work by other researchers has continued in the direction MAC based pedestrian travel analysis, demonstrating the promise of this particular direction.

**2) Development of novel pedestrian data collection approach**

In addition to the application of MAC sensing technology to pedestrian data collection, this dissertation also introduces a novel means of collecting such data using a mobile sensing framework, via an app. At the time of development, no such approach existed and to the author's knowledge no such work has yet to be done from a pedestrian data perspective. The main concept is the use of smartphones to collect MAC address sightings and their corresponding locations in spacetime. The sightings are then matched across multiple sensors to recreate trajectories. This allows for network-wide observation using no additional sensing infrastructure.

**3) MAC collection analysis tools and framework**

Innovative sensing tools had to be developed to accomplish the new types of experiments described above. A portable static Bluetooth sensor was developed to help with transit and vehicle travel studies, as at the time of experiment design no suitable sensors were available on the market. The sensor is based on a custom designed board and contains a number of novel

features (for example, compact size and long battery life) that have just recently began appearing appeared in commercially available products. Performance of the sensor is currently being evaluated against a number of commercially available alternatives. Additional sensing tools created include an app designed to capture MAC addresses, much like the static sensor. This app allowed for the evaluation of the mobile sensing approach. In addition to the app, a simulation module was developed to test the approach.

To ingest and analyze data produced by the tools described above, a MAC data uncertainty analysis framework was designed and implemented within the DriveNET platform, containing modules for both static Bluetooth sensor corridor evaluation and mobile app-based travel analysis. The modules are capable of ingesting, processing, evaluating and visualizing MAC address data in a variety of formats. The static sensor data visualization tool has also been expanded to enable comparisons with other travel time sensor types.

**4) Novel algorithms for reduction of uncertainty in MAC address data**

Within the uncertainty analysis framework described above are a number of filtration techniques and algorithms designed to improve the quality of the MAC address data interpretation. However, the primary contribution lies within the path reconstruction algorithm that is capable of reducing the spatial uncertainty by estimating the paths taken by detected entities through the network, thus providing spatial and temporal continuity. By weighing the

network according to popularity and imputing sufficiently long trajectories, it becomes possible to impute the probable locations of individuals on the network. The developed algorithm was compared against a static sensor approach, and while the improvements in MAC matching were modest, the underlying concept was validated.

## 7.3 Future Research

Because this dissertation has been primarily explorative in nature, numerous open questions remain within this budding field. The future research directions can be organized in a similar manner as the dissertation itself, further reducing each of the discussed uncertainties in detail.

**Population uncertainty** – the non-discriminatory nature of MAC sensing is both an opportunity and a problem. While it provides the capacity to collect data from a variety of modes, it is also difficult to distinguish between them. Some of the nascent strategies for doing so have been introduced in this dissertation, but additional work is necessary to improve mode classification capabilities of MAC sensing techniques. In particular, the use of additional identifiers besides the MAC address itself (such as device type, available in Bluetooth 4.0) could improve classification, as some device types are likely to be exclusive to a particular mode. Furthermore, demographic studies regarding the actual ownership rates of visible MAC devices owned by different social groups would allow for some bias correction regarding which

segments of the population are actually being captured. Any methodology developed to deal with this issue would have to account for the constantly changing device ownership rates, device usage (when is the device visible? is there a particular app that a particular group is using that increases visibility?) and even change in communication protocols.

**Temporal uncertainty** – most of the issues with temporal uncertainty in MAC address data revolve around the detection time of the device. Interpreting signal strength indicators, such as RSSI as a potential discriminatory factor to improve the accuracy is a direction that has been thoroughly explored in the WiFi spectrum. Bluetooth-based RSSI readings tend to be less reliable, but work relating them to travel time accuracy has also begun (Araghi et al, 2012). Expanding the current generation tools to incorporate signal strength information could yield improved travel time results. Furthermore, the model developed as part of this dissertation could be improved using other data sets form a number of locations and sensor configurations. This would help mitigate some of the intersection delay noise issues present in the current model. An improved model can also yield improved travel time data.

**Spatial uncertainty** – the framework for spatial trajectory reconstruction presented in this dissertation could be improved in a number of ways. For example, the current static thresholds of cutoff distance and popularity could be made dynamic, relying on a fuzzy logic or a similar approach to estimate appropriate thresholds for a given network and flows. Furthermore, the distance threshold can be expanded to include a temporal component. The cost function could

also be improved to better reflect the distribution of the available popularity data – many links do not have assigned popularity values, but only because the selected analysis interval did not happen to have a dynamic sensor on those links at the time. Thus, the popularity weights are zero-inflated and could benefit from an according model to assign non-zero weights to links that did not have measured data.

Besides the specific data accuracy directions discussed above, several practical issues regarding he mobile sensing approach are also of interest. A lingering question is the motivation of mobile sensors to collect data on behalf of the aggregating entity – an equitable and transparent framework must be developed to provide benefit and explain the terms of use of the data collected to the any individuals participating as sensors, if this approach will be scaled beyond campus experiments. The aggregation privacy safeguards discussed in Chapter 6 must also be considered within such a framework, ensuring that no personal data is shared with the individuals and devices acting as sensors.

Finally, another potentially valuable direction for future work could be the development of additional opportunistic sensing techniques that could take advantage of the framework. As discussed earlier, vehicle-based license plate reading using vehicle-mounted rear (backup) or front cameras could be a potentially promising approach and would be readily adaptable to the framework constructed for MAC data, as well as any other opportunistic re-identification data.

# Bibliography

Aad, I. and V. Niemi, “NRC Data Collection and the Privacy by Design Principles,” In *Proceedings of PhoneSense*; *First International Workshop on Sensing Applications on Mobile Phones*, pp.41-45, November 2010.

Abdelzaher, T., Y. Anokwa, P. Boda, J. Burke, D. Estrin, L. Guibas, A. Kansal, S. Madden, and J. Reich, “Mobiscopes for Human Spaces,” *IEEE Pervasive Computing*, vol. 6, Apr. 2007, pp. 20-29.

Aharony, N., A. Gardner, C. Sumter, W. Pan, Y. A. Montjoye, and A. Pentland. "Fünf, Open Sensing Framework.", 2012.

Ahmed, H., M. El-Darieby, B. Abdulhai, and Y. Morgan. “Bluetooth- and Wi-Fi-Based Mesh Network Platform for Traffic Monitoring” In *Transportation Research Board 87th Annual Meeting*. CD-ROM. Transportation Research Board, 2008. Paper 08-1848.

Aloi, D.N., P.E. Dessert, L. Fay, M. Ronning, and M. Willer. GPS Car Talk: Listening to Bluetooth. Vol. 14, 2003.

Alta Planning and Design. “National Bicycle and Pedestrian Document Project.” Alta Planning and Design, Inc. 2006. www.altaplanning.com/. Accessed January 31, 2011.

American Public Transit Association (2008). Public transportation facts at a glance.

Aslam, J., Z. Butler, F. Constantin, V. Crespi, G. Cybenko, and D. Rus, “Tracking a moving object with a binary sensor network,” *Proceedings of the first international conference on Embedded networked sensor systems - SenSys ’03*, 2003, p. 150.

Barberis, C., Carlevato A., Malnati G., Portelli G., Bluetown: Extracting Floating Transport Data from Personal Mobile Devices via Bluetooth, UMDS 2006, Aalborg, Denmark May 15-17, 2006.

Bayir M.A., M. Demirbas, and N. Eagle, “Discovering spatiotemporal mobility profiles of cellphone users,” *2009 IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks & Workshops*, Jun. 2009, pp. 1-9.

Bechler, M., J. Schiller, and L. Wolf. In-Car Communication Using Wireless Technology. In 8th World Congress on Intelligent Transport Systems. CD-ROM. ITS America, 2001.

Bhagwat, P. Bluetooth: technology for short-range wireless apps. Internet Computing, IEEE, Vol. 5, No. 3, 2001, pp. 96-103.

Borgers, A. and H. Timmermans, “A Model of Pedestrian Route Choice and, Demand for Retail Facilities within Inner-City Shopping Areas”, Geographical analysis 18 (1986) no. 2.

Bullock, D., R. Haseman, J. Wasson and R. Spitler. “Anonymous Bluetooth Probes for Measuring Airport Security Screening Passage Time: The Indianapolis Pilot Deployment” In *Transportation Research Board 89th Annual Meeting*. CD-ROM. Transportation Research Board, Washington D.C., 2010. Paper 10-1438.

Chaintreau, A., P. Hui, J. Crowcroft, C. Diot, R. Gass, and J. Scott, “Impact of human mobility on the design of opportunistic forwarding algorithms,” IEEE Transactions on Mobile Computing, vol. 6, pp. 606–620, 2007.

Cortright, J. “Walking the Walk: How Walkability Raises Home Values in U.S. Cities,” August 2009.

Cuff, B.D. and M. Hansen, “Urban Sensing : Out of the Woods,” *Communications of the ACM*, vol. 51, 2008.

Daamen, W. SimPed: A Pedestrian Simulation Tool for Large Pedestrian Areas. Proc., EuroSIW (European Simulation InteroperabilityWorkshop) (CD-ROM). London, June 24–26, 2002.

E.O’Neill, T.Kindberg, A.F.genSchieck, T.Jones, A.Penn,and D.S.Fraser. “Instrumenting the city: developing methods for observing and understanding the digital cityscape”.In Proc. of the 8th International Conferenceon Ubiquitous Computing (UBICOMP), 2006.

Eagle, N. and A. (Sandy) Pentland, “Reality mining: sensing complex social systems,” *Personal and Ubiquitous Computing*, vol. 10, Nov. 2005, pp. 255-268.

Ertoz, L., Steinback M., Kumar V., "Finding Clusters of Different Sizes, Shapes, and Density in Noisy, High Dimensional Data", Second SIAM International Conference on Data Mining, San Francisco, CA, USA, 2003.

Ferris, B., K. Watkins, and A. Borning (2010). "OneBusAway: Results from providing real-time arrival information for public transit." *Association for Computing Machinery Conference on Human Factors in Computing Systems (CHI) 2010.*

Fetiariason, M., G. Flotterod and M. Bierlaire, "Evaluation of Pedestrian Data Collection Methods within a Simulation Framework," In Proceedings of the Swiss Transport Research Conference 2010, September 2010.

Ghys, K., B. Kuijpers, B. Moelans, W. Othman, D. Vangoidsenhoven and A. Vaisman, "Map matching and uncertainty: an algorithm and real-world experiments," in Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. Seattle, Washington: ACM, 2009, pp. 468–471.

Gräßle, F. and Kretz, T. "An Example of Complex Pedestrian Route Choice." In PED2010 Proceedings Book, 2010.

Gumstix Inc. www.gumstix.com. Accessed Oct 10th, 2010.

Haghani, A., M. Hamedi, K.F. Sadabadi, S. Yound and P. J. Tarnoff. "Freeway Travel Time Ground Truth Data Collection Using Bluetooth Sensors" In *Transportation Research Board*

*89th Annual Meeting.* CD-ROM. Transportation Research Board, Washington D.C., 2010. Paper 10-0729.

Haseman, R.J., J. Wasson and D. Bullock. “Real-Time Measurement of Work-Zone Travel Time Delay and Evaluation Metrics Using Bluetooth Probe Tracking” In *Transportation Research Board 89th Annual Meeting.* CD-ROM. Transportation Research Board, Washington D.C., 2010. Paper 10-1442.

Hewlett Packard “Wi-Fi and Bluetooth interference issues”. 2002.

Hood, J., E. Sall and B. Charlton “A GPS-based bicycle route choice model for San Francisco, California” In *Transportation Letters: The International Journal of Transportation Research*. J. Ross Publishing, Inc. 2011.

Hornsby, K. and M. J. Egenhofer. Modeling moving objects over multiple granularities. Annals of Mathematics and Artificial Intelligence, 36(1-2):177–194, 2002.

Inrix, Inc., Public Sector Solutions, June 10, 2011. http://www.inrix.com/publicsector.asp

Ismail, K., T. Sayed and N. Saunier, "Automated Analysis of Pedestrian-Vehicle Conflicts Using Video Data." In *Transportation Research Record: Journal of the Transportation Research Board*, Washington, DC., 2009, Vol. 2140, pp. 44-54

Kanjo, E., “NoiseSPY: A Real-Time Mobile Phone Platform for Urban Noise Monitoring and Mapping,” *Mobile Networks and Applications*, vol. 15, Nov. 2009, pp. 562-574.

Kansal, A., “Location and Mobility in a Sensor Network of Mobile Phones,” CM SIGMM 17th International workshop on Network and Operating Systems Support for Digital Audio & Video (NOSSDAV), 2007.

Kiukkonen, N., J. Blom, O. Dousse, D. Gatica-Perez, and J. Laurila, "Towards Rich Mobile Phone Datasets: Lausanne Data Collection Campaign", in Proc. ACM Int. Conf. on Pervasive Services (ICPS), Berlin, Jul. 2010

Kong, D. D. Gray, and H. Tao (2006). A Viewpoint Invariant Approach for Crowd Counting. *Proceedings of the 18th International Conference on Pattern Recognition*, Vol. 3, 1187-1190.

Kostakos, V., “An empirical study of spatial and transpatial social networks using Bluetooth and Facebook.”

Krumm, J., “Ubiquitous Advertising: The Killer Applications for the 21st Century”, IEEE Pervasive 2010.

L.-Y. Wei, Y. Zheng, and W.-C. Peng, “Constructing popular routes from uncertain trajectories,” in Proceedings of KDD, 2012.

Lane, N.D., E. Miluzzo, H. Lu, D. Peebles, and T. Choudhury, “AD HOC AND SENSOR NETWORKS A Survey of Mobile Phone Sensing,” *IEEE Communications Magazine*, 2010, pp. 140-150.

Lane, N.D., S.B. Eisenman, and A.T. Campbell, "Urban Sensing Systems : Opportunistic or Participatory ?," 2007, pp. 11-16.

Lee, J.-S. and B. Hoh, "Sell your experiences: a market mechanism based incentive for participatory sensing," *2010 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, Mar. 2010, pp. 60-68.

Li, Y., and Tsukaguchi, H. (2005). "Relationships between network topology and pedestrian route choice behaviour." Journal of the Eastern Asia Society for Transportation Studies, 6, 241–248.

Litman, Todd. "Measuring Transportation: Traffic, Mobility and Accessibility." Victoria Transport Policy Institute. November 2008. Accessed January 31st, 2010. http://www.vtpi.org/measure.pdf

Malinovskiy, Y. and Y. Wang "Pedestrian Travel Pattern Discovery Using Mobile Bluetooth Sensors," Presented at the 91st Annual Meeting of the Transportation Research Board, January 2012, Washington, D.C.

Malinovskiy, Y., Lee U., Wu Y. and Y. Wang "Investigation of Bluetooth-Based Travel Time Estimation Error on a Short Corridor," Presented at the 90th Annual Meeting of the Transportation Research Board, January 2011, Washington, D.C.

Malinovskiy, Y., N. Saunier, and Y. Wang (2012). "Pedestrian Travel Analysis Using Static Bluetooth Sensors." publication in *Transportation Research Record: Journal of the Transportation Research Board, No. 2299,* pp. 137-149. Copyright, National Academy of Sciences, Washington, D.C., 2012.

Malinovskiy, Y., Wu Y., Wang Y. and Lee U., "Field Experiments on Bluetooth-based Travel Time Data Collection," Presented at the 89th Annual Meeting of the Transportation Research Board, January 2010, Washington, D.C.

Malinovskiy, Y., Y. Wu and Y. Wang. "Video-Based Monitoring of Pedestrian Movements at Signalized Intersections", In *Transportation Research Record: Journal of the Transportation Research Board*. Washington, DC., 2008, Vol. 2073, pp. 11-17

Manual on Uniform Traffic Control Devices for Streets and Highways. FHWA, U.S. Department of Transportation, 2009. mutcd.fhwa.dot.gov/.

Meehan, B. Transportation Information Management Team, Federal Highway Administration. Travel Times on Dynamic Message Signs. September 28, 2005 – Travel Time Messages on Dynamic Message Signs National Transportation Operations Center (NTOC) Web Casts Archive http://www.ntoctalks.com/web_casts_archive.php

Mizuta, K., Automated License Plate Readers Applied to Real-Time Arterial Performance: A Feasibility Study, Department of Civil & Environmental Engineering: University of Washington, 2007.

Monsere, C., A. Breakstone, R. L. Bertini, D. Deeter, and G. McGill. "Validating Dynamic Message Sign Freeway Travel Times Using Ground Truth Geospatial Data. In Transportation Research Record: Journal of the Transportation Research Board, No. 1959, TRB, National Research Council, Washington, D.C., 2006, pp. 19–27.

Montoliu, R. and D. Gatica-perez, "Discovering Human Places of Interest from Multimodal Mobile Phone Data" In *Proceedings of 9th Int. Conference on Mobile and Ubiquitous Multimedia,* 2010.

Naini, F.M., "Population Sampling Using Mobile Phones," Research Proposal, Ecole Polytechnique Federale de Lausanee. June 29th, 2010.

O'Neill, E., T.Kindberg, A.F.genSchieck, T.Jones, A.Penn,and D.S.Fraser. "Instrumenting the city: developing methods for observing and understanding the digital cityscape". In Proc. of the 8th International Conferenceon Ubiquitous Computing (UBICOMP), 2006.

Open Streets Map Foundation. (2012) "Open Street Map" www.openstreetmap.com.

Otsu, N. A Threshold Selection Method from Gray-Level Histograms. Systems, Man and Cybernetics, IEEE Transactions on, Vol. 9, No. 1, 1979, pp. 62-66.

Paniati, F. J., “Dynamic Message Sign Recommended Practice and Guidance”. FHWA Memorandum, July 16, 2004. Accessed October 10, 2010. http://mutcd.fhwa.dot.gov/res-memorandum_dms.htm.

Passos, L.S. and R.J.F. Rossetti, “Intelligent Transportation Systems : a Ubiquitous Perspective,” In Proceedings of EPIA 2009.

Patterson, D. J., Liao, L., Fox, D. & Kautz, H. A. Inferring High-Level Behavior from Low-Level Sensors, In Proceedings of the Fifth International Conference on Ubiquitous Computing (Ubicomp), pp. 73-89 (Springer-Verlag). 2003.

Paulos, E. and E. Goodman, “The Familiar Stranger : Anxiety , Comfort , and Play in Public Places,” *Culture*, vol. 6, 2004, pp. 223-230.

Perk, V., J. Flynn, et al. (2008). *Transit Ridership, Reliability and Retention*, National Center for Transit Research.

PGRouting. (2012). “PGrouting” www.pgrouting.org.

PIPS Technology. Product Overview. http://www.pipstechnology.co.uk/products.php?section_id=5&article_id=30. Accessed Feb. 12, 2009.

Pokrajac, D.; Borcean, C.; Johnson, A.; Hobbs, A.; Agodio, L.; Nieves, S.; Balbes, M.; McCauley, L.; Tice, J.; Dare, N.; McKie, J.; Lombardo, B.; Self, B.J.; Austin, J.; ,

"Evaluation of automated license plate reader accuracy," *Telecommunication in Modern Satellite, Cable, and Broadcasting Services, 2009. TELSIKS '09. 9th International Conference on* , vol., no., pp.217-220, 7-9 Oct. 2009.

PTV (2012). Vissim User Manual: Version 5.4. Karlsruhe, Germany.

Quayle, S., P. Koonce, D. DePencier, and D. Bullock. "Freeway Arterial Performance Measures Using MAC Readers: Portland Pilot Study" In Transportation Research Board 89th Annual Meeting. CD-ROM. Transportation Research Board, Washington D.C., 2010.

R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Reas, C., Fry, B. and Maeda, J. "Processing: A Programming Handbook for Visual Designers and Artists (1st ed.)". 2007. The MIT Press, pp. 736, ISBN 0-262-18262-9.

Reddy, S., D. Estrin, M. Hansen, and M. Srivastava, "Examining Micro-Payments for Participatory Sensing Data Collections," *Human Factors*, 2010, pp. 33-36.

Reddy, S., J. Burke, D. Estrin, M. Hansen, and M. Srivastava. "Determining transportation mode on mobile phones". In Proceedings of The 12th IEEE Int. Symposium on Wearable Computers, 2008.

Ruppe, S., M. Junghans, M. Haberjahn and C. Troppenz. “Augmenting the Floating Car Data Approach by Dynamic Indirect Traffic Detection,” In Proceedings *Transport Research Arena – Europe 2012*, Berlin, Germany, 2012.

Schneider, R., R. Patton, J. Toole and C. Raborn. “Pedestrian and Bicycle Data Collection in United States Communities: Quantifying Use, Surveying Users, and Documenting Facility Extent”. Report commissioned for the Federal Highway Administration Office of Natural and Human Environment. January 2005.

Special Interests Group (SIG), Core Specification v4.0, June, 30, 2010. www.bluetooth.com

Sweeney, Latanya. "k-anonymity: A model for protecting privacy." International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10, no. 05 (2002): 557-570.

Tarnoff, P.B., Darcy; Young, Stanley; Wasson, James; Ganig, Nicholas; Sturdevant, James. “Continuing Evolution of Travel Time Data Information Collection and Processing,” In Transportation Research Board 88th Annual Meeting. CD-ROM. Transportation Research Board, Washington D.C., 2009.

The PostgreSQL Global Development Group. (2012). “PostgreSQL” www.postgresql.org.

Traffax, Inc. “Bluetooth Traffic Monitoring Technology - Privacy and Legality Concerns”. http://www.traffaxinc.com/content/privacy-concerns. May 17th, 2008.

Travel Time Data Collection Handbook. Office of Highway Information Management, Federal Highway Administration, U.S. Department of Transportation and Texas Transportation Institute, Texas A&M University System. Report FHWA-PL-98-035, March 1998

Tuominen, A. and T. Ahlqvist, "Is the transport system becoming ubiquitous? Socio-technical roadmapping as a tool for integrating the development of transport policies and intelligent transport systems and services in Finland," *Technological Forecasting and Social Change*, vol. 77, Jan. 2010, pp. 120-134.

Turner, S., W.L. Eisele, R.J. Benz, and D.J. Holdener. 1998. Travel Time Data Collection Handbook. Research Report FHWA-PL-98-035 for the Federal Highway Administration. Washington D.C.

U.S. Department of Transportation. Bureau of Transportation Statistics. "Bicycle and Pedestrian Data: Sources, Needs, & Gaps" BTS00-02. Washington, DC, 2000.

Urbanek, Simon. Rserve--A Fast Way to Provide R Functionality to Applications. In DSC Proceedings, 2003.

Versichele, M., Neutens, T., Goudeseune, S., Van Bossche, F., & Van de Weghe, N. "Mobile mapping of sporting event spectators using bluetooth sensors: Tour of Flanders 2011." Sensors, 12(10), 14196-14213. 2012.

Washington State Office of Financial Management. "Rank of Cities and Towns by April 1, 2009 Population Size" April, 2009. http://www.ofm.wa.gov/pop/april1/rank2009.pdf. Retrieved November 29th, 2009.

Wasson, J.S., J.R. Sturdevant, and D.M. Bullock. Real-Time Travel Time Estimates Using Media Access Control Address Matching. ITE Journal, Vol. 78, No. 6, 2008, pp. 20-23.

Weiser, M., "The computer for the 21st century," *Scientific American*, vol. 265, 1991, p. 94–104.

Whitbeck, J., M.D.D. Amorim, and V. Conan, "Plausible Mobility : Inferring Movement from Contacts," *Measurement*, 2010.

Y. Zheng and X. Zhou, Computing with spatial trajectories. Springer, 2011.

Zaiben Chen; Heng Tao Shen; Xiaofang Zhou, "Discovering popular routes from trajectories," *Data Engineering (ICDE), 2011 IEEE 27th International Conference on* , vol., no., pp.900,911, 11-16 April 2011.

Campanella, MC & Daamen, W (2011). Fundamental diagrams for pedestrian networks. In RD Peacock, ED Kuligowski & JD Averill (Eds.), Pedestrian and evacuation dynamics (pp. 255-264). New York: Springer.

Hoogendoorn, S.P., and P. Bovy. "Pedestrian route-choice and activity scheduling theory and models." Transportation Research Part B: Methodological 38, no. 2 (2004): 169-190.

# Reprint Permissions

Permissions have been granted by the publishers to reuse the contents in the papers listed below in this dissertation.

**Yegor Malinovskiy**, Nicolas Saunier, and Yinhai Wang (2012). "Pedestrian Travel Analysis Using Static Bluetooth Sensors." publication in *Transportation Research Record: Journal of the Transportation Research Board, No. 2299,* pp. 137-149. Copyright, National Academy of Sciences, Washington, D.C., 2012.

**Yegor Malinovskiy** and Y. Wang "Pedestrian Travel Pattern Discovery Using Mobile Bluetooth Sensors," Presented at the 91st Annual Meeting of the Transportation Research Board, January 2012, Washington, D.C.

**Yegor Malinovskiy**, Lee U., Wu Y. and Y. Wang "Investigation of Bluetooth-Based Travel Time Estimation Error on a Short Corridor," Presented at the 90th Annual Meeting of the Transportation Research Board, January 2011, Washington, D.C.

**Yegor Malinovskiy**, Wu Y., Wang Y. and Lee U., "Field Experiments on Bluetooth-based Travel Time Data Collection," Presented at the 89th Annual Meeting of the Transportation Research Board, January 2010, Washington, D.C.

# APPENDIX A

The simplified pseudocode for generating and updating the popular paths database is presented below:

```
GeneratePopularPaths() {
	foreach Detectee d in All_Detectees:

		d.filterTrips();

		foreach Trip t in d.Trips:
			t.filterTrajectory();
			LinkWeights = ImputeCertainTrajectory(t);
		end;

		updateLinkPopularity();
		computeCompositeCost();
		changeLinkWeights();
	end;
}


ImputeCertainTrajectory(Trip t) {

	foreach Trajectory trajectory in t:

		foreach (consecutive) TimePoint p1 and p2 in t:


			distance = haversine_distance(p1,p2);
			interval = time_interval(p1,p2);
			speed = distance/interval;

			if(dist < thresholdDistance)
			{
				newPoints  = getCertainTrajectory(a,b);
				trajectory.updatePoints(newPoints);
			}
		end;
```

```
	end;
}
UpdateLinkPopularity() {
	foreach Link l in LinkWeights:
		UPDATE database SET weight = l.popularity WHERE link = l.id
	end;
}



ComputeCompositeCost() {

	result = "SELECT id, popularity, length FROM database";

	foreach link id in result:

		LinkCostFunction lcf = new LinkCostFunction();
		lcf.setPopularityWeight(popularity);
		newWeight = lcf.getNewCost(length, popularity)

		LinkWeights.update(link, newWeight);
	end;
}



ChangeLinkWeights() {

	foreach Link l in LinkWeights:
		UPDATE database SET composite_weight = l.cost WHERE link = l.id;
	end;
}
```