A Framework for Understanding and Addressing Bias and Sparsity in Mobile Location-Based Traffic Data

Kristian C. Henrickson

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2018

Reading Committee:

Yinhai Wang, Chair

Xuegang (Jeff) Ban

Don MacKenzie

Program Authorized to Offer Degree:

Civil and Environmental Engineering

University of Washington

Abstract

A Framework for Understanding and Addressing Bias and Sparsity in Mobile Location-Based Traffic Data

Kristian C. Henrickson

Chair of the Supervisory Committee:

Yinhai Wang, Full Professor

Department of Civil and Environmental Engineering

Traffic data derived from Global Positioning System (GPS) traces of individual travelers is achieving widespread adoption in transportation engineering and planning, practice, and research. Currently, the majority of such data is obtained from commercial sources, who provide little information about the processes and quality control methods that have been applied to address informative missing data patterns and sampling bias. Looking forward to a future of connected and autonomous vehicles, when fixed mechanical sensing will likely be a thing of the past, there is a growing need to highlight this issue and develop methods to address bias in a

principled way. To do this, it is necessary to understand the sampling mechanisms and their impact on missing data and bias.

The goal of this work is to describe the mechanisms leading to bias, inaccuracy, and missing data in GPS-based probe vehicle data, and to quantify the impact of these mechanisms quantitatively. It is most often the case that commercial probe vehicle data is collected from multiple traveler subpopulations, each with a distinct driving profile, data collection technology, and penetration rate. Thus, this work develops a framework for estimating the impact of these factors on data completeness and bias under heterogeneous driver populations and data collection technologies. This framework is validated using microscopic traffic simulation software under a range of sampling and traffic conditions. The implications of the estimation framework are investigated with respect to real-world probe vehicle datasets and transportation applications.

The primary contributions of this work are as follows. First, this work develops a mathematical framework for describing the relationship between observed data and the true on-road traffic conditions under different sampling parameters and mixed vehicle populations. Second, this work presents an in-depth analysis of the impact of sampling and traffic parameters on statistical representation of real-world probe vehicle data. Finally, a set of case studies are presented illustrating how the proposed framework can be used to improve probe vehicle data quality and fidelity, including the development of a methodology for addressing sampling bias. The methods and guidance provided in this work will be of significant value to public agencies wishing to use probe vehicle data for various forms of transportation analysis, and will inform experimental design and data acquisition agreements for future data collection efforts. Further, this work will support future work in missing data imputation and quality assessment.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Acknowledgements

Thanks first and foremost to Dr. Yinhai Wang, my advisor and committee chair for his support and mentorship throughout my time as a PhD student. Dr. Wang has been an unwavering positive force in my academic, professional, and personal growth.

I would also like to extend a great deal of appreciation to my doctoral committee, Dr. Don MacKenzie, Dr. Jeff Ban, and Dr. Anne Vernez Moudon for their insight and guidance through this process. In particular, my conversations with Dr. MacKenzie's have provided much needed perspective on my research work, including that contained in this document.

A number of agencies and organizations supported my work and activities in a variety of ways over the last six years, including but not limited to Washington State Department of Transportation (WSDOT) and The Pacific Northwest Transportation Consortium (PacTrans). This support is indispensable for PhD students like myself, and I greatly appreciate it. I was also lucky enough to be a Valle Fellow during 2016 – 2017, and I am indebted to Dayna Cole and the Valle program for the amazing 10 months I spent in Denmark.

I would be remiss not to thank the many members of the Smart Transportation Application and Research Laboratory (STAR Lab) for their insight, support, and friendship. In particular, John Ash has been a willing, energetic, and capable partner for many of my efforts at the UW, and I am so grateful for his companionship. Zhiyong Cui has also been a faithful co-conspirator and friend. Thanks as well to Wenbo Zhu, Ruimin Ke, Ziqiang Zeng, and Mayuree Binjolkar for helping to create such a positive and welcoming community in the STAR Lab, and for their collaborative spirit in research. Though not members of the STAR Lab, I would also like to thank Dr. Francisco Camara Pereira and my fellow researchers Dr. Filipe Rodrigues, Dr. George Panagakos, and Ioulia

Markou at DTU Management Engineering for their friendship, support, and collaboration during my time at DTU.

Finally, and most importantly, I would like to thank my wife, Alicia, for her devotion, attitude, and hard work over the last ten years of our lives together. I do not believe I would be here without her.

Chapter 1: Overview

In this work, a framework is developed to describe and address some key quality issues present in probe vehicle data. Starting with established queuing theory, this work introduces a mathematical model that can accurately describe the sampling process and the resulting sample size, bias, and completeness of probe vehicle data under different sampling and traffic conditions. The primary contributions of this work a) a framework for estimating sampling bias, sample size, and completeness in heterogeneous vehicle populations, b) an in-depth analysis of the impact of these factors on real-world probe vehicle data, and c) two case studies illustrating how the proposed framework can be used in planning and executing probe vehicle data collection projects. The methods and guidance provided in this work will be of significant value to public agencies wishing to use probe vehicle data for various forms of transportation analysis, and will inform experimental design and data acquisition agreements for future data collection efforts. Further, this work will support future work in missing data imputation and quality assessment.

The remainder of this document is structured as follows: Chapter 2 provides a general background on the problem space and outlines the objectives of this work in detail. Chapter Chapter 3: covers relevant literature on missing data, sampling bias, and previous efforts to address these challenges. Chapter Chapter 4: develops a mathematical framework for estimating sample size, bias, and missing data rates for a given set of traffic and sampling parameters. In Chapter Chapter 5:, the microscopic traffic simulation work completed to validate the framework developed in Chapter Chapter 4: is described. Chapter Chapter 6: develops a set of case studies to demonstrate the application and implications of the proposed framework, including: an analysis of

completeness using real-world probe vehicle data (6.1), development of a probe data collection plan to meet sample size and completeness requirements (6.2), and a set of methods for addressing sampling bias (6.3). Chapter Chapter 7: offers some additional discussion and concluding remarks.

Chapter 2: Introduction

2.1 Problem Statement

Over the past few years the quality and coverage of probe vehicle data has advanced along with improved Global Position System (GPS) accuracy and increased probe vehicle penetration rates. Correspondingly, there has been a rapid increase in the reliance on probe vehicle data in engineering practice and in research. Such data has the potential to offer greater coverage and granularity compared to conventional transportation data, and do so without the high cost and complexity associated with maintaining a large network of on-road sensing hardware.

In a general sense, this work is motivated by potential of probe vehicle data, and eventually data from connected and autonomous vehicles, to increase the scope, granularity, and inclusiveness of transportation planning, management, and analysis activities. However, this potential cannot be fully actualized without addressing the current quality and bias issues. Toward this end, there are three primary motivations for this work. The initial motivation for this work came from a desire to raise awareness and interest in the topic of probe vehicle data sampling bias in public agencies and the transportation field in general. Much of the previous work on validating probe vehicle data quality has been based on empirical studies which largely ignore causal mechanisms that contribute to bias and uncertainty. Second, and more importantly, this work seeks to explain and quantify the causal mechanisms related to probe vehicle bias, sample size, and missingness, to inform experimental design and the development of methods to address missing data and bias.

Finally, this work is motivated by the need to consider sampling and bias in probe vehicle data analysis, and supports this goal by providing a quantitative explanation for many of the idiosyncrasies of this data. These three motivations are discussed in more detail in the following subsections.

2.1.1 *Commercial Probe Vehicle Data – the Black Box*

The majority of prove vehicle data comes from commercial sources who provide little information to the end user regarding how informative missing data patterns and sampling bias are addressed. An intuitive understanding of how this data is collected, as well as some previous work on this topic (Hallenbeck and McCormack 2015), suggests that a certain amount of sampling bias is inevitable, but the causal mechanisms and extent of this bias are not often discussed. As we look forward to a future of connected and autonomous vehicles, when fixed mechanical traffic sensors are all but obsolete, it will become increasingly vital for public agencies to be involved in the data collection, processing, and quality control processes, rather than simply purchasing a pre-packaged product from one or more vendors.

It is no secret that GPS traces are often noisy, with many anomalous and inaccurate values. Furthermore, a substantial amount of work is needed to associated millions of GPS traces with a geospatial representation of the road network and aggregate it to link-level traffic data, a process that is itself imperfect and often leads to additional errors. Commercial probe vehicle data providers are currently in control of and responsible for every aspect of data processing, anomaly detection and quality control, spatial conflation, and aggregation. We can only assume that these companies have methods in place to deal with quality and bias issues. However, understanding the mechanisms contributing to bias is key to addressing it in a principled way. Thus, if the customer and user is not made aware of the processing methods that are applied to the data, it seriously

undermines the confidence that can be placed in the product. Previous work has shown that commercial traffic data generally conforms to the quality requirements placed in the procurement documents in terms of aggregate measures of accuracy and bias (I-95 Corridor Coalition 2018). That said, the significant temporal and spatial variability in data quality maybe in part driven by a lack of considerations for the contributing mechanisms. The customer is in general not able to assess this possibility, because the data is delivered in an aggregate state with little indications of the adjustments and processes that have been applied to it.

Figure 2-1 illustrates the process of scientific inquiry in very general terms, starting with the identifying and defining a research question and proceeding (potentially in a cyclical fashion) to reach a conclusion. In most such inquiries, the notion of drawing conclusions with little knowledge of or control over how the data is collected and processed would be disconcerting. However, as shown in Figure 2-1, this is the reality when commercial probe data is applied in traffic operations and safety analysis. While probe vehicle data has significant benefits over conventional traffic monitoring technologies, its usefulness would be improved dramatically by opening up the "black box" that obscures the data collection and processing procedures performed by data providers. If data consumers (i.e. public agencies and engineers) were allowed more insight and involvement in the data collection, processing, and quality control processes, more efforts could be made to develop and implement improved data processing and quality control methods, and to consider any remaining weaknesses in the data in subsequent analysis. The end result would likely benefit both the consumers and providers because engineers, decision makers, and the public could place greater faith in the resulting analysis and conclusions.

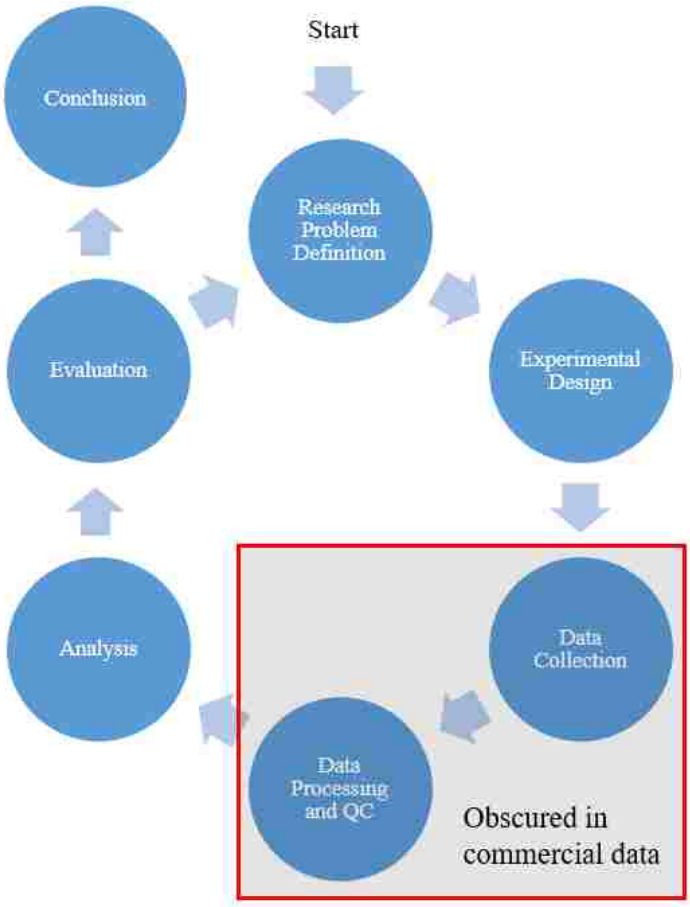

Figure 2-1: An Illustration of Scientific Inquery

2.1.2 *Addressing Bias and Missing Data*

Probe vehicle data is derived from the GPS updates of a contributing subset of all vehicles on the road, and is most often aggregated to link-level mean speeds or travel times before it is delivered to customers. GPS updates are not delivered continuously, instead vehicle locations and/or speeds are updated at (usually) regular intervals. Thus, the expected number of point updates that are represented in a link-level traffic observation over a fixed time period is a function of a) the number of contributing vehicles on the road, b) the sampling frequency distribution of the contributing vehicles, and c) the travel speed distribution of the contributing vehicles (i.e. faster moving vehicles will spend less time on a link, and so be sampled fewer times in a time interval

all else being equal). Commercial probe vehicle data is most often aggregated from multiple provider types and technologies, each of which is associated with a certain penetration rate, sampling rate distribution, and driving profile (Bucknell and Herrera 2014; S. Kim and Coifman 2014). All else being equal, higher sample counts will be associated with slower moving vehicles, higher penetration rates, and more frequent sampling. Understanding the sampling bias and missing data patterns that arise from this mixture of populations is crucial to assessing the quality and suitability of a probe vehicle dataset for various types of transportation analysis. It is the intent of this work to develop a framework to describe these causal mechanisms in quantitative terms, and show how this framework can be used to improve probe vehicle data quality and fidelity.

Fundamentally, there are to factors that contribute to bias in probe vehicle data. First, with a mix of vehicle driving characteristics and sampling rates, certain subpopulations are likely to be over represented. Second, all else being equal, high volume, more congested time periods are more complete than less congested time periods. Figure 2-2 illustrates the sampling process for two vehicles traveling across three road links during a given time interval. It is clear that vehicle B has produced a larger number of samples on all traveled links, which may be due to a) more frequent sampling, b) lower travel speed, or c) a combination of the two. As a result, vehicle B is more likely to be observed on any given road segment and may be weighted more heavily in the resulting speed estimate (depending on how the point speeds are aggregated). The fact that different vehicles, with different speeds and sampling rates, are represented differently in the resulting traffic data is the first source of bias.

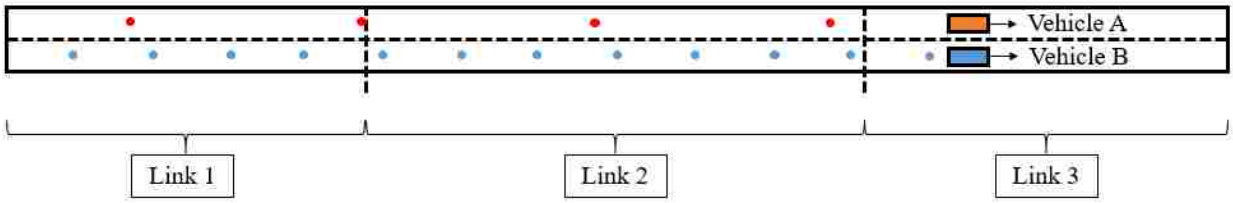

Figure 2-2: Sampling Process for Two Vehicles on Three Road Links

The second source of bias is the fact that higher traffic density, more congested conditions are less likely to be missing from a probe vehicle data set that uncongested conditions. Data can only be obtained for time intervals during which a probe vehicle passed through the road link(s) of interest. Because the probe vehicle population constitutes a small fraction of the overall vehicle population, the probability of a probe vehicle appearing on a road link increases with traffic density as shown in Figure 2-3. Generally speaking, heavier traffic conditions are of the most interest for public agencies, so it could be argued that this is nearly an ideal scenario. However, consider a travel time reliability study based on such data. If a particular road section becomes heavily congested 20% of the time period of interest and is otherwise free-flowing, the congested period will likely be nearly complete while the uncongested time periods will be missing data to a significantly greater degree. It is clear from this discussion that bias and missing data are two closely related problems. Missing data is a general quality issue that, on its own, can have deleterious impacts on analytical results. However, because the pattern of missingness is related to the quantity of interest, missingness tends to bias the observed data toward slower moving conditions, causing further degradation of quality.

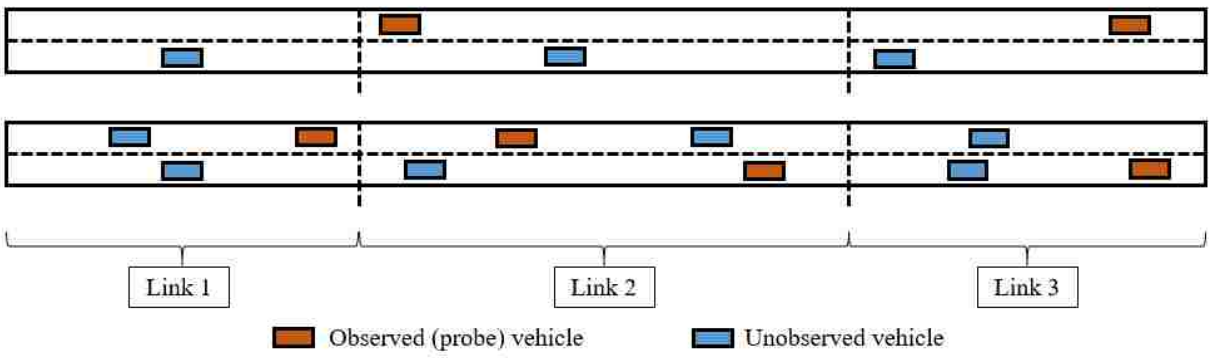

Figure 2-3: Congested (bottom) vs. Uncongeted (top) Sampling Conditions

2.1.3 *Understanding the Characteristics of Probe Vehicle Data*

Intuitively, it seems likely that increasing the overall penetration rate of contributing vehicles will gradually improve the accuracy of the resulting traffic data to the point where it is perfectly representative of the true conditions. However, as this work demonstrates, increased sample size does not necessarily lead to improved accuracy, and in some cases may actually exacerbate sampling bias. Further, many of the quality issues observed in existing probe vehicle data sets can be understood as artifacts of the data collection process. Because of this, understanding the sampling and data collection process is key to informed application of probe vehicle data in engineering analysis.

Consider that a number of past studies have dealt with high missing data rates, errors, and anomalies with little or no consideration given to the causal factors related to the sampling process (e.g. Gong and Fan 2017). In many cases, the anomalies will arise from more or less obvious non-sampling related factors, such as construction activities, inclement weather, and others. It critically important to understand and distinguish erroneous data from anomalous traffic conditions, one being a data quality issue that must be addressed prior to completing any analysis and the other a potential phenomenon worthy of study. A good example can be found in (Jenelius and

Koutsopoulos 2013), where it is hypothesized that slower speeds on shorter road links are attributable to deceleration / acceleration behavior. It is entirely possible that the authors have correctly identified the primary cause of the observed traffic behavior. However, no mention was made of potential biases due to faster vehicles being less likely to be observed on shorter links, especially for low frequency probe vehicle data as was used in this study. Thus, this work is in part motivated by the need to inform transportation analysts and researchers that will be expected to apply probe vehicle data in transportation studies, such that data quality issues are identified and treated in a principled way.

2.2 Background

2.2.1 *A brief History of Crowd Sourced Traffic Data*

The earliest work in mobile GPS as a traffic data source preceded the rise of GPS enabled mobile phones, and so generally assumed that dedicated hardware and communications systems would be required (Shawn M Turner et al. 1998). In fact, the earliest commercially available probe vehicle data came from freight, taxi, and other commercial vehicles equipped with GPS transponders as well as, to a lesser extent, dedicated in-vehicle consumer GPS units. For example, the company INRIX based in Kirkland, WA was founded in 2004, and their initial flag ship product consisted primarily of aggregated data from commercial vehicle-based dedicated GPS (INRIX 2006). NAVTEQ was another early leader in probe vehicle-based traffic data, and began offering real-time traffic data based largely on commercial GPS probes in 2007 (Wolstan 2007). As mobile phones became more capable and ubiquitous, interest in using their combined locationing and communications capabilities to crowd source traffic data grew. Among the first real world applications of crowd sourced cell phone location data for traffic information, Google Maps began offering real-time traffic information in early 2007 (D. Wang 2007; Crackberry 2007). Early work

in mobile phone-based traffic data also included the Mobile Century and Mobile Millennium projects at UC Berkley (Herrera et al. 2010; K. Greene 2008). In the Mobile Millennium project, volunteers were asked to contribute data via a mobile phone app in return for access to real-time traffic condition updates (K. Greene 2008).

As the quality and coverage of probe vehicle data improved, public agencies were increasingly interested the potential of such data to meet their needs at lower cost and without the traffic disruption associated with maintaining fixed mechanical traffic sensors. In 2008, the I-95 Corridor Coalition project selected INRIX as the primary traffic data provider for their Vehicle Probe Project, a multi-state project with the goal of providing coalition members with better access to travel-time and speed data (I-95 Corridor Coalition 2018). The availability and use of commercial probe vehicle data in traffic information systems and analysis increased in the years that followed, and in late 2013 HERE North America (formerly NAVTEQ/Nokia) began providing probe vehicle data for the entire national highway system under contract with the Federal Highway Administration (FHWA). This dataset, titled the National Performance Management Research Data Set (NPMRDS), was obtained through mobile phones, dedicated GPS, and embedded fleet systems and provided by the FHWA free of charge to metropolitan planning organizations (MPOs) and departments of transportation (DOTs) throughout the USA for performance monitoring and planning activities (FHWA Office of Operations 2013). The NPMRDS data provider was switched to a consortium including INRIX in 2017, but the program remains an important source of traffic data for DOTs and MPOs across of the country (FHWA 2018). In 2014, the I-95 Corridor Coalition Vehicle Probe Project was expanded to include data from HERE and TomTom, as well as INRIX (I-95 Corridor Coalition 2018). The data provided by these three vendors has been subjected to

substantial testing and validation, which to date likely constitutes the most comprehensive validation and comparison of commercial probe vehicle data.

Probe vehicle data from the NPMRDS and Vehicle Probe Project, as well as multiple other acquisitions from various vendors, have been extensively applied in traveler information systems, performance monitoring, research, and engineering work. For example, the Texas Transportation Institute's Urban Mobility Report (and subsequently the Urban Mobility Scorecard) has been based in part on INRIX data since 2010, having previously been based exclusively on federal and local transportation agency data (Schrank, Lomax, and Turner 2010; Schrank and Lomax 2009). The Danish Road Directorate uses commercial probe vehicle data for traffic monitoring and management activities, and a number of other transportation agencies around the world are moving in this direction (Bends 2017; Iteris Inc 2018; Mcnamara et al. 2015; EUEIP 2018). In addition to practical applications, it is clear from Figure 2-4 that the use of probe vehicle data in research has risen dramatically over the last decade.

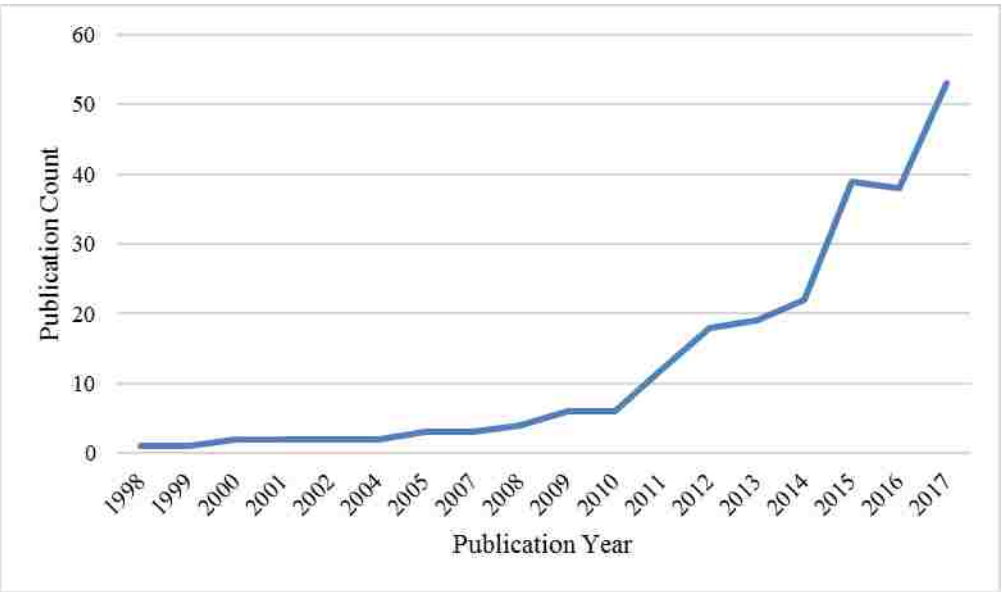

Figure 2-4: Trend for publication topics "floating car" or "probe vehicle" which include the term "GPS" since 1998 (Science 2018)

It is worth noting that the use of traveler GPS traces for collecting traffic information is not limited to on-road vehicles. A great deal of research has applied such data for non-motorized travel and transit. For example, previous work has examined the use of mobile phone GPS traces for bicyclist experience mapping (Eisenman et al. 2009), identifying spatial and temporal trends in sporting activity (Ferrari and Mamei 2013), and combined with fixed sensor data to estimate network-wide bicycle volumes (Strauss, Miranda-Moreno, and Morency 2015; Jestico, Nelson, and Winters 2016). However, GPS data comes with a unique set of bias challenges for non-motorized vehicle traffic, because the travel reporting process most often requires active participation on the part of travelers (Romanillos et al. 2016; Jestico, Nelson, and Winters 2016). For example, it is most often the case for bicycle data that either a) cyclist GPS traces are collected by an athletic performance tracking application, which are disproportionately used by more experienced riders or b) GPS traces are collected by a study-specific data collection application, which are most likely to recruit more enthusiastic and engaged cyclists.

Much of the early work in GPS-based traffic data was in relatively small-scale, carefully controlled experiments. In such experiments, for example the Mobile Century and Mobile Millennium projects, researchers had control over and knowledge of every aspect of data collection and processing, and the results could be evaluated in light of this knowledge. Currently, the majority of such data comes pre-aggregated from commercial sources and, except in a few very limited cases, this knowledge and control has been sacrificed to protect intellectual property and achieve economies of scale. In the short term this would seem to be a relatively good bargain, allowing public agencies to previously unimaginable quantities and coverage of traffic data at comparatively low cost. However, as we look forward to a future of connected and autonomous vehicles, when fixed mechanical sensors will be all but obsolete, allowing this commercial

monopoly on every aspect of traffic data collection and processing to continue seems both perilous and probably impractical. Actualizing a system of fully connected and autonomous vehicles will require technological and strategic partnerships between public and commercial agencies, partnerships that are not likely to be conducive to intellectual property protection at the expense of transparency. Figure 2-5 illustrates the author's view on the progression of probe vehicle data from small scale experiments to ubiquity in research and practice. While it is only one person's perspective, the trajectory suggests increased data transparency will (and should) accompany the transition to a fully connected transportation system.

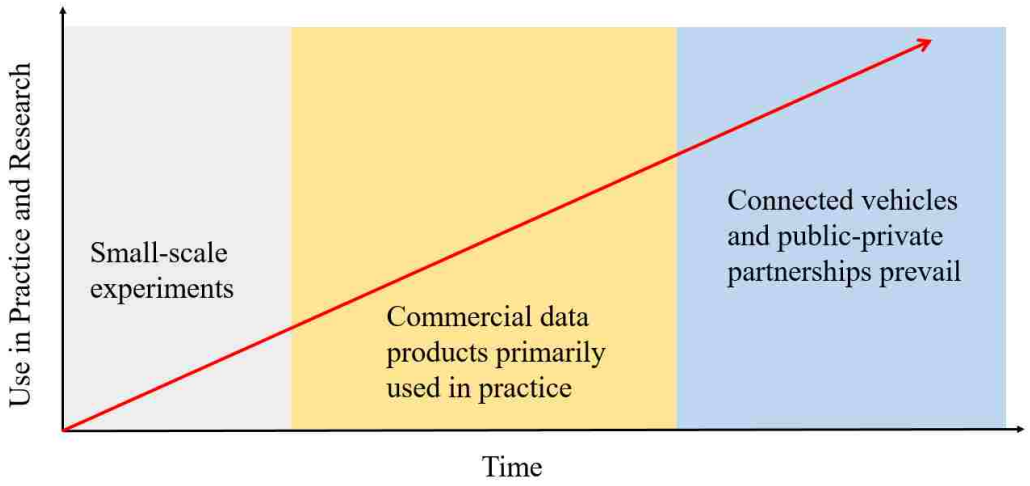

Figure 2-5: The Future of Probe Vehicle data: The Author's View

2.2.2 *Probe Vehicle Data Basics*

In research, there are a number of different ways in which prove vehicle data collection is conceptualized and/or performed. Some have envisioned a system where more of the processing is completed by on-vehicle hardware, in order to reduce the amount of information that is transmitted to the processing center, e.g. (M. Ferman, Blumenfeld, and Dai 2005). The assumptions made in this work regarding the technologies and methods of probe vehicle data

collection draw from a number of sources (Jenelius and Koutsopoulos 2014; Tilley 2012), and to the author's knowledge represent the current state of practice. As shown in Figure 2-6 (adapted from (Tilley 2012; Chen, Chen, and Liu 2013; M. Ferman, Blumenfeld, and Dai 2005), onboard GPS hardware (including mobile phones, consumer GPS, or commercial GPS units) determine the location of the vehicle at discrete points in time via communication with GPS satellites. For a variety of reasons including battery/energy conservation, GPS updates are not delivered continuously, and instead are delivered at regular time intervals. The vehicle's speed and heading is determined from subsequent location updates, and the resulting vector of location, speed, heading, time, and possibly other state information is transmitted to a data processing center via cellular or satellite communications. At the processing center, vehicle state vectors are processed and matched to a map of the road system, and aggregated to produce unique measures for each road link and observation interval. In the case of commercial data providers, the data quality control and aggregation methods are typically proprietary. In addition to link-level speed or travel time data, a variety of other traffic products may be produced at the data processing center including predictive analytics and anomaly detection, origin-destination data, and others. Here, the focus is on link-level speed and travel time data.

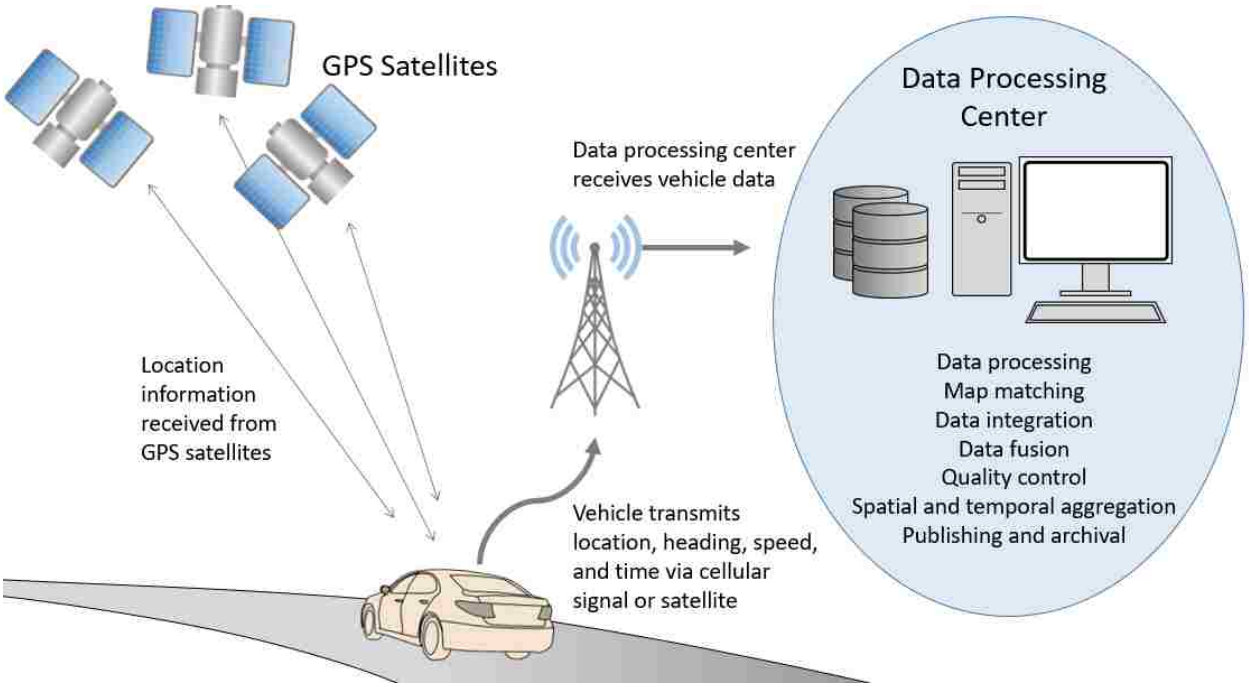

Figure 2-6: Conceptual Illustration of Probe Vehicle Data Collection Process

The population of vehicles that contribute to a dataset represents some sample of the total vehicle population, which may or may not be representative in terms of average speed and/or driving behavior. Further, a variety of GPS and communications devices and vehicle subpopulations are typically represented in this sample, each of which may deliver vehicle state updates at different frequencies (i.e. different sampling rates), and may represent different subpopulations. Thus, there are multiple sampling-related mechanisms that can contribute to bias in probe vehicle data. First, slower moving vehicles are more likely to be observed, and will produce a greater number of samples, all else being equal (Jenelius and Koutsopoulos 2014). As a result, slower moving vehicles may be overrepresented in the observed data, and more congested time periods will be less likely to produce missing observations (Hallenbeck and McCormack 2015; Bitar 2016). Second, it is likely that certain subsets of the overall vehicle population are more likely to contribute to the dataset (i.e. higher penetration rate), which can lead to overrepresentation and bias (Patire et al. 2015). For example, taxis and delivery vehicles are more likely to use some form

of GPS than personal vehicles, and are also likely to differ in their driving behavior and average speed. Finally, if different subsets of the contributing vehicle population are sampled more frequently, the driving profile of this subpopulation will be overrepresented. For example, if delivery fleet vehicles were sampled every second, while the average sampling interval over the entire contributing vehicle population is 30 seconds, the simple mean speed would over-represent delivery vehicles. Though there is significant interaction with the methods used to aggregate point measurements into link-level traffic observations, different combinations of these factors will contribute to different levels of bias. In fact, bias can be seen to vary with traffic conditions, time of day, and location in previous probe vehicle data validation work (I-95 Corridor Coalition 2018).

2.2.3 *Eulerian vs. Lagrangian View*

It is important to note, when comparing probe vehicles to fixed, static sensors as sources of transportation data, the distinction between Lagrangian and Eulerian specifications of traffic flow (Work and Bayen 2008). Fixed sensors observe traffic flow and speed as a time varying process at a fixed location, or an Eulerian view of traffic flow. GPS traces follow individual vehicles as they move through time and space, resulting in a Lagrangian view of traffic flow. Previous work has provided a number of methods for uniting these two perspectives, in order to combine fixed mechanical sensor and probe vehicle data in characterizing traffic flow and to describe the correspondence between Lagrangian and Eulerian representations of existing analytical models of traffic flow, among other things (Matsukidaira and Nishinari 2003; Xia et al. 2017; Yuan et al. 2012).

Though much of the established body of research in traffic flow theory is based on the Eulerian perspective (Duret and Yuan 2017), the Eulerian and Lagrangian perspectives can describe the same underlying physical processes and support the same mathematical conclusions.

However, as a data collection paradigm, both have different strengths and weaknesses. Fixed sensors can provide a complete and unbiased view of traffic state, but do so at fixed point locations. Probe vehicles can provide a more complete picture of the speed profile along a roadway, but come with bias and sparsity challenges. Among these two, probe vehicle data is the only one that is likely to provide full coverage of the road network in the near future at reasonable cost (Hofleitner et al. 2012; Treiber and Kesting 2013). Additionally, the Lagrangian perspective allows the specification of ad hoc trips and travel time estimation based on these trips. Rather than attempting to construct travel times out of point speed measures, often ignoring control delays and slowdowns, with probe vehicle data it is possible to estimate the true travel time as experienced by a vehicle on the road. Due to these and other benefits, probe vehicle data can be expected to constitute an increasing share of the traffic data used in transportation analysis, management, and planning in the future.

2.3 Research Objectives

2.3.1 *Overview*

The objectives of this research are as follows:

- To develop a mathematical framework to describe the mechanisms influencing sample size, missingness, and bias in probe vehicle data. This framework will provide a strong foundation for improving the quality and fidelity of this important source of traffic data.
- Demonstrate how this framework can be implemented analytically and using Monte Carlo simulation
- To validate the proposed framework using microscopic traffic simulation.
- To evaluate and illustrate the impact of different traffic and sampling conditions on sample size, sampling bias, and completeness.

- To build on the proposed framework by applying it to foundational quality and completeness challenges in probe vehicle data including:
    - Estimating sample size and effective sample size (in terms of vehicle and sample count);
    - Computing expected missing data rates under different conditions;
    - Estimating sampling bias under different conditions;
    - Conducting planning level analysis for future probe vehicle data collection efforts;
    - Develop and demonstrate methods to address sampling bias and completeness

2.3.2 *Expected Benefits*

- A more complete understanding of the relationship between traffic state, vehicle type and driving behavior distribution, sampling parameters, and probe vehicle data bias and missingness. This understanding will support informed application of existing probe vehicle data to engineering problems, as well as ongoing work in improving the data collection and quality control process.
- A mathematical model for bias and missing data that can support the development of methods for unbiased data imputation, bias correction, and other quality improvement methodologies.
- A quantitative framework for analysis supporting future probe vehicle and connected vehicle data collection efforts. This framework, and illustrative example provided, will be a useful tool for designing future experiments and assessing the data quality implications of different approaches.

2.4 Study Scope

This research is oriented toward addressing a set of data quality challenges unique to a particular source of data and application scenario. Specifically, it focusses on GPS-based probe vehicle data and the generation of discrete time road link level speed or travel time data. Additionally, this work largely deals with uninterrupted traffic flow, where traffic dynamics are primarily driven by vehicle-vehicle interactions and roadway geometry rather than traffic control systems. While many of the same principles will be at work on signalized facilities, a great deal of additional complexity is introduced that is beyond the scope of this work. That said, the work described here provides a suitable foundation on which to develop models describing a wide range of scenarios, including signalized arterials, non-motorized travel, and others.

There are many factors that can complicate the process of obtaining GPS data from vehicles and turning into a form that accurately represents the traffic state and is useful for transportation analysis. GPS error/accuracy is one such factor, along with communications reliability, application design and performance, and others. The work described here strictly deals with the sampling and data aggregation processes and resulting impacts on the quality and completeness of traffic observations.

2.5 Dissertation Organization

This work described in this dissertation has three major components: 1) the development of a mathematical framework for describing sample size, missing data rates, and sampling bias in mobile location-based traffic data; 2) the development of a simulation model and set of methodologies for applying the proposed framework in estimating sample size, missing data rates, and sampling bias, as well as methods addressing sampling bias; and 3) The design and execution

of several case studies demonstrating the utility of the proposed framework in improving the quality of probe vehicle data and planning data collection efforts.

The remainder of this dissertation is structured as follows: Chapter 3 provides an overview of the state of the art on probe vehicle speed estimation, data quality and bias, and missing data theory and mechanisms. In Chapter 4 the mathematical framework for describing the probe vehicle data sampling process is developed, as well as implementation details. Chapter 5 describes the development of a simulation model and validation of the proposed framework. Chapter 6 offers three case studies showing how the proposed mathematical framework and methods can be used in assessing and improving the quality of probe vehicle data, including the development of a methodology for addressing sampling bias.

Chapter 3: State of the Art

3.1 Missing Data

3.1.1 *Completeness of Probe Vehicle Data*

As for loop detectors and other traffic sensors, missingness is an important quality issue in probe vehicle data. Though the completeness of commercial probe vehicle data has improved over the last decade (Hosuri 2017), it is clear that the primary reason for missing data is the lack of a contributing vehicle on a road segment during a given time interval (Bitar 2016). This means that the probability of data being missing is correlated with both travel time and traffic volume, which suggests that bias can result if missing data is not addressed in a principled way. (Cambridge Systematics and Texas Transportation Institute 2015) analyzed a subset of the National Performance Monitoring Dataset (NPMRDS) over ten eastern states in 2014. They found that interstate highways were 58% complete, while the average completeness for all other road classes was 22%. (S. Kim and Coifman 2014) analyzed a probe vehicle dataset from INRIX on a 14 mile corridor in 2014 and found that, though there were no missing observations (due to processes applied by the provider), there were a great number of repeated measurements over multiple time intervals. This indicates that, at least at the level of temporal aggregation the data was reported, completeness was well below 100%.

Some previous work has investigated the minimum penetration rates required to achieve an acceptable level of completeness. For example, (Boyce, Hicks, and Sen 1991) based their computations on the requirement of having at least one contributing vehicle present on a certain fraction of road segments within a given observation interval, ignoring sampling rate. (Xiaowen Dai, Ferman, and Roesser 2003) develop a simulation model to estimate the penetration rate

needed to achieve fixed levels of accuracy and coverage. This work looked at the influence of observation interval length, but considers only a single sampling rate of 1 sample / 10 seconds. (M. A. Ferman, Blumenfeld, and Dai 2003; M. Ferman, Blumenfeld, and Dai 2005) investigate the relationship between data completeness, relative error of measurement, and reporting interval (not sampling rate). This work represents the only literature identified that, like this work, attempts to develop an analytical formula for data completeness. However, this paper assumes a single distribution of vehicle speeds and a single fixed sampling rate, and ignores the influence of sampling rate on completeness and speed estimation.

3.1.2 *Missing Data Mechanisms*

Much of the body of statistical knowledge on the theory and methods for address missing data were developed for/by psychologists and other scientists who frequently administer surveys (Graham 2009; Rubin 1976; Schafer and Graham 2002). One can consider the probability of a given observation being missing and resulting pattern of missingness as arising from some causal mechanism that may or may not be related to the quantities of interest. In statistical terms, missing data mechanisms are defined as falling into one of three categories (R. Little and Rubin 2002; Rubin 1976). First, "missing completely at random" (MCAR) is a special case of MAR, and refers to situations in which the probability of missingness is independent of both the observed and unobserved values. In practical terms, this means that the observed data is a random sample of the complete data (Raghunathan 2004). "Missing at random" (MAR), describes a situation in which the probability of missingness is independent of the unobserved data conditioned on the observed data. Finally, if data is not MAR or MCAR, it is referred to as "missing not at random" (MNAR) (Graham 2009). MNAR is a non-ignorable missing data pattern, which means that the mechanisms driving missingness cannot be ignored in the methods used for imputation and/or analysis (R. Little

and Rubin 2002; Raghunathan 2004). From this, it is clear that the missing data mechanism can lead to systematic differences between the distributions of the observed and missing data which must be addressed in any missing data treatment (Graham 2009).

Although a variety of methods exist to handle missing data, including but not limited to deleting records with missing values and single and multiple imputation, missing data remains a challenging issue (Schafer and Graham 2002). Specifically, White, Higgins, & Wood, (2008) note that validity of experimental results can be affected when the missing samples are not representative of the population of interest. Further, if analysis methods are used that consider missing data to be MAR or MCAR when neither of these two assumptions describe the missing data pattern, there is strong possibility of non-response bias (i.e., the bias introduced when samples not reporting data have some fundamentally different characteristic(s) than samples reporting data). In addition to non-response bias, another issue associated with missing data is decreased statistical precision and power in cases where missing values exist over multiple variables in a dataset, which may result in researchers making the choice to leave out a significant fraction of all observations (Sterne et al. 2009). Depending on the techniques used to account for missing data in statistical models, bias in model parameter estimates, standard errors, and hence values of test statistics can occur (Allison 2003; Jones 1996; Glasser 1964).

3.2 Sampling and Aggregation Methods

In addition to sampling method and parameters, the method(s) used to aggregate point data from multiple vehicles on a road segment to estimate traffic state have a considerable impact on the resulting accuracy and bias. (Kong et al. 2013) compare a curve fitting method with a vehicle tracking method for estimation of mean link speed, and show that the vehicle tracking method generally produces more accurate results. Though it is not entirely clear in this paper, the superior

performance of the vehicle tracking method seems to be in part due to calculation of individual vehicle trace speeds prior to computing mean link speeds, which makes sense because it would be less likely to overweight slower or more frequently sampled vehicles. (Zhang, Xu, and Liao 2013) introduce a time decaying weighted average schemes which take each GPS record and weights it according to how recently it arrived, and suggest a weighted resampling approach for estimating the current mean speed on a road segment. This method does not acknowledge sampling bias, but in theory could reduce the bias associated with differing sampling rates by assigning the largest weight to the most recent few samples. (Zheng and Van Zuylen 2013) propose a neural network model for link-level travel time estimation using sparse probe vehicle data. However, this method requires ground truth travel time data for model training, which is not likely to be available in most cases.

A significant amount of previous work has assumed that samples are obtained at very high frequency (e.g. 1-3 seconds / sample) or ignored the sampling frequency issue altogether (D'Este, Zito, and Taylor 1999; Long Cheu, Xie, and Lee 2002; Herrera et al. 2010). Under such conditions, the influence of aggregation method on the accuracy of the link-level traffic speed measurements will be minimal. However, there is significant variation in the sampling frequencies in existing probe vehicle data sources, and so there is certainly some interplay between the aggregation method and the resulting accuracy. Some previous work has introduced alternative speed measurement schemes which obviate the need for an explicit aggregation step. For example, (Qing Ou et al. 2011) describes an inflow/outflow model for speed estimation that treats the aggregate speed as a random variable to be estimated from observed flow dynamics.

3.3 Bias and Accuracy

Missing data can certainly increase the complexity of working with probe vehicle data, for example, consider the need to conduct accurate imputation before analysis or before traffic conditions can be reported to the public. However, at the core of the issue is understanding the level of inaccuracy, bias, and uncertainty that may be introduced through the sampling mechanism. If one can quantify the bias and/or uncertainty associated with a given estimate, more informed decisions can be made regarding the development and application of a statistical model. However, making such an estimate requires a firm understanding of the relationship between sampling methods and the accuracy of the observed data.

3.3.1 *Accuracy*

Very little previous work was found which investigated the relationship between sampling rate, penetration rate, and bias / accuracy and completeness in probe vehicle data. (Long Cheu, Xie, and Lee 2002) applied microscopic traffic simulation to study the relationship between speed estimation accuracy and probe vehicle penetration rate. They conclude that 4-5% probe vehicle penetration rate should be sufficient to achieve 5 km/hour in speed accuracy 95% of the time. However, this study did not investigate the impact of sampling frequency, and their results implicitly assume that probe vehicles will be sampled continuously. Similarly, (Cetin, List, and Zhou 2005) applied microscopic simulation to study the impact of penetration rates on travel time estimation accuracy in a mixed arterial and expressway network, but again no mention was made of sampling frequency. (S.M. Turner and Holdener 1995) investigated the minimum probe vehicle sample sizes required to achieve fixed level of accuracy for real-time traffic information, but their study was based on automatic vehicle identification rather than GPS data. It is worth noting,

however, that this work provides some insight into the minimum sample size (assuming systematic bias is not present or has been dealt with) required to achieve an acceptable level of accuracy. (Bucknell and Herrera 2014) investigated the impact of both penetration rate and sampling frequency on speed estimation accuracy using the NGSIM dataset. They conclude that both are closely related to estimation accuracy, and that the relative influence depends on the methods used to compute speed as well as the penetration rate / sampling frequency regime. Using simulation, (Xiaowen Dai, Ferman, and Roesser 2003) showed that both higher penetration rates and longer observation intervals are generally associated with better accuracy, though differing sampling rates and mixed speed distributions were not investigated.

Other studies have investigated the accuracy of probe vehicle data and its relationship to sampling mechanism from the perspective of the reliability of the resulting analysis. For example, Srinivasan & Jovanis (1996) developed an algorithm to determine the probe vehicle sample size needed to obtain certain levels of travel time reliability. They determined that the number of probe vehicles needed increases with prescribed reliability level and decreasing length of time period over which travel times are measured. Additionally, Long Cheu, Xie, & Lee (2002) studied the number of probe vehicles needed to obtain reliable travel time measurements for urban arterials.

Several studies have investigated data fusion methods to combine mechanical traffic sensing data (e.g. loop detectors) with probe vehicle data to improve the accuracy of speed and travel time estimation. Such work as typically either a) ignored sampling frequency altogether (Nanthawichit, Nakatsuji, and Suzuki 2003) or b) assumed homogeneous sampling frequencies and ignored the impact on estimation accuracy (Liu et al. 2016). One exception can be found in (Patire et al. 2015), which applied a Kalman filtering algorithm for data fusion using two different real-world GPS probe vehicle datasets. However, the two datasets had significantly different probe

vehicle penetration rates, and so provided no solid conclusions regarding the influence of sampling frequency on estimation accuracy.

It should be noted that there are a number of factors unrelated to sampling that influence accuracy. For example, some previous work has studied latency, and showed that significant latency can be observed in most common probe vehicle datasets (Z. Wang et al. 2018). Additionally, (Bitar 2016) discussed the impact of travel time reporting quantization on speed estimation accuracy. That is, if travel times are reported in integer seconds (or at low fixed precision), the error of speed estimates will increase at lower travel times. Other issues, such as GPS accuracy and/or urban canyon effect clearly have some impact on accuracy and completeness.

3.3.2 *Bias*

Sampling bias is defined as a case where a sample statistic is not an accurate representation of the underlying parameter in the target population (McCutcheon 2011), and has long been dealt with in a wide range of statistical applications including econometrics (Heckman 1977), medicine (Victor et al. 2004), sociology (Berk 1983; Winship and Mare 1992), and natural science (Driscoll et al. 2012). Various methods have been proposed to address sampling bias, and the selection of methods is largely driven by an understanding of the mechanism driving the bias. For example, inverse probability weighting is often applied when different subsets of a population of interest is sampled unequal selection probabilities (Mansournia and Altman 2016; Seaman and White 2013). Bias that arises from a missing data pattern can often be addressed by applying multiple imputation before completing statistical analysis, which can generally lead to unbiased imputation under MAR missing data mechanisms (R. Little and Rubin 2002; Schafer 1997).

In probe vehicle data, the bias that arises is driven by multiple interacting mechanisms. First, different subpopulations have different penetration rates of contributing vehicles, which can

be interpreted as a weighted sampling mechanism. The extent to which this introduces bias depends primarily on the similarity of the different subpopulations, but in general it violates the assumption of random sampling. The second is that slower moving, and more frequently sampled vehicles produce more samples and are more likely to be observed. This means that a) the quantity of interest (usually speed or travel time) is related to probability of being observed, and b) another form of subpopulation weighting is introduced. The result is that the observed data is potentially biased with respect to the true population statistics in a given time period (Bitar 2016; Hallenbeck and McCormack 2015), and there is a potentially significant difference between the observed and missing data (Henrickson and Wang 2016).

(Jenelius and Koutsopoulos 2014) analyzed the influence of probe vehicle sampling protocols, and showed that time-based sampling (as compared to space-based) is more likely to bias the resulting data toward slower moving vehicles and road segments with lower travel speeds, and making road segments with faster moving traffic less likely to be observed. This bias issue has been identified in other literature, including (Dion, Robinson, and Oh 2011; Lattimer and Glotzbach 2012; S. Kim and Coifman 2014). One of the primary factors that contributes to bias is the fact that slower moving vehicles spend more time on each road section and, at a given time-based sampling frequency, will be more likely to be observed and produce more observations (Hallenbeck and McCormack 2015). The result is that, under time-based sampling, slower moving vehicles are oversampled compared to the population fraction they represent. Despite this, time-based sampling is the primary protocol for commercial probe vehicle data.

The bias attributable to the sampling mechanism in probe vehicle data is not always apparent in aggregate measures, such as overall mean speed error. For example, in the most recent validation work completed as part of the I-95 Corridor Coalition Vehicle Probe Project, results

showed a fairly consistent negative bias across all three major data providers (HERE, INRIX, and TOMTOM) at higher traffic speeds (which represents the majority of the data) but a somewhat variable positive bias at lower speeds (I-95 Corridor Coalition 2018). This could represent data smoothing efforts completed by the vendors, greater likelihood of very low speed data being flagged as anomalous by the vendor quality control algorithms, or a variety of other possibilities. In any case, this result suggests significant interaction between the multiple mechanisms contributing to bias, some of which might tend to cancel each other out. The extent to which this occurs is essentially random unless the causal mechanisms are explicitly considered in subsequent data processing and analysis.

Chapter 4: Model Development

The purpose of the methodology presented in this section is to described the relationship between the true underlying traffic state, the sampling parameters, and the observed quantities in a probe vehicle-based data collection process. Specifically, this section develops a framework to estimate statistical measures describing the sample size, missing data rate, the observed speed based on the traffic volume, sampling frequency distribution, instrumented vehicle fraction, and true speed distribution. This constitutes the core contribution of this work and, by starting from basic physical principles, represents a significant departure from previous efforts in this domain.

At the most basic level, the approach taken here is similar to that of a queuing problem. However, the quantities of interest are not among those typically sought in queuing theory. Queuing theory often deals with optimizing a queuing system to achieve an optimal balance of cost and service quality or assessing the performance of a system, with a focus on service time, waiting time, server utilization, and other quantities related to the operational objectives. Here, we are interested in describing aggregate quantities related to data quality rather than operational performance measures, but the analytical approach is similar to that of a queuing problem. Thus, in addition to operational parameters such as arriving volume, travel time, etc., parameters related to the observation process such as probe vehicle sampling rate and penetration rate are considered to estimate what will actually be observed in a typical probe vehicle dataset.

Section 4.1 describes the framework in general terms, without relying on distributional assumptions. Section 4.2 presents the framework described in Section 4.1 in terms of specific statistical distributions for the instrumented vehicle arrival rate and true traffic speed. Section 4.3 describes the framework as a step-by-step procedure for estimating the observed quantities given the true traffic parameter distributions, sampling parameters, and roadway geometry.

In the following discussion, an *observation interval* refers to the time period over which collected probe vehicle data points are aggregated. That is, an observation interval of 5 minutes indicates that all data points are aggregated to the 5-minute level and delivered to the end user as a unique traffic observation for each road segment and 5-minute time period. A *sampling interval* refers to the time period between subsequent GPS updates. A vehicle on a sampling interval of 30 seconds will report its current location and other state information every 30 seconds.

4.1 General Case

The objective of this section is to show, in general terms, how expressions for vehicle count, sample count, observed speed, and associated statistical measures can be devised based on the probe vehicle sampling process. This framework will be described with respect to specific statistical distributions in subsequent sections. Here, the focus is on the basic physical and statistical principles, with few assumptions regarding the underlying statistical distributions that may describe the quantities of interest.

4.1.1 *Sample Count Distribution*

The first step is to describe the distribution of the number of samples obtained, and the corresponding probability of an observation being missing, as a function of travel time across a road segment (*TT*), the observation interval (*TI*), and the sampling interval (*sr*). The distribution of the number of samples obtained on a road segment by a single vehicle within a single observation interval will be a function of the time spent on the segment during an observation interval and the sampling interval *sr*. Clearly, the time a vehicle spends on a road segment will be the travel time across the segment, *TT*. However, in order to describe the distribution of interest, it is necessary to compute the time spent on the segment *during* an observation interval, which is a

deterministic function of the time of arrival and the travel time. If the time of arrival is referenced to the start of a given observation interval, vehicles arriving at or before *–TT* will spend no time on the segment during the interval because they will have left the road segment as the observation interval starts. For every second later than *–TT* a vehicle arrives, a vehicle will spend an additional second on the road segment up to the maximum value, which is the minimum of *TT* and the length of the observation interval *TI*. Finally, for vehicles arriving within *min(TT,TI)* of the end of the observation interval, their time spent on the segment will be reduced by one second for every second later they arrive until *TI*, after which they will spend no time on the segment. This results in the trapezoidal function shown in Equation 6.1 and illustrated in Figure 4-1.

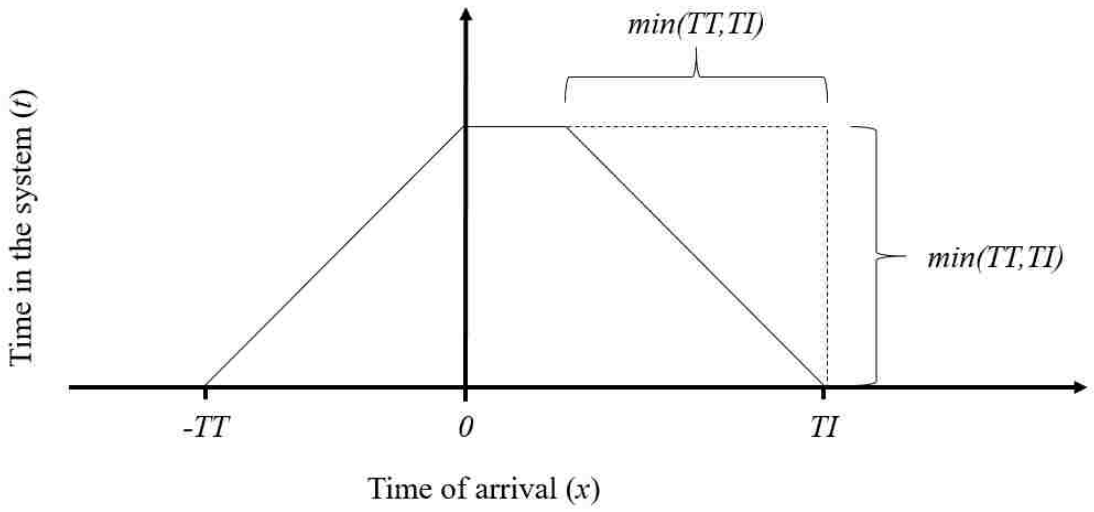

Figure 4-1: Trapezoidal function forTime spent in the system

$$t(x,TT) = \begin{cases} TT + x & if -TT \le x < \min(TI,TT) - TT \\ \min(TI,TT) & if \min(TI,TT) - TT \le x < TI - \min(TI,TT) \\ TI - x & if\ TI - \min(TI,TT) \le x < TI \\ 0 & otherwise \end{cases} \quad (4.1)$$

For a given arrival time with fixed *sr* and *TT*, the timing of the first sample is the only source of randomness, and there are only two possible sample counts. The two possible sample counts are the minimum number and 1 plus the minimum number, and the difference between these two is

determined by the timing of the samples. For example, if the value of $t(x, TT)$ is 15 seconds at some *x*, and *sr = 30*, a vehicle will be produce 1 or 0 samples, each with probability 0.5. If the timing of the first sample is within the 15 second window defined by $t(x, TT)$, the vehicle will be produce 1 sample and 0 otherwise.

To understand the form of Equation 4.1, consider that the value of the function $t(x, TT)$ divided by *sr* represents the expected number of samples for a given arrival time. Or, in different terms, it represents the minimum number of samples plus the probability of 1 plus the minimum number of samples. With this interpretation, an intuition can be gained regarding the form of Equation 4.1. Note the three arrival times x_1, x_2, and x_3 in Figure 4-2. For Point x_1, the minimum number of arrivals is 0 and the maximum is 1. The probability of 1 sample at point x_1 is simply $t(x_1, TT) \ / \ sr$, and the probability of zero samples is. $1 - t(x_1, TT) \ / \ sr$. Similarly for point x_3, the minimum number of samples is 2 and the maximum is 3. The probability of 2 samples at this point is $(t(x_3, TT) - 2 \times sr)/sr$, and the probability of 3 samples is $1 - (t(x_3, TT) - 2 \times sr)/sr$. This is shown in functional form in Equation 4.2.

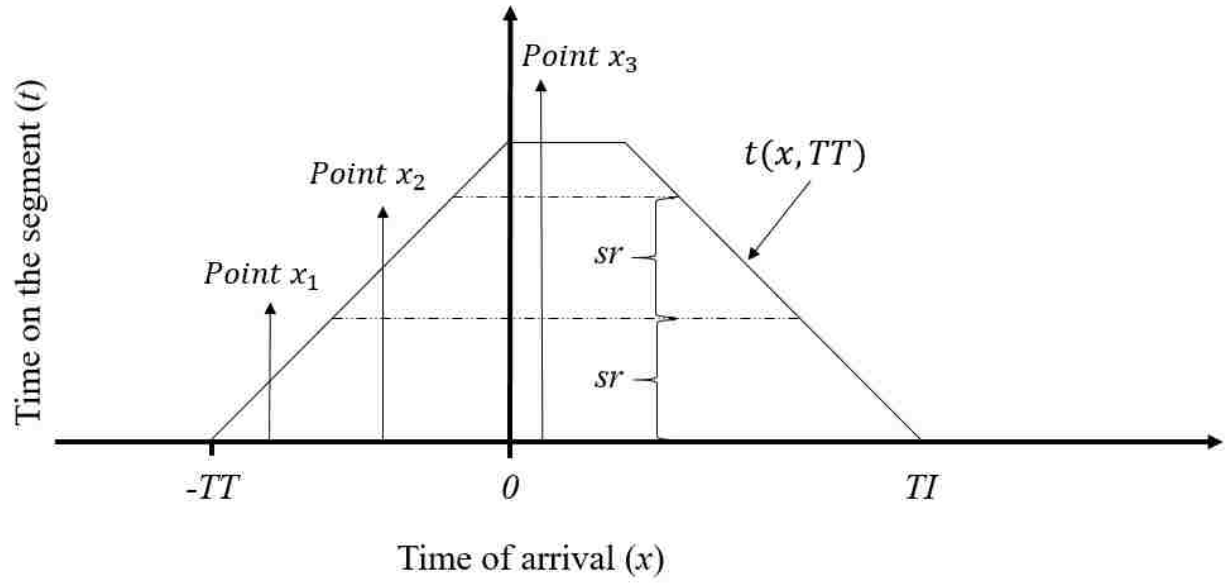

Figure 4-2: Sample Count Distribution by Time of Arrival

$$P(nspv = j|x, TT, sr) = \begin{cases} \frac{t(x,TT)}{sr} - floor\left(\frac{t(x,TT)}{sr}\right) & if\ j = ceil\left(\frac{t(x,TT)}{sr}\right) \\ 1 - \frac{t(x,TT)}{sr} + floor\left(\frac{t(x,TT)}{sr}\right) & if\ j = floor\left(\frac{t(x,TT)}{sr}\right) \\ 0 & else \end{cases} \quad (4.2)$$

Because vehicle subpopulations are defined here in terms of unique combinations of speed or travel time distribution parameters and sampling interval, the objective is to express the distribution of the number of samples per vehicle conditioned only on these quantities. Thus, the categorical distribution describing the number of samples per vehicle conditioned only on travel time and sampling interval can be found by integrating Equation 4.2 over the time of arrival of as shown in Equation 4.3, where $s_X(x)$ is the PDF for the time of arrival *x*. The exact form of the integral in Equation 4.3 will depend on the form of the distribution $s_X(x)$ (more details on this in Section 4.2). Furthermore, the probability of at least 1 sample (or 1 – the probability of 0 samples) is the expression $\min\big(sr, t(x,TT)\big)/sr$ integrated over *x*.

$$P(nspv = j|TT, sr) = \int_x P(nspv = j|x, TT, sr) s_X(x) dx \quad (4.3)$$

Allowing contributing vehicles to be drawn from a mixture of $i \in \{1,2,..,k\}$ subpopulations, each with a mean travel speed probability density function $f_{V_i}(v)$, and sampling interval sr_i, we get the expression shown as Equation 4.4 for the distribution of the number of samples per vehicles for each subpopulation *i*. Note $TT(v)$ indicates the travel time as a function of mean vehicle speed v.

$$P(nspv_i = j|sr_i) = \int_v P(nspv_i = j|TT(v), sr_i) f_{V_i}(v) dv \quad (4.4)$$

Note that the distribution of $nspv_i$ conditioned only on sr_i is identical for all arriving vehicles in subpopulation *i*. Or, similarly, an expression for the expected number of samples per vehicle for subpopulation *I* as shown in Equation 4.5.

$$E(nspv_i|sr_i) = \int_v E(nspv_i|TT(v), sr_i) f_{V_i}(v) dv \quad (4.5)$$

A general count distribution $q_{N_i}(n)$ is assumed for the number of vehicles within a subpopulation appearing during an observation interval and $E(n_i)$ is the expected value. Equation 4.6 shows the distribution of the total number of samples per observation interval for subpopulation *i*. In this expression, $C_{ns_i,n}$ indicates the weak composition of ns_i, or the set of all possible integer tuples of length *n* which sum to ns_i. In words, this expression is the sum of probabilities over every possible number of arriving vehicles and sample counts per vehicle that produces ns_i samples.

$$P(ns_i|sr_i) = \sum_n \left[q_{N_i}(n) \sum_{m \in C_{ns_i,n}} \left(\prod_{j \in m} P(nspv_i = j|sr_i) \right) \right] \quad (4.6)$$

More simply, the expected value of ns_i can be computed as the product of the expected value for vehicle count n_i and the expected value for the number of samples per vehicle $nspv_i$ as shown in Equation 4.7.

$$E(ns_i|sr_i) = E(n_i)E(nspv_i|sr_i) \quad (4.7)$$

The distribution of the total number of samples ns_{all}, or the sum over $i \in \{1,2,\dots,k\}$ of ns_i can be expressed as shown in Equation 4.8.

$$P(ns_{all}|sr) = \sum_{c \in C_{ns_{all},k}} \left[\prod_{i \in \{1,2,\dots,k\}} P(ns_i = c_i|sr_i) \right] \quad (4.8)$$

With a corresponding expected value as shown in Equation 4.9 .

$$E(ns_{all}|sr) = \sum_{i \in \{1,2,\dots,k\}} E(ns_i|sr_i) \quad (4.9)$$

Finally, the probability of no vehicles being observed (i.e. a missing observation) will be the product over all subpopulations of the probability of no vehicles being observed within each subpopulation. A general count model of the form $q_{N_i}(n)$ is assumed for the number of vehicles *n* appearing on a road segment during an observation interval for a single subpopulation *i*. The probability of observing no vehicles in an observation interval for a single subpopulation is the sum over all possible number of vehicles appearing for subpopulation *i* multiplied by the probability that all of the appearing vehicles produce zero samples as shown in Equation 4.10. The

probability of observing no vehicles over all subpopulations is the product of this expression over all subpopulations, as shown in Equation 4.11. In other words, it is the probability that the sum of ns_i over all subpopulations is equal to zero.

$$P(R_i = 0) = \sum_{n\in\{0,1,2,\ldots,\infty\}} q_{N_i}(n)\left(P(nspv_i = 0|sr_i)\right)^n \quad (4.10)$$

$$p(R = 0) = \prod_i [p(R_i = 0)] \quad (4.11)$$

This formulation assumes that, if a vehicle does not report its state at least once on a road segment, it will not be considered to have been observed on that segment. This would be the case for any point-wise speed estimation method. However, it is also possible to compute speed in terms of update point pairs, in which case any vehicle which appears on a road segment during an observation interval will be observed. For this scenario, the vehicle count and probability of missingness is strictly determined by the speed of the vehicle and the count distribution $q_{N_i}(n)$, as will be discussed in Section 4.2.

4.1.2 *Observed Speed Distribution*

The observed speed for a given subpopulation depends to a great extent on the method that is used to combine individual vehicle state updates to aggregate speed measures for each road segment and observation interval. For example, it is intuitively clear that taking the simple mean speed over all sampled speeds will give different results in general from first aggregating the speed for each vehicle and then taking the mean over all vehicle speeds. Correspondingly, one might calculate a distance-weighted mean of travel speed between GPS point pairs, which would produce different results from point speed-based measures. The three methods considered in this work are listed below:

- Vehicle-wise mean point speed
- Sample-wise mean point speed

- Distance weighted mean speed between point pairs

We begin with the vehicle-wise mean point speed, because sampling-based methods will the most efficient to estimate the observed speed for the sample-wise mean speed case (sampling methods will be discussed in Section 4.3.3). The expected value of the observed speed for a single subpopulation can be expressed as shown in Equation 4.12. This is the conventional expected value formula, or the integral over mean vehicle speed of the product of speed, the probability density function for speed, and the probability of being observed, divided by a normalizing term.

$$E\left(v_{obs,i}\middle|sr_i\right) = \frac{\int_v v \times f_{V_i}(v) \times P(nspv_i > 0|sr_i, TT(v))dv}{\int_v f_{V_i}(v) \times p(nspv_i > 0|sr_i, TT(v))dv} \quad (4.12)$$

Estimating the expected observed speed overall subpopulations requires the population weighting factor shown in Equation 4.13. This expression represents the expected number of vehicles in a given subpopulation which are observed within a time interval. As previously noted, $E(n_i)$ is the expected number of vehicles appearing on the road segment during a single observation interval.

$$M_i = E(n_i) \times P(nspv_i > 0|sr_i) \quad (4.13)$$

The expected mean speed over all subpopulations can be computed as shown in Equation 4.14. This formula represents the weighted average of the expected observed speeds, weighted by the expected number of observed vehicles, over all subpopulations. In other words, this can be interpreted as the expected value formula for a mixture distribution, where the mixture weights are equal to $\frac{M_i}{\sum_{i\in\{1,2,\dots,k\}} M_i}$. To understand this representation, consider that the mixture weights represent the long-run population fraction for each subpopulation, such that the aggregate speed in each observation interval represents a random sample from a *k*-component mixture distribution.

$$E(\bar{v}_{obs}|sr) = \frac{\sum_{i\in\{1,2,\dots,k\}} M_i \times E\left(v_{obs,i}\middle|sr_i\right)}{\sum_{i\in\{1,2,\dots,k\}} M_i} \quad (4.14)$$

To compute the variance for the observed travel time, it is no longer sufficient to consider the distribution of the mean vehicle speed across a road segment. This is because, especially at low penetration rates and/or infrequent sampling, the observed speed will be based on very few point-wise observations. The measured variance, then, will reflect the variability in speed across the segment for individual vehicles as well as the between-vehicle variance in mean travel time. The Law of Total Variance can be used to combine these two sources of variability, but it requires estimates of the variance of the mean as well as the variance of the observations given the mean. In words, the Equation 4.15 (from the law of total variance (Wooldridge 2010)) is the sum of the expected value of the observed conditional variance and the variance in the conditional expected value.

$$Var\left(v_{obs,i}|sr_i\right) = E_v(Var\left(v_{obs,i}\middle|v, sr_i\right)) + Var_v(E\left(v_{obs,i}\middle|v, sr_i\right)) \qquad (4.15)$$

Provided some estimate of the variance of a single vehicle's speed over a road segment $\sigma_{v_i}^2$ within subpopulation *i*, the observed variance conditioned on the mean will be a function of the number of samples obtained. That is, using the formula for the variance of the mean, the observed variance can be described as the true variance divided by the sample size. Summing the product of the probability of each sample size and the corresponding variance of the mean over all possible sample sizes gives the formula shown in Equation 4.16. Note that the interest is only in the observed data, so zero count observations are excluded. Because of this, the normalizing term is included to insure that the probabilities sum to 1.

$$Var\left(v_{obs,i}\middle|v, sr_i\right) = \frac{\sum_{j\in\{1,2,\dots\infty\}} P(nspv_i=j|sr_i,TT(v))\frac{\sigma_{v_i}^2}{j}}{\sum_{j\in\{1,2,\dots\infty\}} P(nspv_i=j|sr_i,TT(v))} \qquad (4.16)$$

The expectation of this variance can be found by taking the integral over *v* of Equation 4.16, where $\sigma_{v_i}^2$ is the variance of a single vehicle's speed over the road segment (ignoring the between-vehicle

variance of the mean) and $\sigma_{v_i}^2/j$ is the variance of the sample mean when *j* samples are obtained. This is shown in Equation 4.17.

$$E_v(Var(v_{obs,i}|v, sr_i)) = \int_v \frac{\sum_{j\in\{1,2,\dots\infty\}} p(nspv_i=j|sr_i,TT(v))\frac{\sigma_{v_i}^2}{j}}{\sum_{j\in\{1,2,\dots\infty\}} p(nspv_i=j|sr_i,TT(v))} \times f_{V_i}(v)dv \quad (4.17)$$

The variance in the expected value of v_{obs}, can be found using the conventional formula for variance. This expression, shown in Equation 4.18, is the integral of the squared difference between the mean vehicle speed and the expected mean vehicle speed multiplied by the product of the speed PDF and the probability that at least one sample point is obtained. Again, a normalizing constant is included to insure that the probability term sums to 1.

$$Var_v(E(v_{obs,i}|v, sr_i)) = \frac{\int_v (v-E(v))^2 \times f_{V_i}(v) \times P(nspv_i>0|sr_i,TT(v))dv}{\int_v f_{V_i}(v) \times P(nspv_i>0|sr_i,TT(v))dv} \quad (4.18)$$

This provides a useful expression for the observed variance between vehicles, but a single observation represents the mean over all vehicles within a subpopulation observed in a single observation interval. To compute the true observed variance, the formulas shown in Equation 4.19 and 4.20 are used. In words, the weighting term $W_i(j)$ represents the un-normalized probability that *j* vehicles are observed in an observation interval. Inside the summation term is the probability that *n* vehicles arrived, and that *j* were observed. This is represented by the product of the arriving vehicle count PMF evaluated at *n* and the binomial expression for exactly *j* out of *n* being observed. The binomial distribution is appropriate to describe the distribution of *j* because the probability of *j* observed vehicles out of *n* total vehicles arises from a series of *n* binary outcomes with fixed probability for a given subpopulation.

$$W_i(j) = \sum_{n\in\{j,j+1,\dots,\infty\}} \left[q_{N_i}(n) \times \binom{j}{n} \times P(nspv_i > 0|sr_i)^j P(nspv_i = 0|sr_i)^{n-j}\right] \quad (4.19)$$

$$Var\left(\bar{v}_{obs,i}|sr_i\right) = \frac{\sum_{j\in\{1,2,\ldots,\infty\}} \frac{W_i(j)}{j} \times Var(v_{obs,i}|sr_i)}{\sum_{j\in\{1,2,\ldots,\infty\}} W_i(j)} \quad (4.20)$$

The final $Var\left(\bar{v}_{obs,i}|sr_i\right)$ indicates the observed variance in mean travel speed for subpopulation *i*. An expression for the combined variance of all vehicle subpopulations can be derived, but it is computationally intensive because it requires enumerating all possible combinations of arriving vehicle counts.

For the distance-weighted mean pair-wise speed, the primary factor influencing inaccuracy and bias is the overlap of point pairs between neighboring road segments. That is, for a vehicle with a discrete sampling interval *sr*, it will often be the case that the first point in the pair will fall on an upstream road segment, or that the second point in the pair will fall on a downstream segment. For internal point pairs, in which both the first and second point fall on the segment of interest, the pairwise speed will likely be a more accurate and unbiased estimate compared to point-wise speeds. For those pairs that overlap a downstream or upstream road segment, the speed associated with the pair will in part reflect the travel speed on the adjacent segment(s).

Despite this, there are three primary reasons why the distance-weighted mean pair-wise speed will often provide more accurate and unbiased results. First, pair-wise speed measures are typically more accurate than GPS point speeds due to the nature of GPS technology. Second, each probe vehicle crossing a road segment can be represented in the segment mean speed estimate, even no samples are obtained on the segment of interest. That is, if a vehicle is sampled on both an upstream and downstream road segment, the speed across the segment of interest can be assigned to the pairwise mean speed between the two segments. Finally, by weighting the speed measures according to the distance traveled, the estimated speed better reflects the contribution of each vehicle to the segment mean speed. In this way, the distance weighting approach can be considered a combination of the favorable elements of vehicle-wise and point-wise mean speed

calculation, in that the relative contribution of each vehicle is considered without over-weighting slower moving vehicles. Estimating completeness is trivial for this approach, because any vehicle which appears on a road segment during an observation interval will be observed. Thus, the count distribution $q_{N_i}(n)$ can fully describe this quantity. An analytical form of the relationship between sampling and traffic parameters and the observed speed is left for future work.

4.2 Special Case: Poisson Arrivals

Following a general introduction to the proposed modeling framework, here the framework is described with respect to specific distributional assumptions that follow from established traffic literature. It is worth noting that the formulation presented here ignores many of the complexities of traffic flow including phase transitions (Kerner and Rehborn 1997), traffic bunching (Nagatani 1995), and inflow/outflow of vehicles along a road segment. However, it provides an interpretable and computationally tractable solution for a variety of traffic conditions that should provide adequate fidelity for planning future experiments and evaluating bias and accuracy in terms of aggregate traffic measures.

In this section, the number of observable arriving vehicles within a time period is assumed to follow a Poisson distribution. The distribution of the number of vehicles that are actually observed, and corresponding number of vehicle updates received, are based on this assumption and the probabilistic relationships described in the previous section. The mean vehicle speeds are assumed to follow a lognormal distribution, though this choice offers no particular advantage in terms of computational complexity. Other distributions of interest are derived with respect to these assumptions, and the implications are discussed.

4.2.1 *Poisson Model for Vehicle Presence*

The probability of *n* arrivals during an observation interval with rate λ (in units of vehicles or devices per observation interval) is given by the Poisson probability mass function (PMF) as shown in Equation (4.21. Assuming a M/G/infinity queuing system, infinite service channels will insure that all incoming users are served as they arrive. This is reasonable for all but the most severely congested conditions on a controlled access facility, because vehicles enter a road segment as they arrive rather than waiting in queue (though congestion does increase the service time and reduce arrival volume). Thus, from Little's Law, the time average number of users in the system (η) is expressed as shown in Equation 4.22, where *E(TT)* is the mean or expected service time, and μ is the departing rate or inverse of *E(TT)* (Adan and Resing 2001; J. D. C. Little 1961).

$$P(n) = \lambda^n/(n!)\exp(-\lambda) \quad (4.21)$$

$$\eta = \lambda/\mu = \lambda E(TT) \quad (4.22)$$

From the PASTA (Poisson Arrivals See Time Averages) property of Poisson distributed arrivals, it can be said that the probability of finding the system in a given state is equal to the fraction of time spent in that state. Without going into the full derivation (see Adan & Resing, 2001), this allows us to represent the probability of finding the system in state n at any given time as a Poisson probability as shown in Equation 4.23, where $\eta = \lambda/\mu$ as before:

$$p_n = \eta^n/(n!)\exp(-\eta) \quad (4.23)$$

This gives an expression for the Poisson PMF that describes the number of vehicles in the system at any given time, but the objective is to estimate the total number in the system over an observation interval. To do this, we can express the probability of a vehicle being present within an observation interval as the sum over two time periods; first, the number remaining in the system at the start of the observation interval (Equation 4.23), and second the number arriving over the observation

interval (Equation (4.21). *TI+E(TT)* is simply the time length of the observation interval plus the expected travel time. Thus, the expected number of users is simply the number of arrivals over the time (expected travel time + observation interval). The distribution for the total number of vehicles appearing in the system over a single observation interval is as shown in Equation 4.24, and the expected number of arrivals a shown in Equation 4.25.

$$P(n|\lambda, E(TT)) = (\lambda(TI + E(TT)))^n/(n!)\, exp\left[-\lambda\left(TI + E(TT)\right)\right] \quad (4.24)$$

$$E(n|\lambda, E(TT)) = \lambda(TI + E(TT)). \quad (4.25)$$

These expressions can describe the number of vehicles appearing for a single subpopulation by replacing the arriving volume and expected travel time with the values for that subpopulation, as shown in Equation 4.26. This gives a form for the vehicle count distribution PMF $q_{N_i}(n)$ described in the Section 4.1, which becomes a Poisson PMF with parameter $(TI + E(TT_i))$.

$$q_{N_i}(n|\lambda_i, E(TT_i)) = (\lambda_i(TI + E(TT_i)))^n/(n!)\, exp[-\lambda_i(TI + E(TT_i))] \quad (4.26)$$

4.2.2 *Sample Count Estimation*

With the PMF for the number of vehicles appearing in the system over an observation interval, it remains to consider sampling frequency and the corresponding sample count distribution. Assuming uniform arrivals over the time period $TI + E(TT)$ (which is true in general for a Poisson process), the integral shown in Equation 5.2 makes it possible to provide a specific form for Equations 4.4 and 4.5, as well as other related expressions described in the previous section.

Using the function $t(x, TT)$ shown as Equation 6.1 and illustrated in Figure 4-1, it is possible to describe the number of samples obtained for a single vehicle as a categorical distribution conditioned on travel time and sampling interval.

To solve the integral of this distribution over arrival time *x*, note that the distribution of arrival times is uniform and therefor a constant equal to $\frac{1}{TI+TT}$ (this follows from the Poisson distribution). Combining this knowledge with area formulas for the geometry of an Isosceles trapezoid, the distribution of the number of samples obtained can be described by the categorical distribution shown in Equation 4.27. In these formulas, $a\%b$ indicates "$a\ modulo\ b$" or the remainder of the integer division a/b. $floor(y)$ indicates the floor function or the largest integer less than or equal to *y*, and $ceil(y)$ indicates the ceiling function or the smallest integer greater than or equal to *y*.

$$P(nspv = j|TT, sr) = \begin{cases} 1 - S1 & if \quad j = 0 \\ \dfrac{2 \times sr}{TI + TT} & if \quad 0 < j < floor\left(\dfrac{\min(TT, TI)}{sr}\right) \\ \dfrac{S2 - S3}{sr \times (TI + TT)} & if \quad j = floor\left(\dfrac{\min(TT, TI)}{sr}\right) \\ \dfrac{S3}{sr \times (TI + TT)} & if \quad \dfrac{\min(TT, TI)}{sr} < j < ceil\left(\dfrac{\min(TT, TI)}{sr}\right) \\ 0 & else \end{cases} \quad (4.27)$$

Where *S1*, *S2*, and *S3* are defined as shown in Equations x, x, and x.

$$S1 = \frac{(TI + TT)\min(sr, TI, TT) - \min(sr, TI, TT)^2}{(TI + TT)sr} \quad (4.28)$$

$$S2 = sr \times \left(TI + TT - 2 \times floor\left(\frac{\min(TT, TI)}{sr}\right) \times sr + sr\right) \quad (4.29)$$

$$S3 = (\min(TT, TI)\ \%sr) \times \left(TI + TT - 2 \times \min(TT, TI) + (\min(TT, TI)\ \%sr)\right) \quad (4.30)$$

This expression can be combined with Equation 4.4 to compute the overall probability of a given number of samples for a single subpopulation *i* given only sr_i. Substituting Equation 4.27 as the expression for $P(nspv_i = j|TT(v), sr_i)$ in Equation 4.4 gives a formula for the distribution of the number of samples obtained per vehicle conditioned only on the sampling interval.

Some simplifications can be made to Equation 4.27 to derive expressions for the expected number of samples and the probability of observing a vehicle given it has arrived. First, as noted previously, the expected number of samples for a given *x* and *TT* is simply the value of the function $t(x, TT)$ divided by *sr*. From this, a trapezoidal area formula can be used to compute the expected number of samples per vehicle (*nspv*) given *TT* and *sr* as shown in Equation 4.31. This formula can be combined with Equation 4.5 to compute the expected number of samples per vehicle for each subpopulation given only sr_i and the speed distribution $f_{V_i}(v)$, resulting in Equation 4.32.

$$E(nspv|TT, sr) = \frac{(TI + TT)\min(TI, TT) - \min(TI, TT)^2}{(TI + TT)sr} \quad (4.31)$$

$$E(nspv_i|sr_i) = \int_v \frac{\big(TI + TT(v)\big)\min\big(TI, TT(v)\big) - \min\big(TI, TT(v)\big)^2}{\big(TI + TT(v)\big)sr_i} f_{V_i}(v)dv \quad (4.32)$$

Finally, to get the distribution of the total number of samples over all possible vehicles, Equations 4.4, 4.6, and 4.27, are combined to give the expression shown as Equation 4.33.

$$P(ns_i|sr_i, \lambda_i, E(TT_i)) = \sum_n \left[q_{N_i}(n|\lambda_i, E(TT_i)) \sum_{m \in C_{ns_i,n}} \left(\prod_{j \in m} P(nspv_i = j|sr_i) \right) \right] \quad (4.33)$$

The corresponding expected value expression is shown as Equation 4.34, which combines the Poisson expected number of vehicles with the expected number of samples per vehicle from Equation 4.32.

$$E(ns_i|sr_i, \lambda_i, E(TT_i)) = \lambda_i(TI + E(TT_i) \times E(nspv_i|sr_i) \quad (4.34)$$

From these expressions, it is possible to compute the expected number of samples over all subpopulations, the missing data probability, and the expected observed speed by substituting these expressions into the formulas described in the previous section.

4.2.3 *Considering Heterogeneous Vehicle Populations*

In many cases, significant variation will be present in the sampling parameters and speed distributions over different vehicle subpopulations. As noted previously, the total observable vehicle volume can be interpreted as arising from a set of *k* distinct subpopulations, each with a fixed value of sr and speed distribution. Under this interpretation, observable vehicles are drawn from a categorical distribution, where each category *i* is associated with a fixed value of sr_i and speed distribution $f_{V_i}(v)$.

$$sr_i, f_{V_i}(v) \sim Cat(k, \pi) \quad (4.35)$$

Where $\pi_i, i \in \{1,2, \dots, k\}$ is the parameter vector for the categorical distribution, each *i* associated with a distinct subpopulation. In words, π_i is the probability of a single randomly selected vehicle belonging to subpopulation *i*, given that it is a member of an observable subpopulation. Thus, for each subpopulation i, the arriving vehicle volume can be expressed as the product of the overall observable vehicle volume and the associated distribution parameter π_i.

$$\lambda_i = Q \times frac \times \pi_i \quad (4.36)$$

In this expression, $frac$ indicates the total penetration rate, or the overall penetration rate of probe vehicles considering all subpopulations. Q indicates the total arriving vehicle volume, in units of vehicles per observation interval. As before, the arriving vehicle volume λ_i can be used to calculate the Poisson parameter describing the expected value and variance for the observable vehicle count for subpopulation *i* appearing in the system during an observation interval.

$$E(n_i) = \lambda_i\big(TI + E(TT_i)\big) \quad (4.37)$$

The total expected number of observable vehicles appearing in the system in an observation interval can then be expressed as shown in Equation 4.38.

$$E(n) = \sum_{i=1}^{k} \lambda_i\big(TI + E(TT_i)\big) = \sum_{i=1}^{k} E(n_i) \quad (4.38)$$

There are no specific constraints on the subpopulation speed distributions, except that they faithfully represent the true (not measured) distribution of vehicle mean speeds across the road section of interest. However, in most cases the speed distributions for all subpopulations will be in the same family, and so will be defined by the distribution parameters. For example, if the speed distributions are all assumed to be lognormal, the speed distribution $f_{V_i}(v)$ will be defined as shown in Equation 4.39. In this expression, μ_{vm_i} and σ_{vm_i} are the lognormal distribution parameters, or the location and scale parameters for the normally distributed natural log of mean vehicle speed.

$$f_{V_i}(v) = Lognormal\left(\mu_{vm_i}, \sigma^2_{vm_i}\right) \quad (4.39)$$

Note that $f_{V_i}(v)$ describes the distribution of mean vehicle speeds for subpopulation i, so $\sigma^2_{vm_i}$ is distinct from $\sigma^2_{v_i}$, which describes the variance of a single vehicle's speed along a road segment given the mean. Some estimate of $\sigma^2_{v_i}$ is required to estimate the variance in the observed speed measurement. In simulation, this quantity can be estimated by first calculating the variance of each vehicle's speed across a road segment within a single observation interval, and then taking the subpopulation-wise mean of the variance (this is how it is done in Section Chapter 5:). In practical applications, there are a variety of ways this quantity could be estimated. For example, if speed sensing hardware is placed along a road segment, it may be simpler to estimate $\sigma^2_{v_i}$ from a synthetic speed profile based on measurements from this hardware.

4.3 Methodology

The formulas presented in Sections 4.1 and 4.2 make some assumptions about what is known, and what *can* be known, about the true traffic state and sampling parameters. That is, estimating the distribution of the observed speed and sample sizes requires that the sampling interval, speed

distribution, and contributing fraction of all subpopulations are known or can be estimated for the location(s) and time period(s) of interest. In this form, the proposed framework would be useful in planning new data collection efforts to minimize the bias and/or variance in the observed speed and achieve some minimum sample size, or in evaluating the quality of existing probe vehicle data. For example, one could estimate the true speed distribution for a range of traffic conditions using mechanical sensor data and apply the methodology to prescribe a minimum penetration rate or sampling frequency distribution to meet data quality requirements.

Another likely objective would be to estimate the true speed and corresponding sampling bias given the observed speed and sampling parameters. This is more challenging. On one hand, computing the overall sampling bias from these quantities would be feasible. However, at present this would not be very useful for correcting sampling bias in probe vehicle data. Consider that, especially at low penetration rates, the error or difference between the true traffic speed and the observed speed is comprised of two components: 1) random and 2) systematic. Random error is driven by the fact that the measured data is derived from a small number of vehicles relative to the total population, and is based on relatively sparse point measurements. Systematic error is driven by the extent to which the sampling parameters systematically give higher weight to certain vehicle subpopulations and travel speeds. As the overall penetration rate increases, the random component decreases relative to the systematic component as else being equal. Thus, at low penetration rates, the bias is small relative to the random error and adjusting for it using some form of correction factor could have deleterious impacts on the accuracy of the data. As the penetration rate increases and systematic error increasingly dominates, a correction factor becomes a more feasible solution.

In this work, two possible implementation scenarios are considered. First, the case where the true speed distribution(s) and sampling parameters are known (or can be estimated) for all

vehicle subpopulations, and the objective is to estimate the completeness, observed speed or sampling bias, and sample sizes. This would be the most likely case for planning a future data collection project. In the second case, only the observed speeds and sampling parameters are known, and the objective is to correct for sampling bias. As noted previously, correcting for bias is difficult when the number of contributing vehicles is small and random error is high relative to systematic error. For this reason, implementing such a methodology is discussed in detail in a separate part of this dissertation (Chapter 6.3).

4.3.1 *Parameter Definitions*

This subsection explains how to apply the proposed methodology to estimate sample sizes, missing data rates, and observed speed, focusing on the Poisson arrival case described in Section 4.2. It is assumed here that the probe vehicle population consists of *k* non-overlapping subpopulations, each with a fixed penetration rate and distinct combination of sampling interval and speed distribution. Listed below are the parameters needed to estimate statistical measures describing sample size, missing data rate, and observed speed. These quantities define an analysis scenario.

Probe Vehicle Population Parameters:

- *Overall Penetration Rate* ($frac$): this is the fraction of the total vehicle population that is contributing to the probe vehicle dataset. For multiple non-overlapping subpopulations, it is the sum penetration rate over all subpopulations. If the intent is to investigate the impact of different penetration rates on sample sizes and/or bias, this value can be adjusted over a grid of values.
- *Subpopulation Fractions* ($\pi_i \ \{i \in 1,2,\dots,k\}$): This is the fraction of the overall probe vehicle population the comprises each vehicle subpopulation. In other words, the expected observable traffic volume for each subpopulation *i* is the product of the total vehicle

volume (Q), the overall penetration rate ($frac$), and the fraction associated with the subpopulation of interest (π_i).

Traffic Parameters:

- *True Mean Speed Distribution* ($f_{V_i}(v) \quad \{i \in 1,2,\dots,k\}$): This is the distribution of the true per-vehicle mean speed for each subpopulation. For a parametric distribution family, it is defined by a set of distribution parameters for each subpopulation. For example, a plausible assumption would be that the mean speed is normally distributed. In a microscopic traffic simulation scenario, the distribution parameters (mean and variance) can be computed by first computing the per-vehicle mean speed for all vehicles, and then computing the mean and variance of this speed over all vehicles. For a hypothetical real-world scenario, this quantity can be estimated from existing mechanical sensor data or based on prior research.
- *True Expected Travel Time* ($E(TT_i) \quad \{i \in 1,2,\dots,k\}$): This is the expected travel time for each vehicle subpopulation. It can be computed from the mean speed distribution $f_{V_i}(v)$ and the road segment length L.
- *Traffic Volume* (Q): This is the traffic volume for the scenario of interest, including all vehicles (not just probe vehicles).
- *Speed Variance Along the Road Segment* ($\sigma_{v_i}^2 \quad \{i \in 1,2,\dots,k\}$): This is an estimate of the variance of a single vehicle's speed along the road segment of interest. It does not include the between-vehicle variance of the mean, only the variance along the route for a single vehicle. In this work, this quantity is estimated by first computing the per-vehicle variance for all vehicles within each observation interval, and then taking the mean of per-vehicle variance over each subpopulation. This would underestimate the effective variance due to

autocorrelation, so the observed data is down-sampled to match the according to the sampling frequency.

Probe Vehicle Sampling Parameters:

- *Sampling Interval* ($sr_i \ \{i \in 1,2,\dots,k\}$)*:* The sampling interval is the length of time between subsequent vehicle state updates for each subpopulation. If the intent is to investigate the impact of different sampling interval distributions, one could consider multiple sampling interval distributions (defined by a combination of subpopulation fractions and associated sampling intervals) to investigate the impact on sample sizes, missing data rates, and sampling bias.
- *Observation Interval* (*TI*): This is the length of time over which individual vehicle state updates are aggregated to discrete time traffic observations. Multiple observation interval lengths may be tested to investigate the impact on sample sizes, missing data rates, and sampling bias.

Site Parameters:

- *Road Segment Length* (*L*): This is simply the length of the road segment. It is needed to make the conversion between vehicle speeds and travel times.

4.3.2 *Analytical Approach*

This first approach applies the formulas described in previous sections to analytically compute the observed speed, sample size, and missing data rate. This approach will more straight forward compared to the sampling method described in the following subsection, but it is less flexible and in general cannot be used to estimate the full distribution of the observed speed. To explain, the distribution of the observed speed will depend on the method used to combine point vehicle updates into aggregate measures for each road segment and time period. If pointwise updates are

first aggregated to a single speed value for each vehicle, and these vehicle speeds are averaged for each road segment and time period, then an analytical solution for the expected observed speed is possible. If the simple average of pointwise updates for each time period and road segment are used, then sampling methods will be required to estimate the distribution of the observed speeds. Furthermore, as noted previously, sampling methods will be the most efficient way to estimate the variance of the observed speed when multiple probe vehicle subpopulations are present.

The following steps describe the process for estimating sample size distributions, missing data rates, and expected observed speed. Expected observed speed estimation is described for the case when pointwise updates are first aggregated to a single speed value for each vehicle, and these vehicle speeds are averaged for each road segment and time period (sample sizes and missing data rates do not depend on the speed aggregation method).

1) Estimate the arriving volume for each subpopulation using Equation 4.36:

$for\ i \in \{1,2,\dots,k\}$:

$$\lambda_i = Q \times frac \times \pi_i$$

2) Compute the expected travel time for each subpopulation using the road segment length and mean vehicle speed distribution. In most cases, numerical methods will be required to integrate over the full support for $f_{v_i}(v)$. For the lognormal case described in Section 4.2.3, this can be represented as:

$for\ i \in \{1,2,\dots,k\}$:

$$E(TT_i) = \int_v \frac{L}{v} \times f_{v_i}(v)dv = \int_v \frac{L}{v}\left[\frac{1}{v \times \sigma_{mv,i}\sqrt{2\pi}} e^{\frac{(\ln(v)-\mu_{mv,i})^2}{2\sigma^2_{mv,i}}}\right]dv$$

3) The full distribution of the number of samples per vehicle for each subpopulation can be estimated using Equations 4.4 and 4.27. In most cases, numerical methods will be required to integrate over the full support for $f_{v_i}(v)$. For the Lognormal case, this can be represented as:

$for\ i \in \{1,2,\dots,k\}$:

$$f_{v_i}(v) = \frac{1}{v \times \sigma_{mv,i}\sqrt{2\pi}} e^{\frac{(\ln(v)-\mu_{mv,i})^2}{2\sigma_{mv,i}^2}}$$

$for\ j \in \{0,1,\dots,\infty\}$:

$$P(nspv_i = j|sr_i) = \int_v P(nspv_i = j|TT(v), sr_i) f_{v_i}(v) dv$$

4) The expected sample size (ns_i) for each subpopulation can be estimated using previously described results along with Equations 4.32 and 4.34. In most cases, numerical methods will be required to integrate over the full support for $f_{v_i}(v)$. For the Lognormal case, this can be represented as:

$for\ i \in \{1,2,\dots,k\}$:

$$f_{v_i}(v) = \frac{1}{v \times \sigma_{mv,i}\sqrt{2\pi}} e^{\frac{(\ln(v)-\mu_{mv,i})^2}{2\sigma_{mv,i}^2}}$$

$$E(nspv_i|sr_i) = \int_v \frac{\big(TI + TT(v)\big)\min\big(TI, TT(v)\big) - \min\big(TI, TT(v)\big)^2}{\big(TI + TT(v)\big)sr_i} f_{v_i}(v) dv$$

$$E(ns_i|sr_i, \lambda_i, E(TT_i)) = \lambda_i(TI + E(TT_i) \times E(nspv_i|sr_i)$$

5) The full sample size distribution for each subpopulation can be estimated using previously described results along with Equations 4.4, 4.6, 4.26, and 4.27. For the Poisson case, this can be described as shown here, where $C_{l,n}$ indicates the weak composition of l or the set of all possible integer tuples of length n which sum to l.

$for\ i \in \{1,2,\dots,k\}$:

$$q_{N_i}(n) = \frac{\left[\lambda_i\left(TI + E(TT_i)\right)\right]^n e^{\lambda_i\left(TI+E(TT_i)\right)}}{n!}$$

$for\ l \in \{0,1,\dots,\infty\}$:

$$P(ns_i = l|c, sr_i) = \begin{cases} \sum_{m \in C_{l,c}} \left(\prod_{j \in m} P(nspv_i = j|sr_i) \right) & if\ n > 0 \\ 1 & if\ l = n = 0 \\ 0 & otherwise \end{cases}$$

$$P(ns_i = l|sr_i, \lambda_i, E(TT_i)) = \sum_{c=0}^{\infty} \left[q_{N_i}(c) P(ns_i = l|c, sr_i) \right]$$

6) The full sample size distribution over all subpopulations can be estimated using previously described results along with Equation 4.8 as shown here:

$$P(ns_{all} = s|sr) = \sum_{m \in C_{s,k}} \left[\prod_{i \in \{1,2,\dots,k\}} P(ns_i = m_i|sr_i, \lambda_i, E(TT_i)) \right]$$

7) The expected observed speed for each subpopulation can be estimated using Equations 4.39, 4.27, and 4.12. For the Lognormal/Poisson case, this can be shown as follows, where again the integral over v is likely to require numerical integration:

$for\ i \in \{1,2,\dots,k\}$:

$$f_{v_i}(v) = \frac{1}{v \times \sigma_{mv,i}\sqrt{2\pi}} e^{\frac{(\ln(v) - \mu_{mv,i})^2}{2\sigma_{mv,i}^2}}$$

$$P(nspv_i > 0|TT(v), sr_i) = \frac{(TI + TT(v)) \min(sr_i, TI, TT(v)) - \min(sr_i, TI, TT(v))^2}{(TI + TT(v)) sr_i}$$

$$E\left(v_{obs,i}|sr_i\right) = \frac{\int_v v \times f_{v_i}(v) \times P(nspv_i > 0|TT(v), sr_i) dv}{\int_v f_{v_i}(v) \times P(nspv_i > 0|TT(v), sr_i) dv}$$

8) The expected observed mean speed over all subpopulations can be calculated using previously described results along with Equations 4.13 and 4.17 as shown here:

$for\ i \in \{1,2,\dots,k\}$:

$$M_i = \lambda_i\big(TI + E(TT_i)\big) \times (1 - P(nspv_i = 0|sr_i))$$

$$E(\bar{v}_{obs}|sr) = \frac{\sum_{i\in\{1,2,\dots,k\}} M_i \times E\big(v_{obs,i}\big|sr_i\big)}{\sum_{i\in\{1,2,\dots,k\}} M_i}$$

In this work, this process is implemented using the Python programing language. Using widely available open source statistical and numerical integration libraries, computation time for a single scenario is on the order of a few seconds.

4.3.3 *Sampling Approach*

To estimate the distribution of the observed speed when the average of all point speed observations is taken as the observed speed, it is necessary to use sampling methods. Actually, sampling is a much more general approach that can be used to estimate the full distribution of the observed speed under a variety of aggregation methods. The problem can be represented graphically as shown in Figure 4-3, and allows forward sampling to generate samples from the posterior distributions of interest. The shaded circles represent observed quantities, the non-shaded circles represent unobserved variables, and the wide bordered squares indicate repeated variables or groups of variables. For example, there are *k* values for n_i and $\sum_{i\in\{1,2,\dots,k\}} n_i$ values for $v_{i,c}$ and $nspv_{i,c}$. The representation in Figure 4-3 is somewhat of an abuse of graphical notation because the number of vehicles and the number of samples per vehicle are themselves random variables (rather than fixed values, which is customary for the use of plate notation in graphical model representation). Some methods to depict directed graphical models with variable numbers of nodes have been introduced

such as (Mjolsness 2004). However, there is no widely accepted notation for this, and so the standard plate notation was adapted to the task.

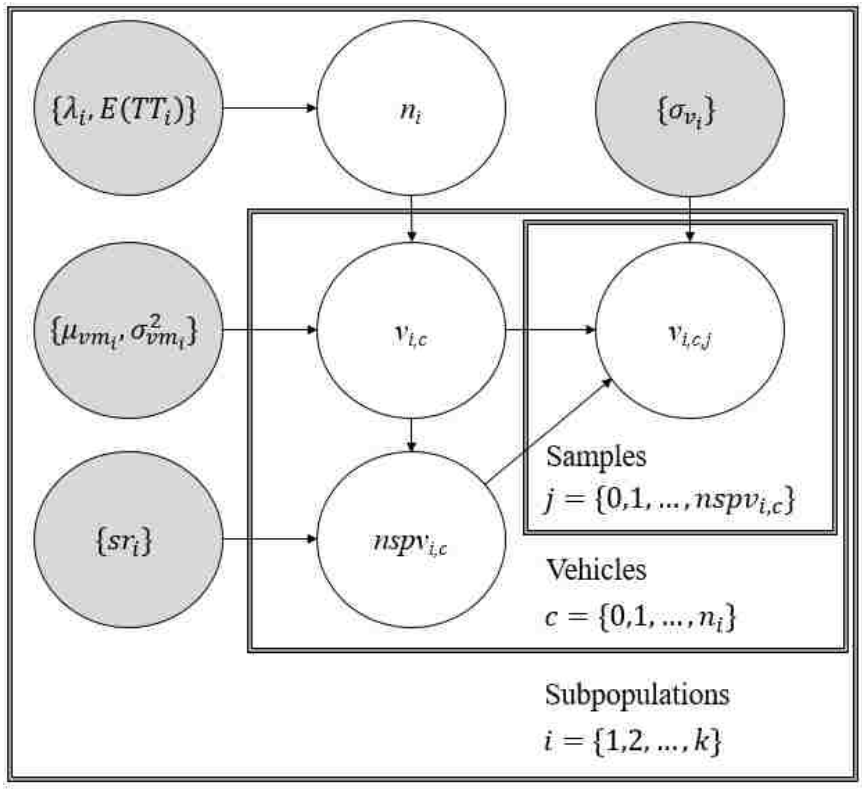

Figure 4-3: Directed Graphical Model Representation

Forward sampling can be used to derive empirical estimates of the expected value, variance, and other measures. To obtain a single sample representing an aggregate speed measurement for a single time period, the following steps are completed.

9) Draw vehicle counts for each subpopulation $i \in (0,1, \dots, k\}$ from the corresponding PDF $q_{N_i}(n)$. For the Poisson case described in Section 4.2.1, this can be represented as:

$$for\ i\ \in \{1,2, \dots, k\}:$$

$$n_i \sim Pois\left(\lambda_i\left(TI + E(TT_i)\right)\right)$$

10) For each vehicle *c* in each subpopulation *i*, draw a mean speed $v_{i,c}$ from the distribution $f_{V_i}(v)$. For the lognormal case described in Section 4.2.3, this can be represented as:

$for\ i\ \in \{1,2,\dots,k\}$:

$for\ c \in \{1,2,\dots n_i\}$:

$$v_{i,c} \sim Lognormal\left(\mu_{vm_i}, \sigma^2_{vm_i}\right)$$

11) For each vehicle *c* in each subpopulation *i*,, draw a sample count $nspv_{i,c}$ from the categorical distribution defined by Equation 4.27.

$for\ i\ \in \{1,2,\dots,k\}$:

$for\ c \in \{1,2,\dots n_i\}$:

$$nspv_{i,c} \sim P\left(nspv_i \middle| TT(v_{i,c}), sr_i\right)$$

12) For each sample *m* for vehicle *c* in each subpopulation *i*, draw an observed speed using the previously sampled mean $v_{i,c}$ and variance σ_{v_i}. Any continuous distribution that can be parameterized by the mean and variance can be used, here it is shown as a Normal distribution. Note that measurement errors can be incorporated into this step by applying a second sampling step, which considers the variance in the measured speed given the true speed ($v_{i,c,j}$).

$for\ i\ \in \{1,2,\dots,k\}$:

$for\ c \in \{1,2,\dots n_i\}$:

$for\ j\ \in \{1,2,\dots,nspv_{i,c}\}$:

$$v_{i,c,j} \sim Normal\left(v_{i,c}, \sigma^2_{v_i}\right)$$

13) Aggregate the speed of all vehicles according to the desired aggregation method, resulting in a single observation for travel speed or travel time across the road segment. If the simple mean of all point vehicle state updates is taken as the observed value, the formula shown as Equation 4.40 can be used. Alternatively, if the per-vehicle mean is taken first, and the mean of per-vehicle speed is taken as the observed value, Equation 4.41 can be used.

$$\bar{v}_{obs} = \frac{1}{\sum_{i=1}^{k}\left[\sum_{c=1}^{n_i} nspv_{i,c}\right]} \sum_{i=1}^{k}\left[\sum_{c=1}^{n_i}\left(\sum_{j=1}^{nspv_{i,c}} v_{i,c,j}\right)\right] \quad (4.40)$$

$$\bar{v}_{obs} = \frac{1}{\sum_{i=1}^{k} n_i} \sum_{i=1}^{k}\left[\sum_{c=1}^{n_i}\left(\frac{1}{nspv_{i,c}} \sum_{j=1}^{nspv_{i,c}} v_{i,c,j}\right)\right] \quad (4.41)$$

This sampling is completed many times, and the results combined to provide estimates of the desired quantities including expected value, variance, empirical quantiles, etc. In addition to speed, other quantities of interest such as the vehicle and sample counts, missing data rate, and associated statistical measures can be found by aggregating the assignment of the variable of interest over all samples generated in this process. For example, indexing the samples by $t \in \{1,2, \ldots T\}$ with a total sample count of T, the expected missing data rate can be computed as shown in Equation 4.42.

$$P(R = 0) = \frac{\sum_{t=1}^{T} I_{R_t=0}}{T} \quad (4.42)$$

Where $I_{R_t=0}$ is an indicator variable taking the value 1 if $\left[\sum_{i=1}^{k}\left(\sum_{c=1}^{n_i} nspv_{i,c}\right)\right]_t = 0$ and 0 otherwise. Similarly, the expected number of observed vehicles $E(n_{obs})$ and Expected number of samples $E(ns)$ per observation interval can be computer as shown in Equations 4.43 and 4.44, respectively, where $I_{nspv_{i,c}>0}$ is an indicator variable taking the value 1 if $nspv_{i,c} > 0$ and 0 otherwise.

$$E(n_{obs}) = \frac{1}{T}\sum_{t=1}^{T}\left[\sum_{i=1}^{k}\left(\sum_{c=1}^{n_i} I_{nspv_{i,c}>0}\right)\right]_t \quad (4.43)$$

$$E(ns) = \frac{1}{T}\sum_{t=1}^{T}\left[\sum_{i=1}^{k}\left(\sum_{c=1}^{n_i} nspv_{i,c}\right)\right]_t \quad (4.44)$$

Chapter 5: Validation

5.1 Experimental Set up

The validation work described here supposes that, if the model accurately represents the underlying sampling process, then the predicted speed, completeness, and sample size should agree with the results obtained using microscopic traffic simulation. The proposed model is designed to represent the sampling process based on aggregate quantities (mean and variance of speed, mean arriving volume, mean penetration rate, etc.), and describe the outcome in terms of link-level averages over a reasonably large set of observation intervals. Thus, the validation should be based on mean speed and between – observation interval speed variance, mean completeness, and mean vehicle count and sample size. To do this, a series of simulations were completed using PTV's VISSIM microscopic traffic simulation software. A location on Interstate 5 in the north Seattle, WA area was selected for the simulation. A mode was constructed with 6 adjoining road links, all with a fixed speed limit of 60 mph and four traffic lanes in the northbound direction. A single lane slowdown section was added in link 7 to introduce some heterogeneity in traffic conditions. Table 5-1 shows the link lengths and ID numbers.

Figure 5-1 shows the location from which the test scenario was adapted. There are some differences between the Interstate 5 site and the simulation model, this location was selected simply as an example of a basic freeway section. No on or off ramps were used in the test scenario, and all links were designed with 4 mainline lanes (ignoring HOV lanes). By allowing multiple vehicle subpopulations, HOVs are considered implicitly, so including the HOV lanes would not introduce any additional complexity. Allowing a significant number of vehicles to enter and exit

the traffic stream along a link would also be straightforward, but was not considered in this experiment.

Table 5-1: VISSIM Model Road Link Definitions

Link ID	Length (miles)
1	0.325
3	0.097
4	0.226
6	0.155
7	0.526
9	0.870

Figure 5-1: Location of Test Site

In all cases, the simulation model is run for 10080 seconds for each of five random seed values (42, 43, 44, 45, and 46). With the output of each simulation run, vehicles are selected at random according to the selected penetration rate, and assigned to a sampling rate using a draw from the multinomial distribution for *sr*. The measured speed, true speed (over all vehicles), sample count, and missing data rate are then computed for each roadway link. The vehicle selection is completed 50 times for each simulation run, and results combined over all simulation runs and vehicle assignments. The first 5 minutes of each simulation run is discarded, in order to insure that all road links are populated before data collected begins.

The simulated scenario was roughly adapted from the probe data described in (Patire et al. 2015) which describes the sampling rate distribution and penetration rate for two real-world probe vehicle datasets. The first provider, here referred to as Provider A, has a lower mean desired speed and a higher average sample rate. The second provider, Provider B, has a faster mean desired speed and a lower average sample rate. The overall vehicle population is strictly comprised of vehicles with the driving characteristics of these two subpopulations, but only a fraction of all vehicles are probes. In all validation scenarios, $1/7^{th}$ of all vehicles have the desired speed distribution of provider A, and $6/7^{th}$ of all vehicles follow the desired speed distribution of provider B. The within population penetration rates is the same for both subpopulations in all cases, though the overall penetration rate (as a fraction of the total vehicle volume) differs. It is clear that this will not be the case in general, and in fact a real world vehicle population may contain vehicle subpopulations with speed distributions that differ significantly from those represented in the probe vehicle population. Thus, the simulated scenario could be naively interpreted as a "best case scenario" in that the probe vehicle population is representative of the true traffic population with respect to desired speed distribution.

To explore the bias that can result when two providers have different speed distributions, synthetic desired speed distributions were devised for both providers. Entering traffic volume is set to $4500\ veh/hour$, which gives arriving volumes for Providers A and B of $\lambda_A = penetration\ rate \times {}^{1}/_{7} \times 4500\ veh/hour$ and $\lambda_B = penetration\ rate \times {}^{6}/_{7} \times 4500\ veh/hour$, respectively. The desired speed distributions for the two simulated data providers are shown in Figure 5-2.

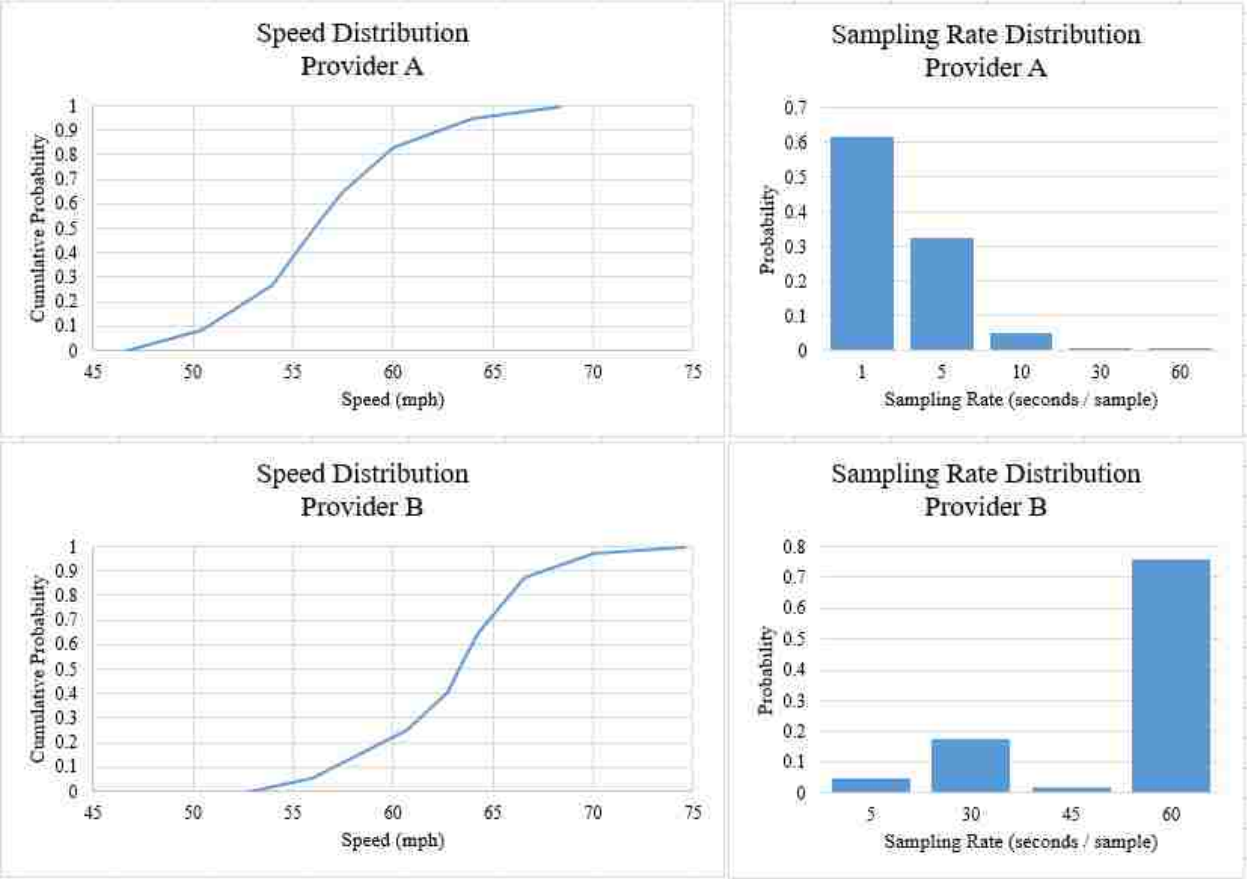

Figure 5-2: Desired speed distributions and sampling rate distributions for the two simulated data providers.

5.2 Sample Size and Completeness

Figure 5-3 shows the completeness rate (or, conversely, missing data rate) estimation results for an overall penetration rate of 0.02 (2%) and observation interval of 2 minutes. Note that

completeness is higher, all being equal, for longer road segments due to higher average travel time. Over 40% of the observation intervals were missing for the shortest segment and nearly 25% of all observation intervals were missing. As this figure shows, the analytical and Monte Carlo (MC) methods both provide very high accuracy.

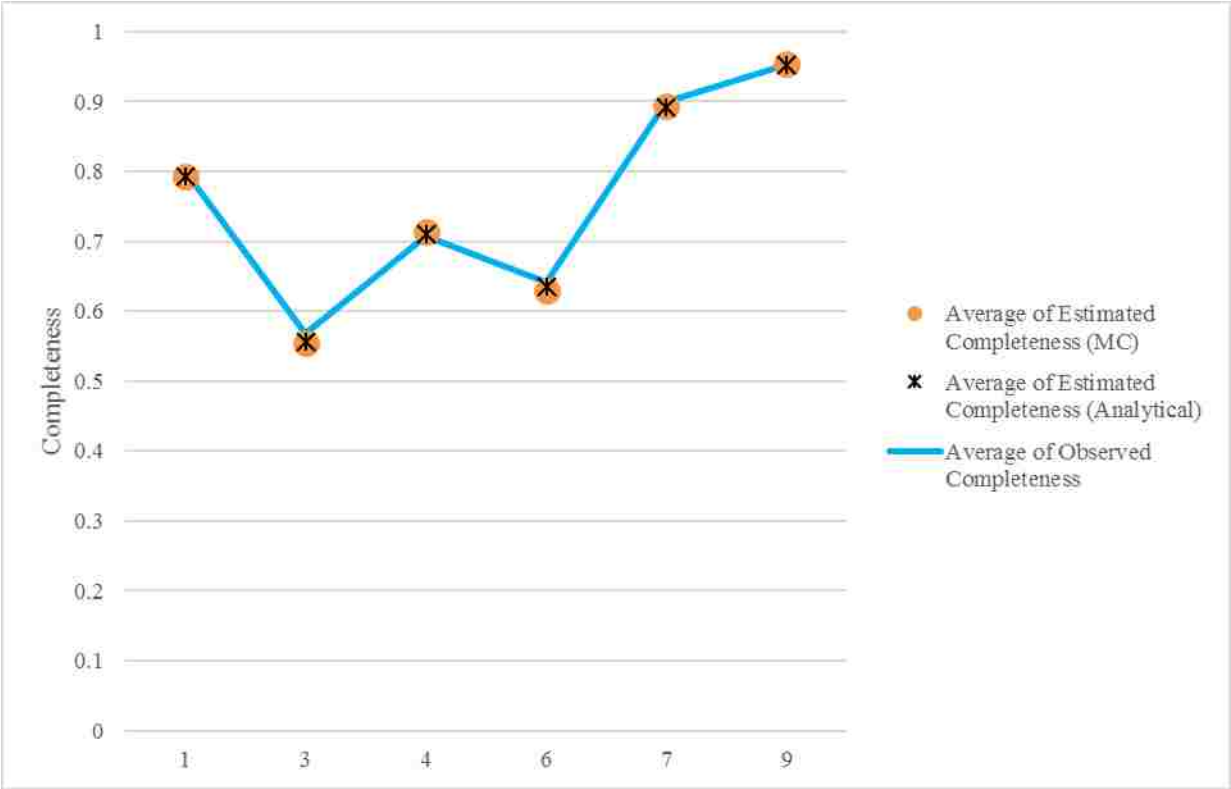

Figure 5-3: Completness Estimation Results for the Base Scenario

Figure shows the estimated and observed completeness for one road segment over a range of penetration rates. Again it is clear that the analytical and Monte Carlo methods give nearly identical results, and both accurately reflect the observed completeness. Figure 5-5 shows the observed and estimated vehicle count (per observation interval) for road segment 3. Because traffic state is approximately identical for all penetration rates, the observed vehicle count is approximately a multiplicative fraction of the probe vehicle volume. Because of this, the relationship between penetration rate and vehicle count is expected to be linear. However, the completeness does not

increase linearly with penetration rate, because of the increasing likelihood of multiple vehicles being present and the number of vehicles present in a single observation interval also increases with penetration rate.

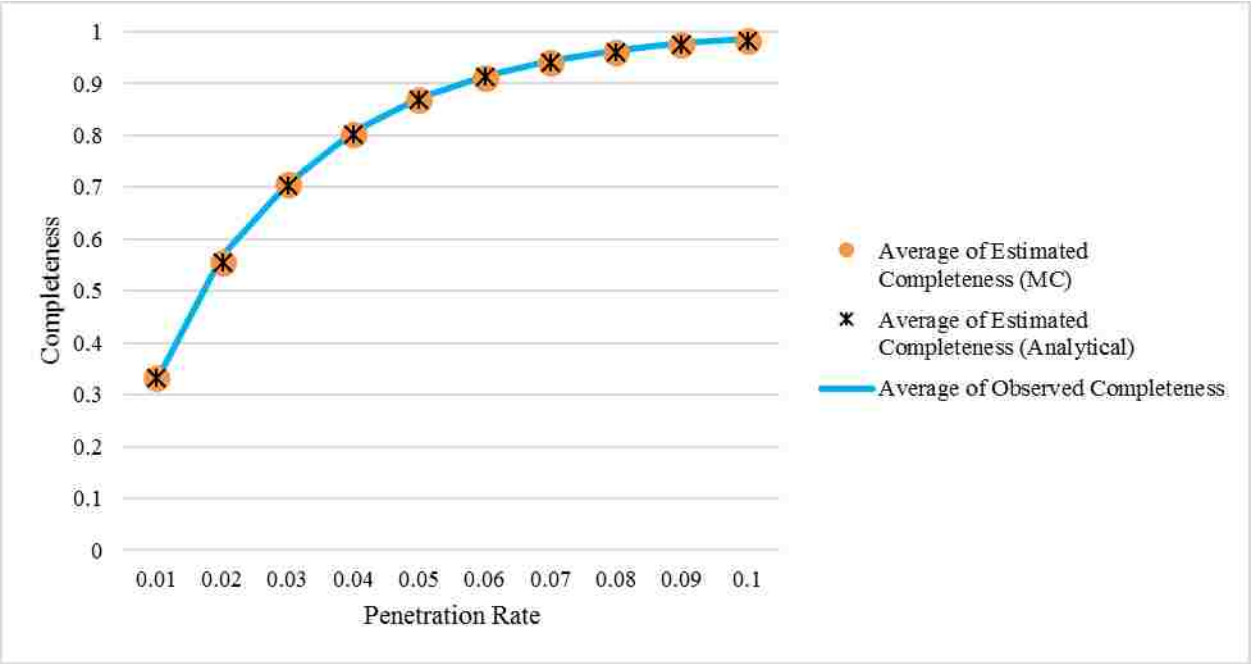

Figure 5-4: Missing Data vs. Penetration Rate, Road Segment 3

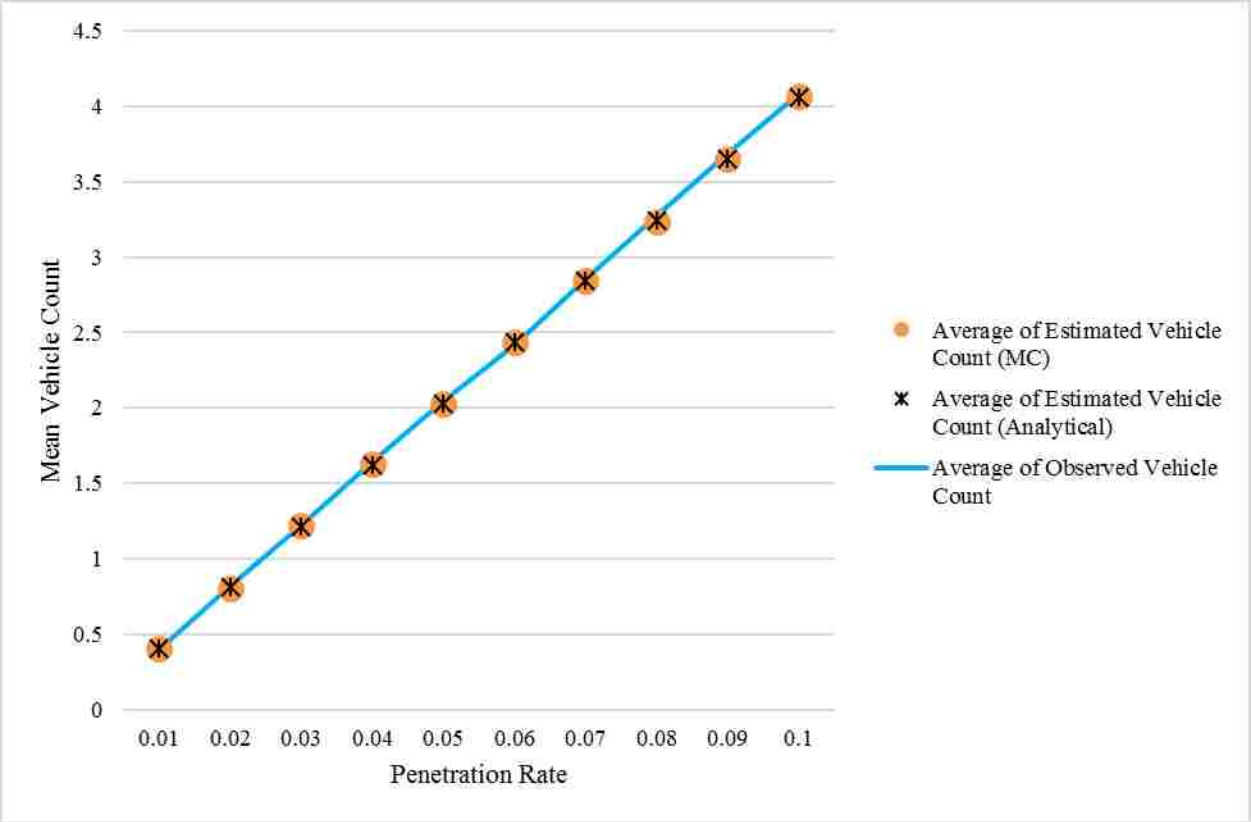

Figure 5-5: Observed Vehicle Count vs. Penetration Rate, Road Segment 3

Figure 5-6 shows the mean sample count (per observation interval) for road segment 3 over a range of penetration rates. Again, the relationship between sample count and probe vehicle population is expected to be linear, which is reflected in these results.

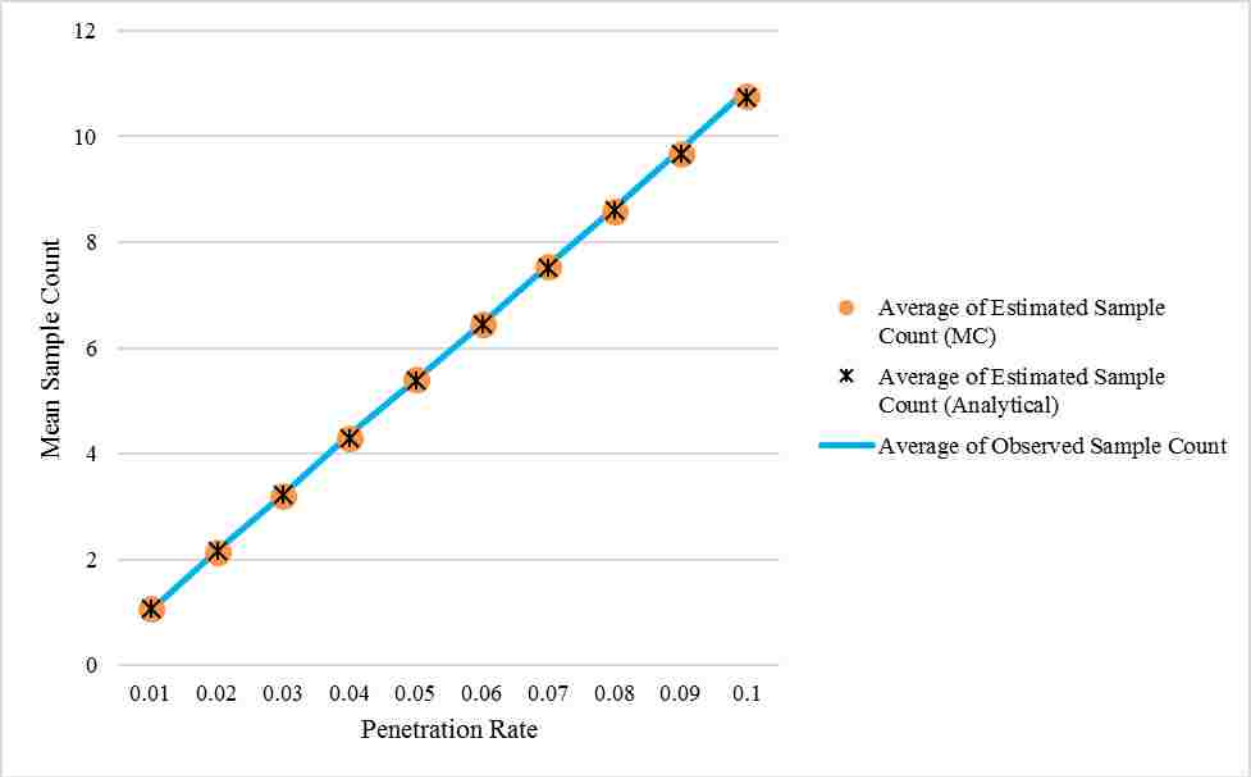

Figure 5-6: Mean Sample Count vs. Penetration Rate, Road Segment 3

5.3 Measurement Bias and Variance

Figure 5-7 shows the speed estimation results for all road sections for all road segments with an overall penetration rate of 0.02 and observation interval of 2 minutes. In this case, speed is first aggregated by vehicle and then by time period, and the estimated measured speed is calculated analytically and using the Monte Carlo (MC) method. The true speed considers all vehicles at all times, and as such represents the actual mean speed over all vehicles and time periods (including both probe and non-probe vehicles). This figure shows a mean bias of just over 1 mph, though there is variation between different road segments. The lowest bias is observed on link 9, the longest link. This largely reflects the fact that link 9 has a mean travel time of nearly

one minute, there is very low occurrence of vehicles with long sampling intervals being missed completely.

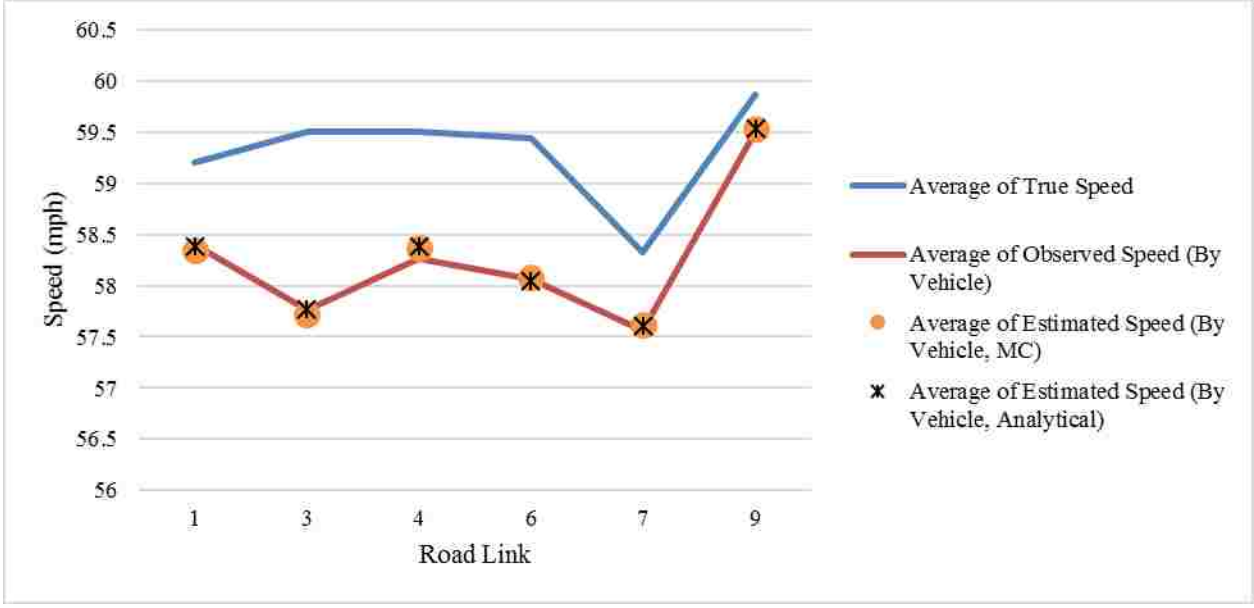

Figure 5-7: Speed Estimation Results for Penetration Rate of 0.02

To explore the relationship between observation interval and sampling bias, a range of observation intervals is used. Consider that, with 1-minute observation intervals, there would not be a great deal of difference between aggregating by vehicle first or taking the overall mean of vehicle speed updates in each time interval for the penetration rates used here. This is because very few vehicles pass through each road segment in 1 minute, and so most observations are simply a single vehicle. By increasing the length of the observation interval, it allows more vehicles to be present in each time step, with the result that the number of samples obtained per vehicle has significantly more influence that for shorter observation intervals. In this section, aggregating first by vehicle and road link and then by time period will be referred to as method 1, and taking the simple mean of updates for each time period and road link will be referred to as method 2. For method 2, the Monte Carlo method is used. For method 1, the analytical approach is used. (both methods are described in detail in Section Chapter 4:).

Figure 5-8 shows that, for method 1, bias is most severe for shorter observation intervals and quickly reaches a relatively static level. On the other hand, the speed bias generally increases with the length of the observation interval for method 2 for observation intervals greater than 60 seconds. As noted previously, using the simple mean of all vehicle updates in a time interval generally leads to more emphasis on slower moving vehicles and those with shorter sampling intervals. Comparing the two, it can be seen that bias decreases somewhat between the 30 second and 120 second observation intervals in both cases. Below 60 seconds, bias is primarily driven by the increased probability of observing slower moving or more frequently sampled vehicles rather than sample count, so it makes sense that this can be observed in both cases. It should be pointed out that the trends observed here are hardly general, and would be significantly different under different sampling and/or traffic conditions.

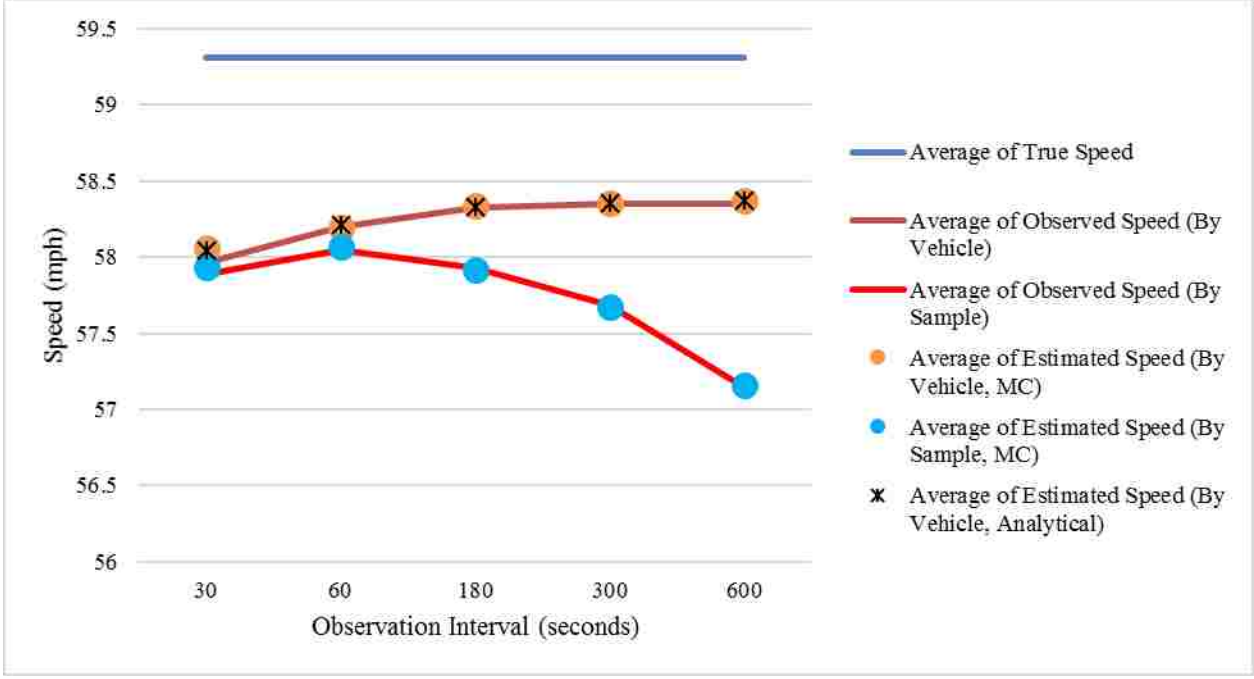

Figure 5-8: Observed Speed vs. Observation Interval Length

Referring to Figure 5-9 and Figure 5-10, it is clear that the difference between method 1 and method 2 is small for shorter observation intervals, and increases as the length of the observation

interval increases. At an observation interval of 30 seconds, there is very little difference at all between the two.

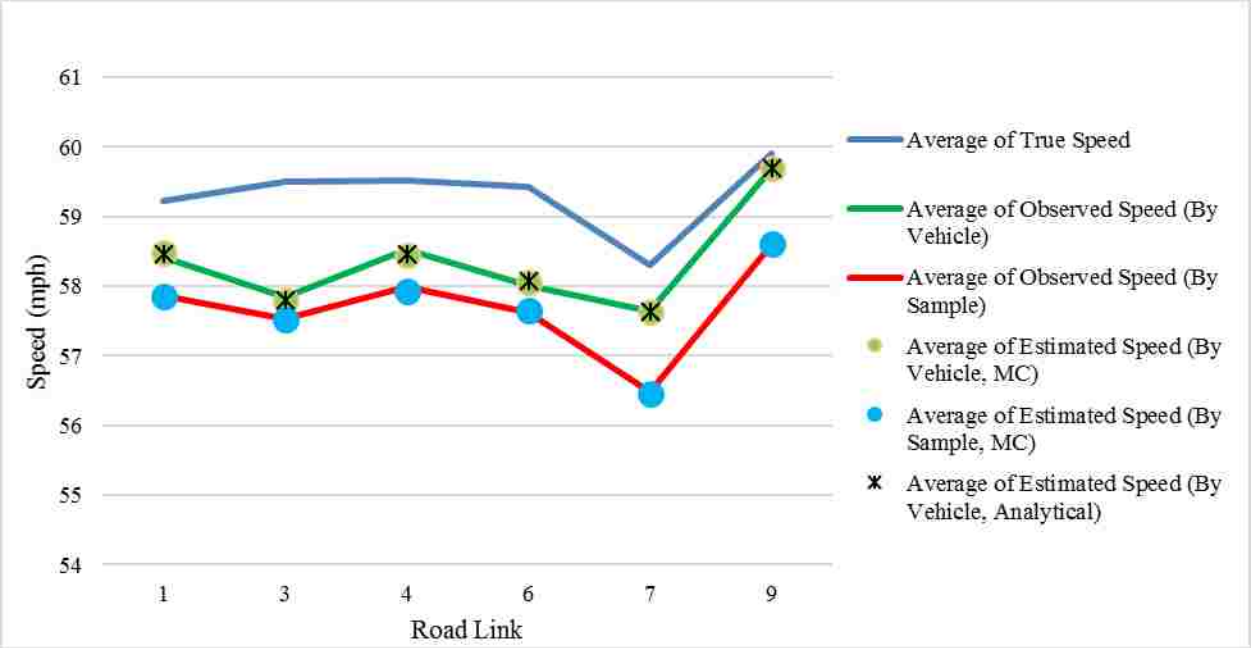

Figure 5-9: Observed Speed for all Road Segments, Observation interval = 300 Seconds

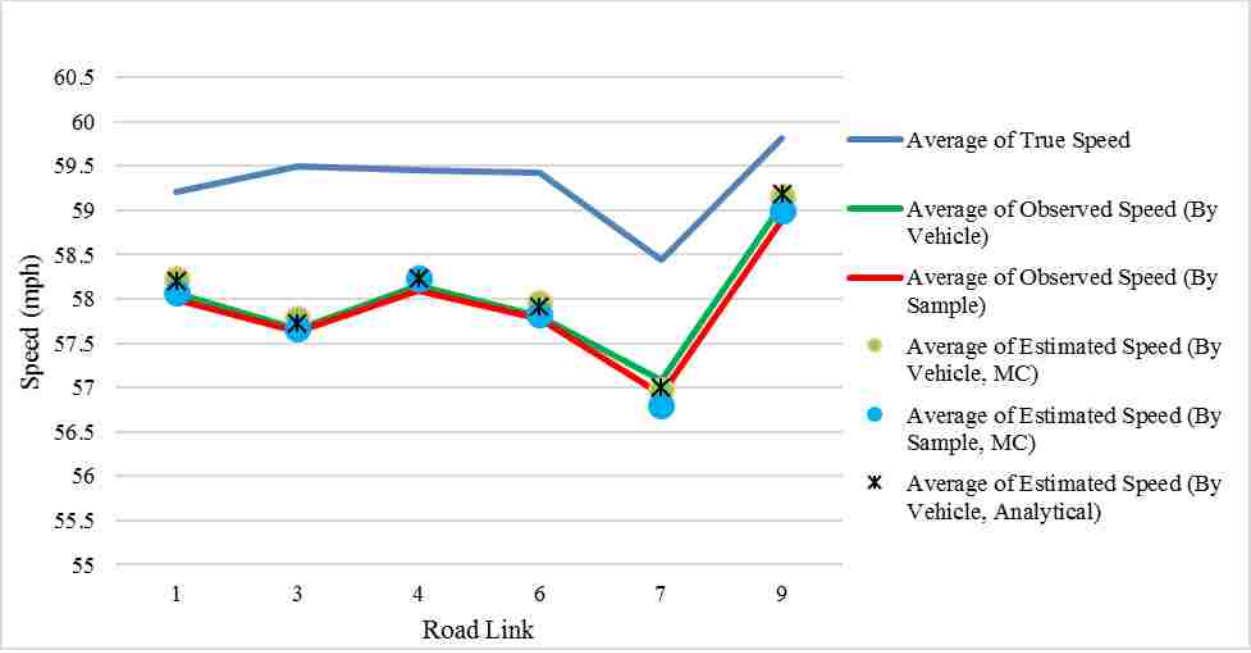

Figure 5-10: Observed Speed fo all Road Segments, Observation Interval = 30 Seconds

In this scenario the observation interval is held at 2 minutes and the penetration rate allowed to vary. As the penetration rate, or the length of the observation interval increases, the measured speed using method 2 gradually approaches the weighted mean, where the weighting parameter is the expected sample count. As this work shows, higher penetration rates do not necessarily lead to improved accuracy, and in some cases can actually exacerbate sampling bias.

Figure 5-11 shows the true, observed, and estimated speeds, averaged over all road links, over a range of penetration rates. Note that, for method 2, the observed speed decreases with penetration rate. As previously mentioned, this is driven by the fact that, as the number of contributing vehicles increases, the number of samples produces by slower and more frequently sampled vehicles has a greater impact on the observed speed. For method 1, observed speed is essentially static, which reflects the increased probability of observing certain subpopulations and slower moving vehicles. Also, referring to Figure 5-12, it is clear that method 1 produces significantly lower bias overall for the test scenarios. Figure 5-4 shows the missing data rate (or 1 – completeness) for all road segments over a range of penetration rates. The decreasing trend is expected, as higher penetration rates will in general produce few missing observations. However, it is worth noting that the bias will in some cases be worse under higher penetration rates, as shown in Figure 5-11.

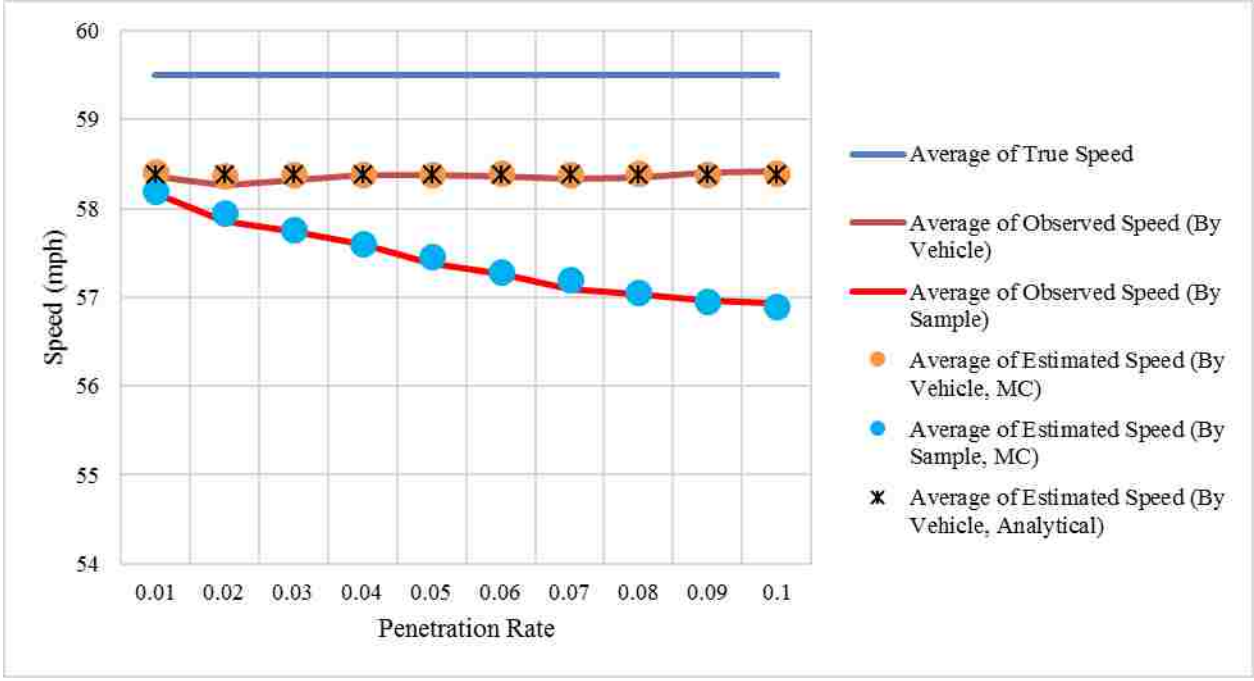

Figure 5-11: Observed Speed vs. Overall Penetration Rates, Road Segment 4

Figure 5-12 shows the observed and estimated speeds using both aggregation methods over all penetration rates and road links. It is clear from this illustration that the proposed methods accurately reflect the true sampling results, and the differences between different traffic and road characteristics. This figure also shows that the sample mean aggregation method is more sensitive to differences in driving and sampling conditions, and to a greater extent over-represents frequently sampled and slower moving vehicles.

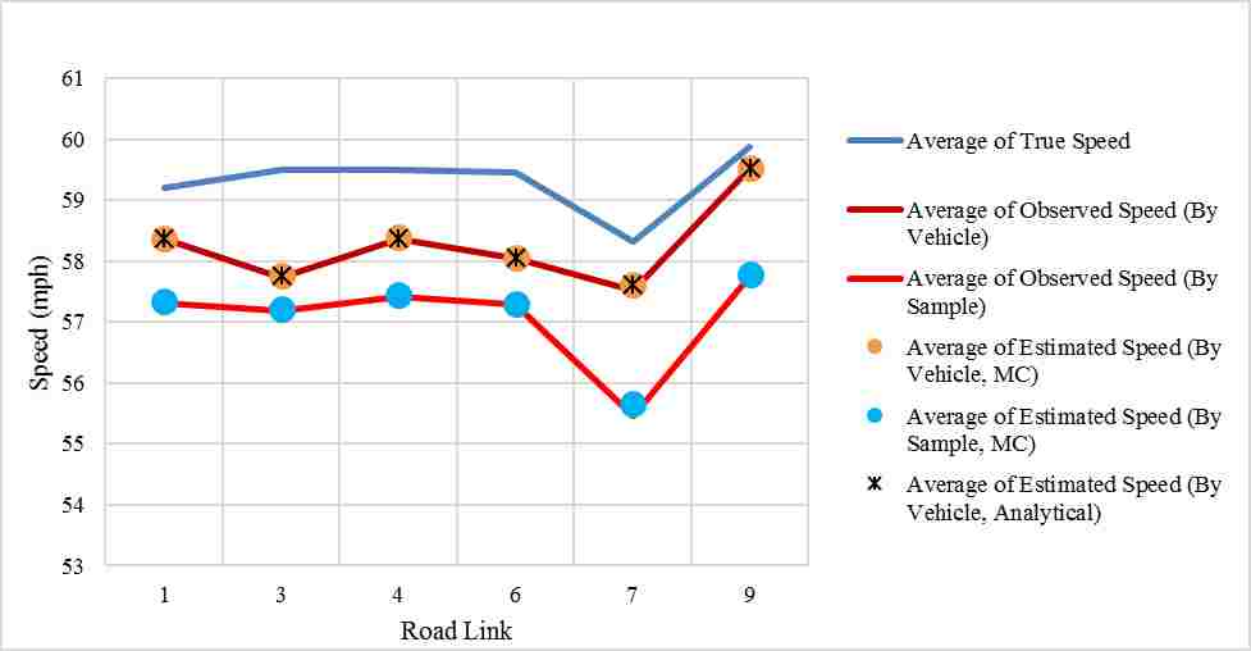

Figure 5-12: Speed vs. Road Link, All Penetration Rates

Figure 5-13 and Figure 5-14 show the observed variance in the mean observed speed across observation intervals over a range of penetration rates for road segment 9 (using aggregation methods 2 and 1, respectively). Because the traffic conditions are essentially the same for all penetration rates, the decreasing variance largely reflects the increasing sample size present in each observation interval. The variance is estimated using the Monte Carlo method. This figure illustrates the fact that, in addition to sample bias and completeness, the variance in the observed speed is key sampling-related factor influencing probe data quality. Likewise, the variance obtained using aggregation method 2 is significantly worse than that obtained using method 1. This can be explained in part by the fact that the presence of a slow moving and frequently sampled vehicle can significantly impact the mean observed speed during a single observation interval for the sample mean aggregation method. The vehicle mean first method would give such a vehicle much lower weight in the aggregation process, thereby reducing the variance. Of course these

results ignore non-sampling sources of error, including GPS noise. It is likely that, in many cases, increasing the sample size would help to further smooth out such sources of randomness.

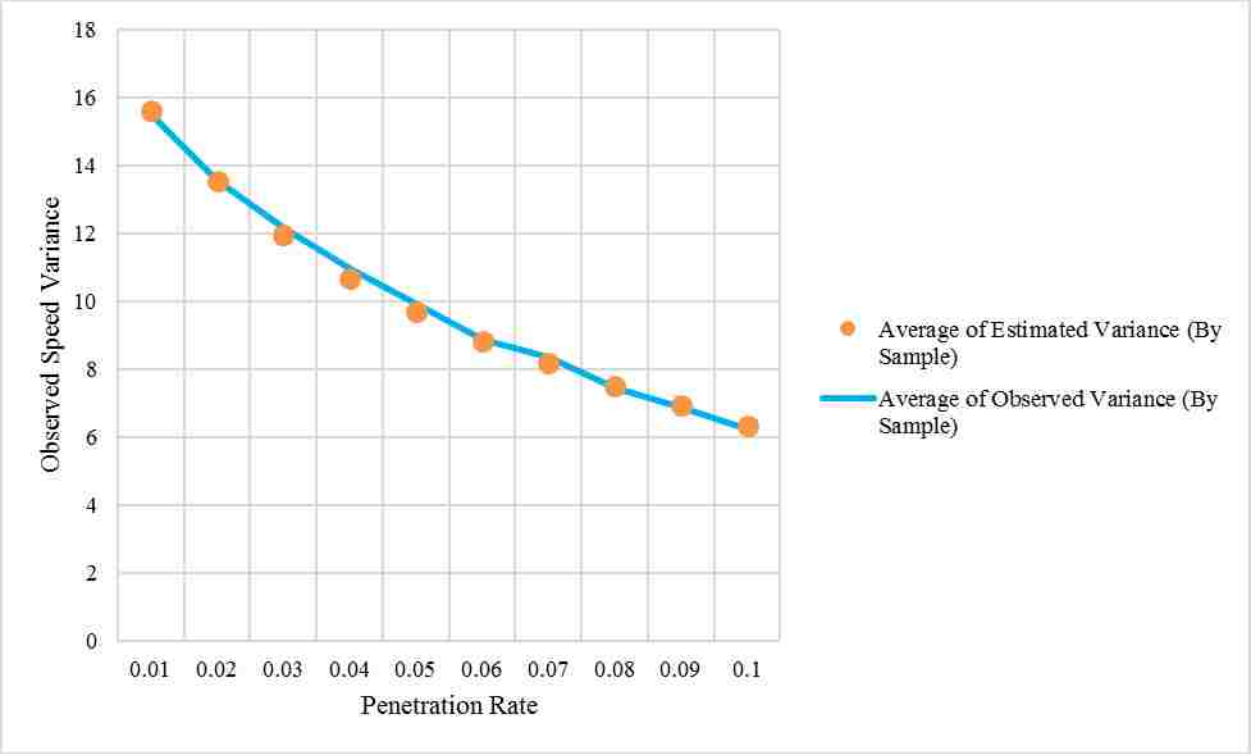

Figure 5-13: Variance in Observed Speed (Method 2) vs. Penetration Rate, Road Segment 9

Figure 5-15 shows the mean variance computed using the analytical approach over all vehicle subpopulations for road segment 9. No method was discussed here for combining the variance obtained for each vehicle population into a single observed speed variance estimate, so this illustration is the simple mean across all subpopulations and sampling intervals. It does not constitute an estimate of the observed variance, but does show that the subpopulation-wise variance estimation method proposed here is representative of the expected observed speed variance for the respective subpopulations. Also note that the mean subpopulation variance is substantially higher than the mean observed variance shown in Figure 5-14. This is because the sample sizes for each subpopulation is much lower, than the overall sample size, resulting in higher mean observed variance.

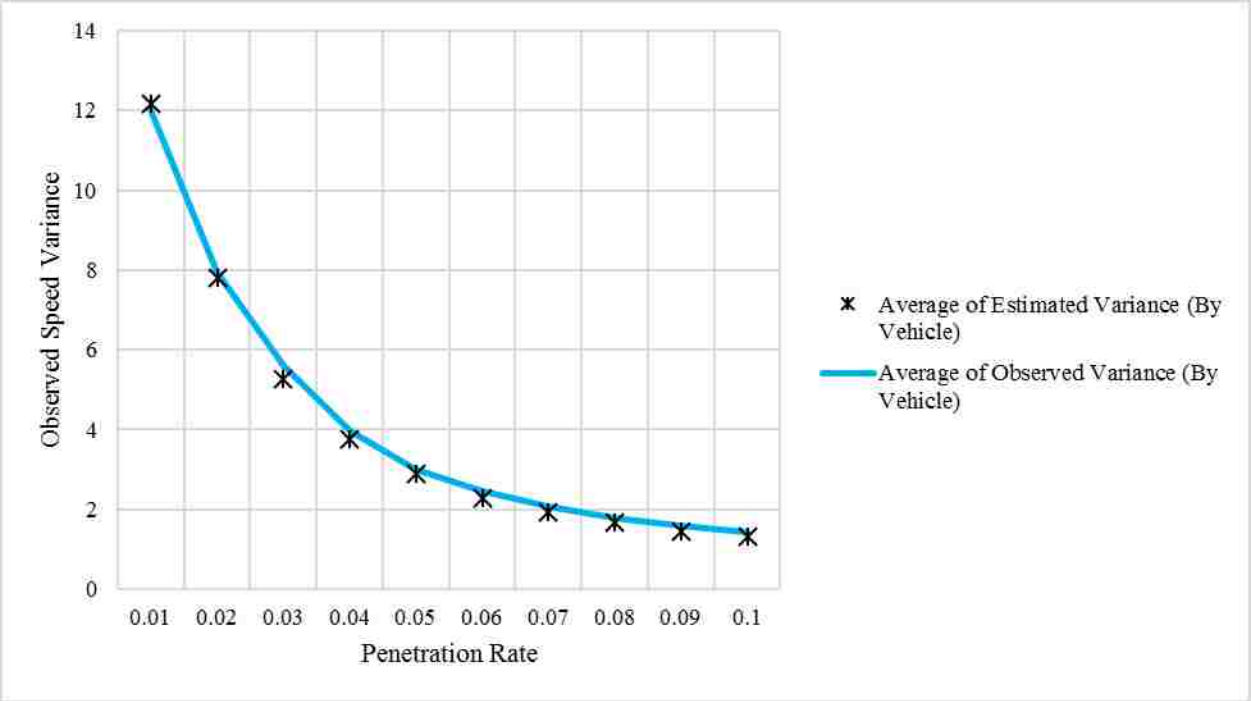

Figure 5-14: Variance in Observed Speed (Method 1) vs. Penetration Rate, Road Segment 9

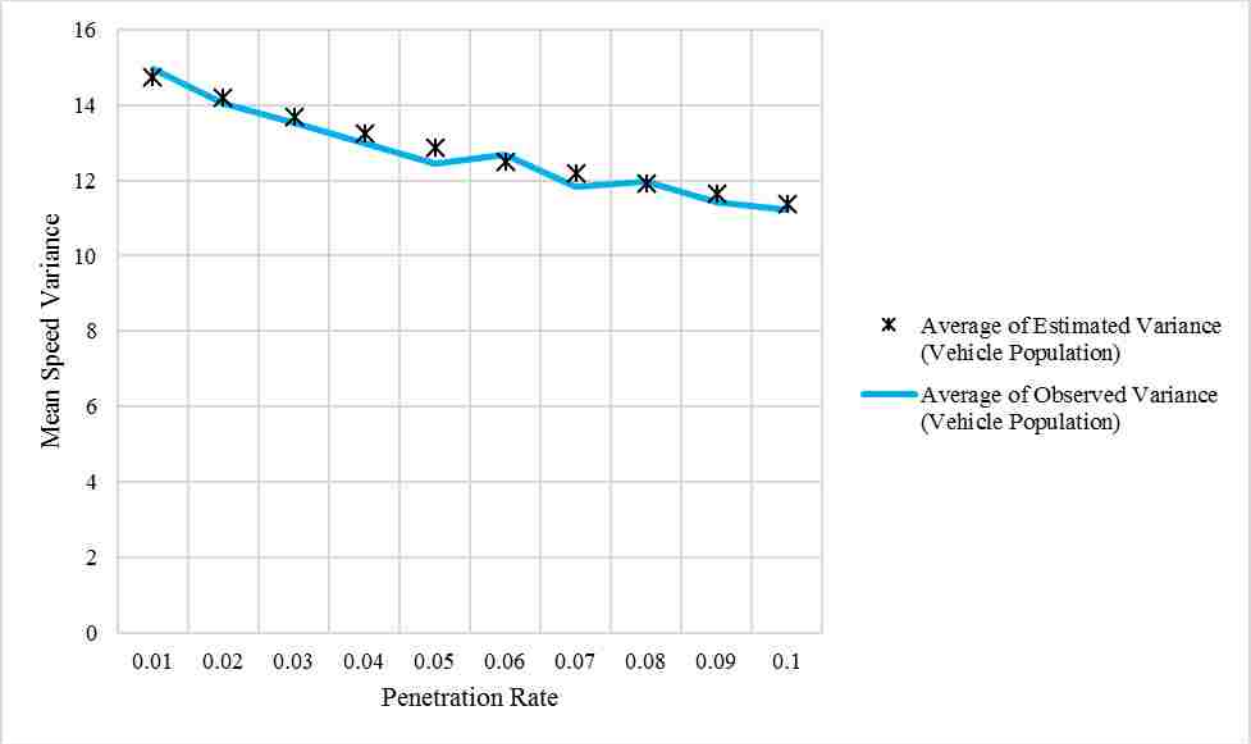

Figure 5-15: Mean Observed Speed Variance Across All Vehicle Populations, Road Segment 9

5.4 Discussion

From the results presented in this section it is clear that the estimation methods are consistently highly accurate for describing high level trends in data completeness, sample size, bias, and variance in the measured values. Nearly identical results were obtained from the Monte Carlo and analytical methods proposed, which supports the physical interpretation of the data collection process that underpins the proposed model. However, it should be stressed that these methods do not constitute a model of the stochastic, time varying dynamics of traffic conditions. For example, little consideration is given to temporal correlation between observations from a given vehicle and the events in consecutive observation intervals. Instead, the results obtained through these methods represent what would be obtained from random samples from a steady state system. The methods presented here are a suitable mathematical model through which to understand quality issues in probe vehicle data, and a suitable methodology for conducting higher-level analysis of existing and planned data collection processes.

The outcome of this validation work suggests that the proposed model would be useful not only in characterizing missing data rates and sample sizes, but also the bias and variance attributable to the sampling mechanism. Thus, there are a number of ways in which the model can be used in practice. First, it provides a mathematical basis for evaluating the data quality impacts of different combinations of probe vehicle source data. With some modifications, this framework can also be used to assess anomalies and quality issues in existing probe vehicle datasets. Further, research in a variety of fields has investigated joint modeling of outcomes and missingness in an informative missing data process. By describing the relationship between the quantity of interest (speed or travel time) and the probability of missingness, the formulation proposed here could provide a suitable foundation on which to develop predictive models that are robust to non-random

missing data patterns. Such models can be applied in data imputation as well as general predictive modeling tasks. More generally, the model described here can be considered a lens through which to view common quality issues in commercial probe vehicle datasets, for example, bias that varies by time and location, missing data prevalence and patterns, and excessive variance values that often occurs during low volume time periods. A more in-depth (though not comprehensive) discussion of potential applications is given in Section 6.

Chapter 6: Applications

6.1 A Predictive Analysis of Probe Vehicle Data Completeness

Without access to a large collection of raw GPS trace data and corresponding commercial link-level traffic data, it is difficult to validate the proposed framework against real-world data. However, this section describes the development of a predictive model for probe vehicle data completeness that broadly supports the framework. In addition, this section shows how the bias and completeness issues outlined in this work can lead to a better understanding of probe vehicle data quality. As this work shows, even ignoring the within-observation bias, assuming the data that is present constitutes a representative sample from the true traffic speed distribution can lead to significant overrepresentation of slower moving time periods.

6.1.1 *Data Description*

The data used in this analysis is from the Federal Highway Administration (FHWA) National Performance Management Research Dataset (NPMRDS). This dataset has been acquired and made available by the FHWA under contract with HERE North America, a commercial data provider. It contains nation-wide link level travel time at five minute intervals, along with road segment (referred by unique Traffic Message Channel or TMC codes) GIS shape files. Unlike some other common probe vehicle datasets, the NPMRDS dataset is raw and unprocessed, and has some obvious quality and consistency issues. In addition, quite a large number of records are missing entirely. In the procurement documents, the FHWA explicitly stipulates that the data should represent only measured travel times rather than some combination of measured and imputed values (Crowder 2012). While the unprocessed and incomplete form of the data may seem

undesirable at first glance, it allows the user to retain full control of quality control, imputation, and uncertainty estimation which should result in more informed analysis and decision making. It should be noted that the data used in this study is from the first version of the NPMRDS, the contract for which ended in 2017.

The data used in this study is from Interstate 5 in western Washington State between Tacoma and the Canadian border. A total of 65 road segments or TMCs, all in the northbound travel direction, are included on this corridor, ranging in length from 0.25-4.6 miles. Data from the month of June, 2015 was used and, with five-minute observation intervals, this constitutes 8640 observations for each road segment. Not all road segments on this corridor were used, as segments had to be selected such that reliable loop detector data was available for all lanes on each segment. A significant number of time intervals are missing in all cases as noted previously. In addition to the NPMRDS travel time data, traffic volumes and travel speeds were obtained for the analysis sections from loop detectors owned by Washington State Department of Transportation. Speed and volume data from loop detectors were aggregated over all sensors present on each TMC segment. A description of the road segments used in this work, including segment length and rate of missingness, is provided in Table 6-1.

Table 6-1: Description of Road Segments

	Missing Rate	**Length (miles)**
Mean	16.5%	1.56
Std. dev.	10.3%	1.00
Min	2.9%	0.25
Max	61.7%	4.65
25th Percentile	9.7%	0.84
50th Percentile	13.4%	1.31
75th Percentile	22.0%	1.87

6.1.2 *A Brief Discussion of NPMRDS Data Completeness*

A few things should be clarified before introducing the modeling methodology. As discussed previously, assuming a constant penetration rate for contributing devices, the completeness of available probe vehicle data for a given segment or TMC is a function of total vehicle time, which is itself a function three interrelated factors: a) the length of the road section or TMC, b) the volume of vehicles on the roadway, and c) the travel speed on the segment. Though sample frequency is also an important factor, at a fixed sampling frequency completeness will still be a function of total vehicle time. As is typically the case for probe vehicle data, the NPMRDS only provides length and segment-wise speed or travel time information. Providing data in this aggregate form, rather than raw vehicle GPS traces, simplifies the task of applying such data in transport analysis, and provides a layer of privacy protection for individual travelers represented in the data. For this reason, we expect that such pre-aggregated data will continue to represent the majority of probe vehicle data used by transport agencies well into the future.

To illustrate these basic characteristics of missingness and related causal factors in probe vehicle data discussed in this work, a series of plots were developed for a section of I-5 in Washington State using only the described NPMRDS dataset. Figure 6-1 shows the relationship between completeness, represented as the percent of 5-minute intervals for which data are available, and road segment length. As expected, Figure 6-1 shows that data completeness is dependent to a large extent on the length of the segment.

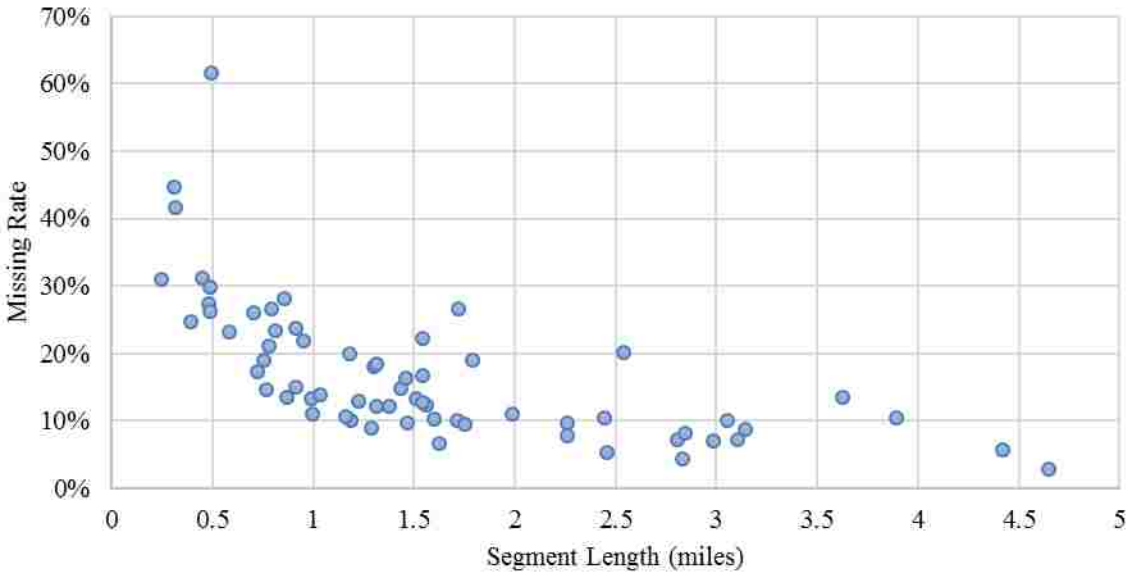

Figure 6-1: Plot of Data Complenteness vs. Segment Length for I-5 Corridor in Western Wasington

Figure 6-2 shows the data completeness versus hourly average speed. In this case, loop detector speeds were aggregated into hourly averages, and plotted against the probe vehicle data missing rate for the corresponding time period and segment. Speeds from loop detectors were used to insure that all time periods and locations were included, as many hour-long time periods are missing entirely from the probe vehicle dataset. It is clear from Figure 6-2 that there is an overarching inverse relationship between speed and completeness, which again is as expected. In comparing Figure 6-1 and Figure 6-2, it should be clear that the there is significant variation between the missing rates of different TMCs, even at a given travel speed, which is not apparent from the aggregate view presented in Figure 6-2 alone. In viewing these plots, it is important to note that there is significant heterogeneity in terms of the relationship between traffic state (i.e. level of congestion) and missingness along the study corridor due to differences in the number of lanes, segment length, and daily traffic characteristics

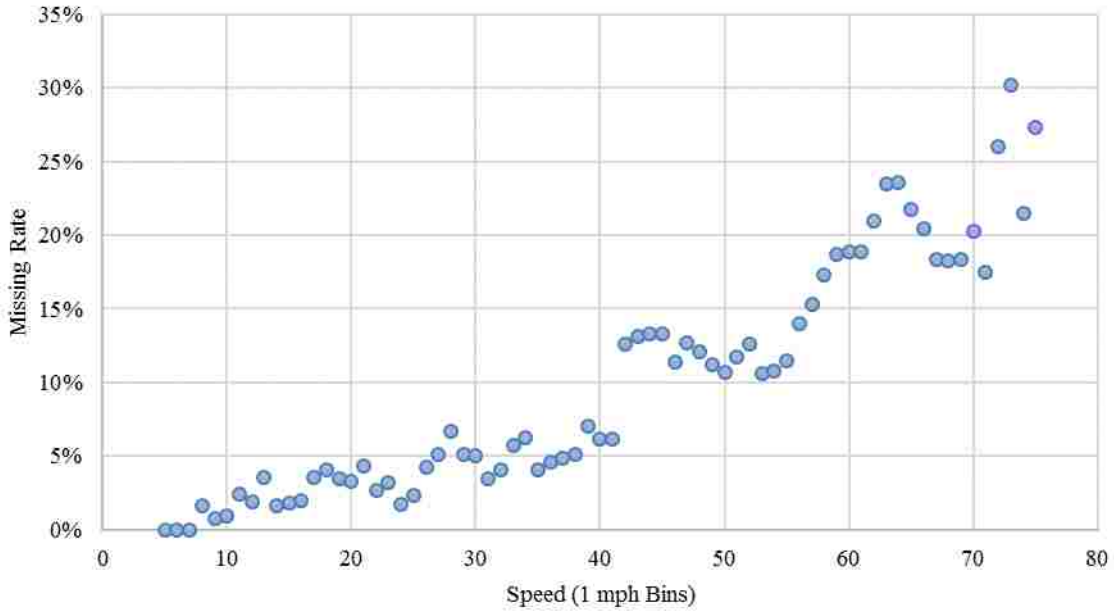

Figure 6-2: Plot of Data Completeness vs. Travel Speed for I-5 Corridor in Western Washington

6.1.3 *Model Development*

This section develops a censored count regression model, starting with the Poisson PMF expression for the number of vehicles appearing on a road segment during an observation interval. Developing a regression model describing the data observation process and missing patterns in real-world probe vehicle data is somewhat problematic because the number of contributing vehicles is not known. Instead, the true number of vehicles is known only when the count is zero, otherwise it can only be said with certainty that 1 or more vehicles were present. To address this, we develop a censored Poisson regression model where the count is censored at 1. In this way, we can estimate the true number of reporting vehicles, as well as the probability of data being observed for various combinations of travel time and traffic volume. Some will recognize this methodology as analogous to censored survival analysis, similar at least in concept to (J. Kim, Mahmassani, and

Dong 2010). However, here we consider the Poisson approach produces good results and is more interpretable given that the data is both left and right censored from a survival analysis perspective.

Here a brief summary of the censored Poisson model is given; for a more in-depth discussion please refer to (Cameron and Trivedi 2012). First, given some threshold c for which the observed count $y_i^* = c$ if the true count $y_i \geq c$, an indicator variable is defined as shown in Equation 6.1

$$d_i = \begin{cases} 1, & y_i \geq c \\ 0, & otherwise \end{cases} \quad 6.1$$

To write the log likelihood expression, first observe that $Pr(y_i \geq c)$ can be computed as 1 - $Pr(y_i < c)$ which can be described as shown in 6.2 as a Poisson process:

$$\Pr(y_i \geq c) = \sum_{j=c}^{\infty} \rho^j/(j!)\exp(-\rho) = 1 - \sum_{j=0}^{c-1} \rho^j/(j!)\exp(-\rho) \quad 6.2$$

This gives the expression shown in Equation 6.3 for the Poisson regression log likelihood, where 1 - $Pr(y_i < c)$ is defined as in Equation 6.2 and $\rho_i = exp(\beta' x_i)$ arises from the Poisson log link function with predictor set X, response variable Y, and model coefficient vector β (Famoye and Wang 2004; W. H. Greene 2005). The result is that the conventional log likelihood expression is nonzero when the count is observed (the first term in Equation 6.3), and the 1 - $Pr(y_i < c)$ expression is nonzero when the count is equal to or larger than the threshold (second term in Equation 6.3).

$$LogL(\beta|X,Y) = \sum_{i=1}^{n}\{(1-d_i)[y_i \beta x'_i - \exp(\beta x'_i)] + d_i Log(\Pr(y_i \geq c))\} \quad 6.3$$

The model parameters can then be estimated using one of several methods, in this work we rely on the VGAM package in R (Thomas W. Yee 2015; T. W. Yee and Wild 1996), which applies the Newton-Raphson algorithm.

With a relatively homogeneous set of locations and time periods, the only predictor variable in the case should be the product of the incoming vehicle volume and the expected travel time across the segment. This has intuitive interpretation on the basis of a Poisson arrival process (see section 4.2.1), and removes the need to explicitly consider the difference in length and speed profiles across the various road segments. Under the assumption that the product of expected travel time and arriving contributing vehicle count is linearly related to the product of expected travel time and overall vehicle volume, we take the log of this term as the predictor. We may expect that the overall percentage of contributing vehicles on the roadway traffic characteristics and time period, and to account for this we include both a) dummy variables representing am/pm and weekday/weekend time periods, and b) interaction between the time period dummy variables and the log of the product of arriving volume and expected travel time.

It bears noting that an assumption inherent to this model formulation is that, if a contributing vehicle is present at any point in a time period, it will be observed. In reality, probe vehicles report their location or speed at discrete time intervals rather than continuously. Thus, a vehicle that is reports its state information less frequently is less likely to be observed, all else being equal. However, the true sampling frequencies are not known and, for this reason, we follow the above described formulation expecting that the rate parameter lambda will indicate the expected sample count, rather than the vehicle count. Note that the total number of samples over a time interval represents superposition of multiple independent point processes (that is, one or more vehicles producing updates at different time intervals), which makes intuitive sense as a Poisson process (M. Ferman, Blumenfeld, and Dai 2005). In any case, the sampling rate implied by this model formulation can be meaningfully interpreted as an effective sampling rate, and the estimated observation probability should be unbiased.

6.1.4 *Modeling Results*

The regression model estimated as described in the previous section is summarized in Table 6-2. All terms are significant at the 0.01 level, and coefficient estimates generally make intuitive sense. For example, we can interpret these coefficients as indicating the probability of observing a contributing vehicle increases with ρ, on weekdays, and in the evening.

Table 6-2: Regression Model Summary

Term	**Coef. Estimate**	**Std. Error**	**z value**	**Pr(>|z|)**
Intercept	-1.7381	0.0304	-57.10	<2E-16
log(ρ)	0.5464	0.0074	73.52	<2E-16
weekday	0.1418	0.0275	5.15	2.65E-07
hour	-0.1596	0.0223	-7.15	8.84E-13
log(ρ) x weekday	0.0204	0.0070	2.94	0.00329
log(ρ) x hour	0.0686	0.0056	12.25	<2E-16

An important step in this work is to find a way to validate the proposed methodology in a way that is instructive to the problem at hand, namely to answer the question of whether this method can accurately quantify the relationship between volume, travel time, and data completeness. Though there are a range of statistical test that have been applied to such models, prediction accuracy cannot be assessed because the true counts are not observed. Here we have approached this problem by generating empirical cumulative distribution (CDF) curves (Pr(vehicle count $\geq$ 1)) from the observed data missing rates, and plotting these against the predicted CDF from the Poisson regression. This is done for each possible combinations of weekday/weekend and AM/PM, a subset of which is shown in Fig. 3. The regression model is trained using 50% of the available data, and the CDF plots generated from the remaining 50%. The empirical CDF is calculated as the average completeness (as a fraction) for each ρ range, with bin sizes of 20 (in terms of $vehicles \times travel\ time\ in\ minutes$).

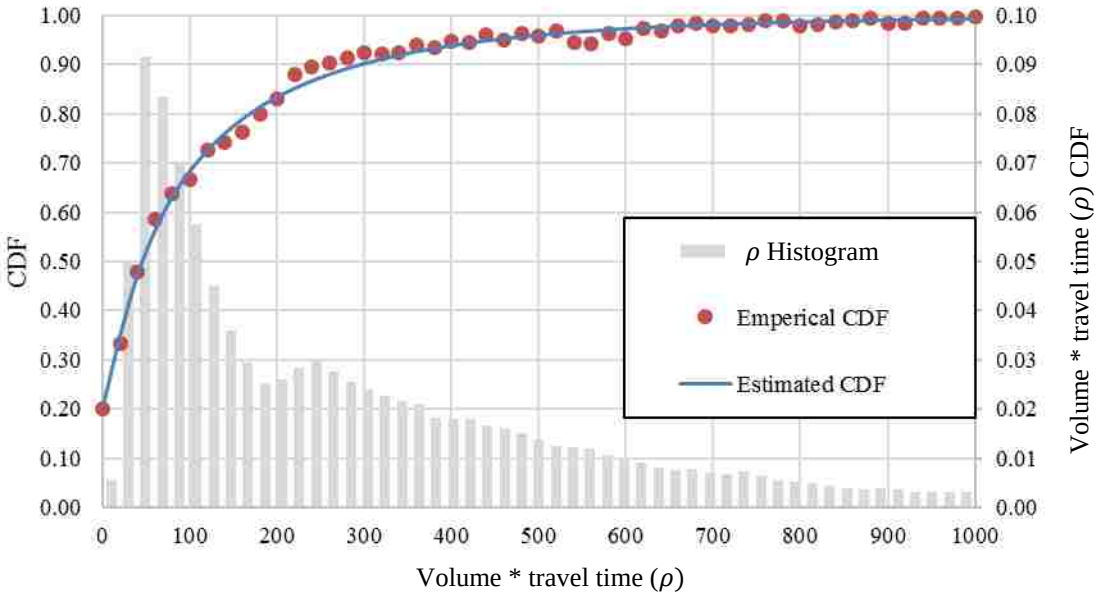

Figure 6-3: Predicted and Emperical CDF for Data Complenteness During Weekday AM Time Period, With ρ Histogram

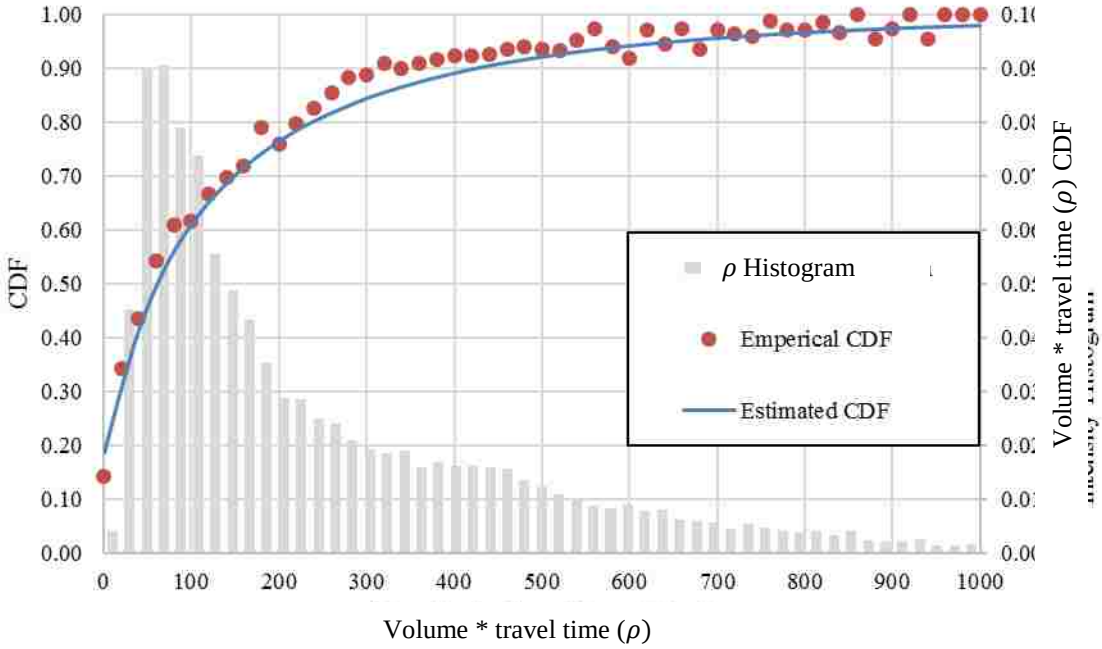

Figure 6-4: Predicted and Emperical CDF for Data Completness During Weekend AM Time Period, With ρ Histogram

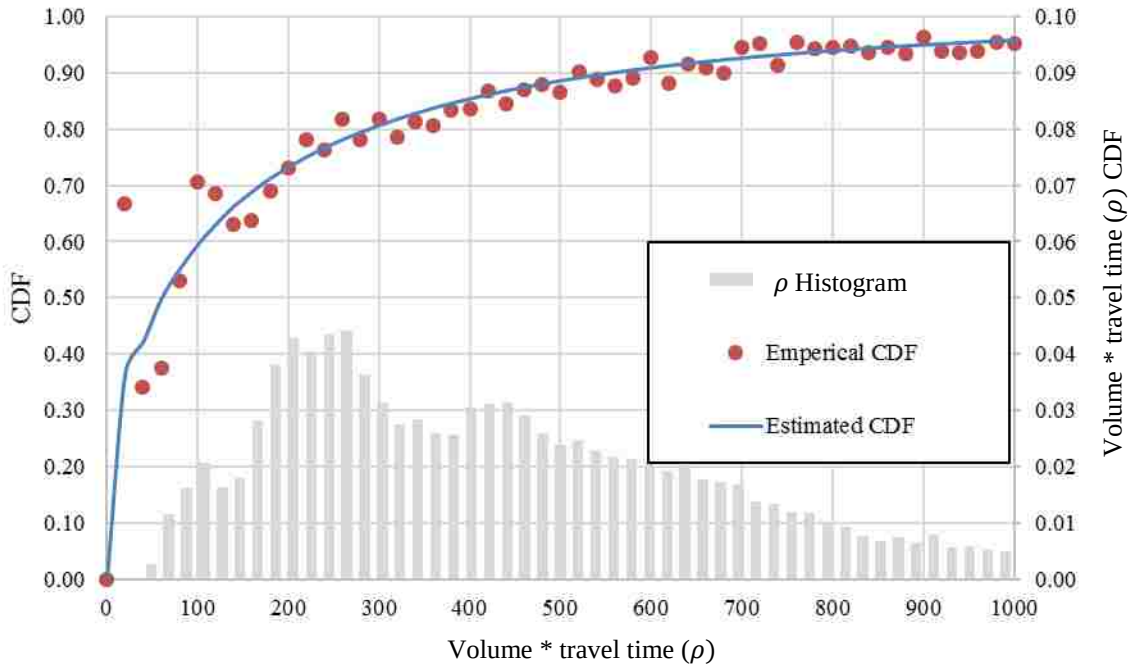

Figure 6-5: Predicted and Emperical CDF for Data Completness During Weekend PM Time Period, With ρ Histogram

These plots show a good fit with the empirical data, and demonstrate that the CDF curve is well-described by the Poisson distribution. The least convincing fit is the weekend PM time period, which may be the result of behavioral and temporal heterogeneity of weekend evening travelers. The ρ histograms indicate that, to varying degrees, the majority of observations fall in regions that will be expected to be missing a significant number of observations.

To present the probability of observing data as a function of both volume and travel speed, contour probability plots are used as shown in Figure 6-6 and Figure 6-7. To understand these results, consider a segment 1 mile in length with 2 travel lanes, an average speed of 60 mph, and entering volume of 1000 veh/hr/ln. Observations for this scenario would be under 77% complete for a 2 lane road, and over 84% for a 3 lane road. Were the speed reduced to 48 mph, the completeness would increase to 81% for a 2 lane road and 87% for a 3 lane road. Thus, especially

in lower volume time periods when travel time measurements are made based on a small number of vehicles, it is easy to see how the data can be biased toward slower moving vehicles.

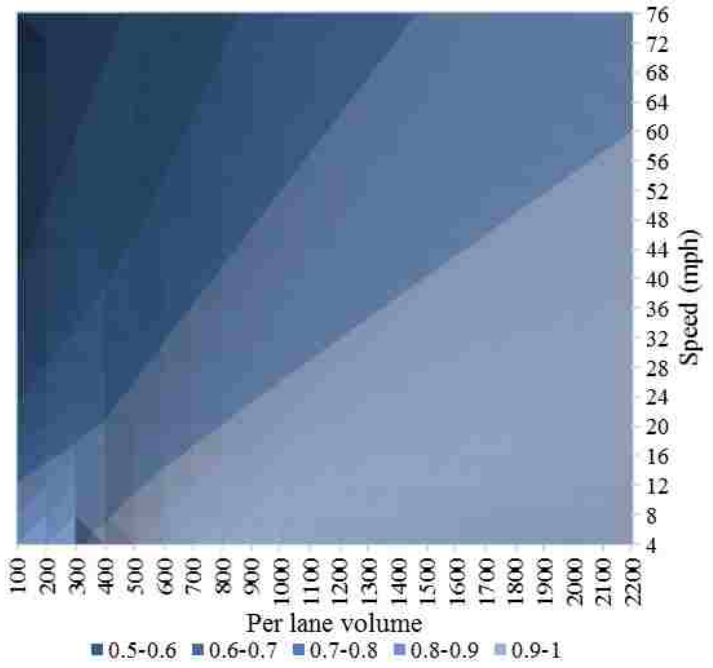

Figure 6-6: contour Plot of Data Completeness as a Function of per-lane Traffic Volume and Speed for 1 mile, 2 lane Road Segment

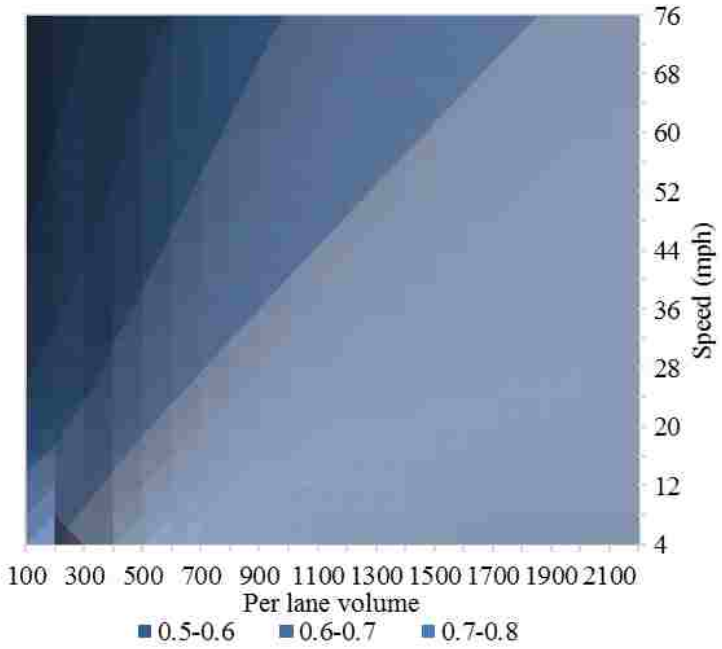

Figure 6-7: Contour Plot of Data Completeness as a Function of Per-lane Traffic Volume and Speed for 1 mile, 3 lane Road Segment

Another result that is likely to bias travel time reliability and other engineering analysis occurs during moderate to high volume periods. That is, as the demand volume approaches the capacity of a facility (say, 2000 veh/hr/ln on a freeway), the probability of breakdown (traffic entering a congested state) increases dramatically (Brilon, Geistefeldt, and Regler 2005). If, for example, half of all peak period travel (> 1700 veh/hr/ln) results in significant slowdowns and a resulting decrease in travel speed from 75 to less than 15 mph, the congested data will be nearly 100% complete whereas the non-congested data will be less than 91% complete. This may seem counterintuitive, as we might expect the total number of vehicles passing through a segment to be higher when traffic is not congested and, if a certain fraction of all vehicles are reporting data, the likelihood that a contributing vehicle pass through the segment should increase. However, the maximum headway between contributing vehicles for the data to be complete in most cases is just

below the time interval length plus the travel time. Thus, up to the point of traffic breakdown, an increase in volume will result in a decrease in the measured headway-as well as an increase in the maximum headway allowable for data completeness if the travel time increases correspondingly. As traffic enters a congested state, if the classical quadratic form of the volume-speed relationship can be assumed, the increase in travel time will be fractionally less than the drop in time headway. Thus, we may expect the rate of data missingness to increase somewhat, especially for short segments. That said, it cannot be assumed that the fraction of vehicles on the road contributing to the location services is static. While this phenomenon was not studied in this work, it might be assumed that drivers stuck in traffic will look to mobile location services to find a better route, or simply to check the severity and extent of the congestion. This distinction in missingness mechanisms is important, because the observations on the high end of ρ could represent either relatively stable traffic conditions on a long road segment, or near breakdown conditions for a road segment that is shorter or with fewer lanes. Resolving this heterogeneity may help to explain the variability in the empirical data in more congested traffic.

6.1.5 *Discussion*

This section presented a modeling methodology which describes the probability of observing data using the Poisson distribution. The justification for the choice of models is based in part on an intuitive understanding of vehicle arrivals as a Poisson process, and supported mathematically by established queuing theory. The results of this analysis suggest that, as expected, the missing patterns are very closely related to both travel time and vehicle volume, which could have significant impacts on any statistical analysis based on this dataset. This also adds a layer of complexity to imputation efforts, as the factors describing the missing patterns may not always be available. For example, in most cases, volume measurements are only available for roadways with

fixed mechanical sensors.

In our analysis, it is clear that there is some temporal heterogeneity that is not sufficiently accounted for in the model. Additionally, a number of anomalies were identified in the dataset which could add additional complexity to any analysis based on this data. Understanding this, and investigating the missing rates under moderate to high traffic conditions (which are of the greatest interest in many cases), is a subject for future work. The work described in this section provides a justification for the mathematical formulation presented in Chapter Chapter 4:, and offers a more complete understanding of the problem at hand and the potential impacts, and constitutes a starting point for more informed transportation analysis using mobile location data. This type of data overcomes many of the shortcomings of fixed mechanical sensing but, like fixed sensors, it comes with a unique set of considerations for data quality and uncertainty that must be addressed.

6.2 Planning for Completeness and Sample Size

This case study illustrates how the data completeness for a set of real-world road links can be assessed for a range of sampling scenarios. This example is intended to show how the proposed methodology can be used to design future experiments involving probe vehicles for generating real-time traffic information. A number of previous publications have attempted to describe the minimum penetration rate of probe vehicles needed to achieve some fixed level of completeness (S.M. Turner and Holdener 1995; Boyce, Hicks, and Sen 1991). However, such work is generally empirical in nature, relying on real-world vehicle probe data or simulation. One notable exception can be found in (M. A. Ferman, Blumenfeld, and Dai 2003), but the model provided in this paper was quite simple and ignored several key factors, including the influence of sampling rate and road segment length. In this work, the data completeness for any number of hypothetical scenarios can be estimated without the need for extensive preliminary data collection or simulation.

6.2.1 *Data Description*

The data used in this study is the same as used in Section 6.1, collected on Interstate 5 near Seattle, WA during the month of June, 2015. Here, only the weekday evening hours from 4:00PM to 8:00PM are considered. The loop data is used to obtain traffic volumes and speeds as inputs to the proposed methodology. For each road segment defined by the NPMRDS TMC standard, the data is divided into 10 ρ (traffic volume * travel time) quantiles and the average volume, speed, and completeness are computed for each quantile. These quantile bins are intended to represent a reasonable demarcation of expected traffic states over the study period and, though there is obviously some variation in volume and speed within each bin, should be adequate for planning level analysis. In addition, using the same TMC road segment definitions, a range of hypothetical scenarios are investigated.

6.2.2 *Experimental Set Up*

In this study it is assumed that multiple vehicle subpopulations will be present, each with a fixed sampling frequency and penetration rate, as well as potentially a unique desired speed distribution. The objective is to investigate a range of sampling scenarios, such that it will be possible to show the minimum requirements to achieve a given level of completeness or average number of samples for each traffic state bin (as defined by the ρ quantiles discussed previously). To do this, the methodology described in Section Chapter 4: is applied to a grid of penetration rates, observation interval, and sampling frequency distributions. The inputs to this methodology includes the sampling rate distribution, the overall penetration rate, and the observation interval as well as, for each subpopulation, the mean and variance of the speed distribution. Thus, it is trivial to estimate

completeness and sample counts for a range of scenarios as well as to estimate sensitivity to different inputs.

All of the scenarios investigated are based on the probe vehicle population shown in Table 6-3. Note that each subpopulation is defined by a distinct combination of sampling rate and speed distribution, so each subpopulation could be a combination of multiple providers. Likewise, two subpopulations need only differ in terms of speed distribution, sampling frequency, or both to be considered distinct. Thus, the penetration rate and population fraction may in fact be the weighted mean and sum, respectively, over multiple smaller, non-overlapping subpopulations. The scenario is arbitrary, but is intended to show how a mixed vehicle population can be considered in the methodology. Note that the population fractions and speed adjustments are designed such that the true mean speed over all subpopulations is simply the overall observed mean speed. This is done to insure that the scenario represents a reasonable approximation of reality. This table shows six distinct vehicle subpopulations, each of which is associated with the following characteristics:

- Vehicle Class: Vehicle type; passenger car (PC), Commercial Vehicle (CV), or Transit Vehicle (TV)
- Sampling Interval: Number of seconds between samples
- Subpopulation Fraction (all): the fraction of the entire vehicle population made up of vehicles from this subpopulation
- Population fraction (probes): The fraction of vehicles in the probe vehicle population that is made up of vehicles from this subpopulation
- Effective Penetration Rate: The fraction of the overall vehicle volume made up of probe vehicles from this subpopulation

- Speed adjustment: The average deviation of the subpopulation mean speed from the population mean speed. That is, if the population mean speed is 62 mph, a subpopulation speed adjustment of -2 will result in a subpopulation mean speed of 60 mph.
- Speed standard deviation: The standard deviation of the mean vehicle speed for each subpopulation (assumed to be fixed)

Because the penetration rate varies between the different subpopulations, the probe vehicle population is not strictly representative of the overall vehicle population in this scenario. This is scenario is based on the assumption that commercial and transit vehicles are more likely to be equipped with mobile GPS and contribute to the dataset. Thus, though they constitute a smaller fraction of the overall vehicle population, these vehicles represent a comparatively large fraction of the probe vehicle population.

It is highly likely that the vehicle type, sampling frequency, and desired speed distributions will shift significantly depending on the time of day and/or day of the week, as well as gradually over longer time horizons. Because of this, only the evening hours between 4:00PM and 8:00PM are considered here, though in practical application it would be advisable to consider all potential scenarios.

Table 6-3: Base Scenario

Parameter	**Vehicle Subpopulation**					
	1	2	3	4	5	6
Vehicle Class	CV	TV	PC	CV	PC	PC
Sampling Interval (sec/sample)	30	60	5	10	30	60
Subpopulation Fraction (all)	0.05	0.02	0.15	0.20	0.30	0.28
Subpopulation Fraction (probes)	0.457	0.365	0.034	0.027	0.062	0.054
Effective Penetration Rate	0.0228	0.0073	0.0051	0.0054	0.0186	0.0151
Speed Adjustment (mph)	-2	-2.5	0	-1.81	0.51	1.28
Speed Standard	3	3.5	4	3.6	3	4.2

For each 5 minute epoch in the measured data, the mean speed of each subpopulation described in Table 6-3 is computing by adding the speed adjustment to the population mean speed. The scenarios to be considered are as follows.

Selecting an Optimal Observation Interval: In this scenario, the base case is considered (Table 6-3), and a range of observation interval values are explored. The objective in this case is to select an observation interval to insure a minimum number of samples and completeness is achieved.

Selecting Data Providers: In this scenario, an additional data provider is considered in addition to the base case providers. The objective is to determine the value of the additional provider, decide whether it is necessary to purchase the additional data.

For the purpose of determining completeness and sample size, only vehicles that produce at minimum a single state update on a road segment during a time period are observed. That is, vehicles that pass through a segment without being observed because of the scheduling of state updates are not considered to be observed in sample size or completeness calculations. It is entirely possible that the speed of a vehicle can be computed between subsequent updates and the results interpolated between segments. However, this analysis assumes that vehicles must be observed on a segment in order for the state information to be relevant to that segment. This is one possible scenario; it would of course be trivial to include all vehicles passing through a segment during an observation interval.

6.2.3 *Results: Scenario 1*

For the base scenario, the first step is to define a set of criteria for selecting the most appropriate observation interval. These could be based on the analytical requirements of the project, cost, and

any number of other considerations that will vary from project to project. For this work, the criteria are (rather arbitrarily) selected as follows:

- The minimum (over all road sections) average vehicle count should be above 3 for the top 50% of ρ bins
- The minimum (over all road sections) average sample count should be above 5 for the top 50% of ρ bins
- No more than 5% of all observation intervals should be missing due to lack of data
- The minimum (over all road sections) average completeness should be above 85% for every ρ quantile bin

Figure 6-8 shows the minimum average completeness for every combination of observation interval and ρ quantile. It is clear from this plot that an observation interval 0f 180 seconds is the minimum length that satisfies the minimum completeness requirements (above 85% for all ρ quantile bins). Similarly, Figure 6-9 shows the minimum average vehicle and sample count for the top 50% of ρ bins for each observation interval, and it is clear that 180 seconds is the minimum length that satisfies the sample size requirements.

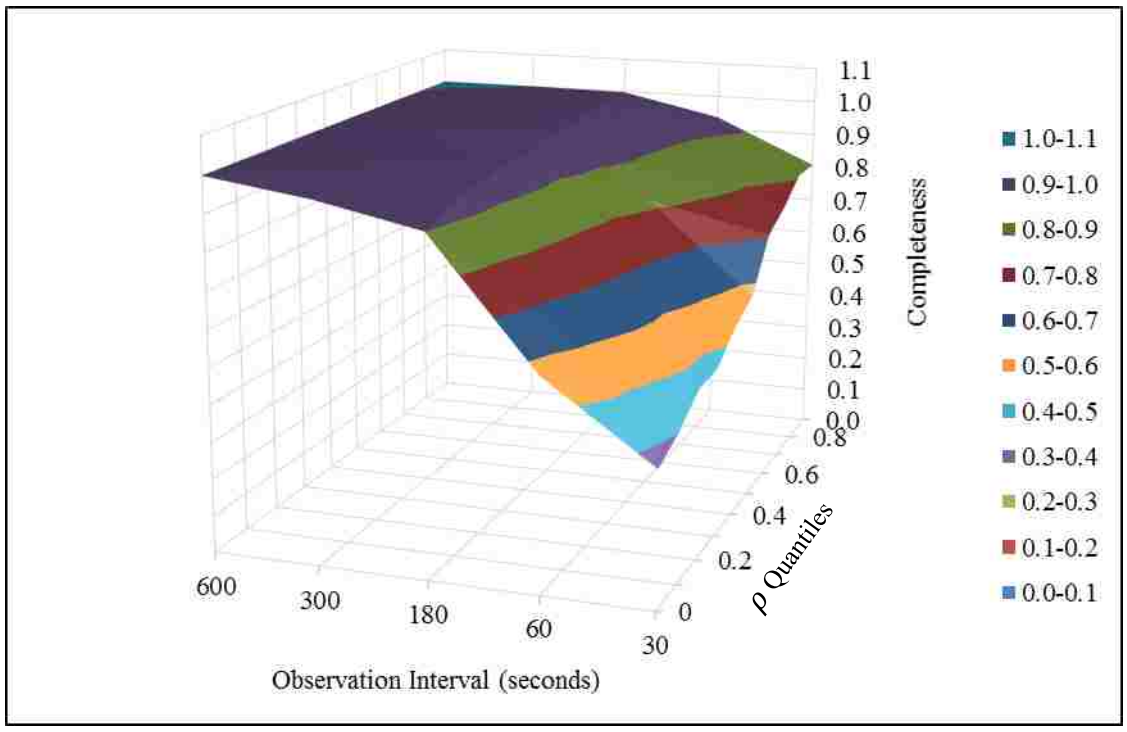

Figure 6-8: Minimum Average Completeness vs. Observation Interval and ρ

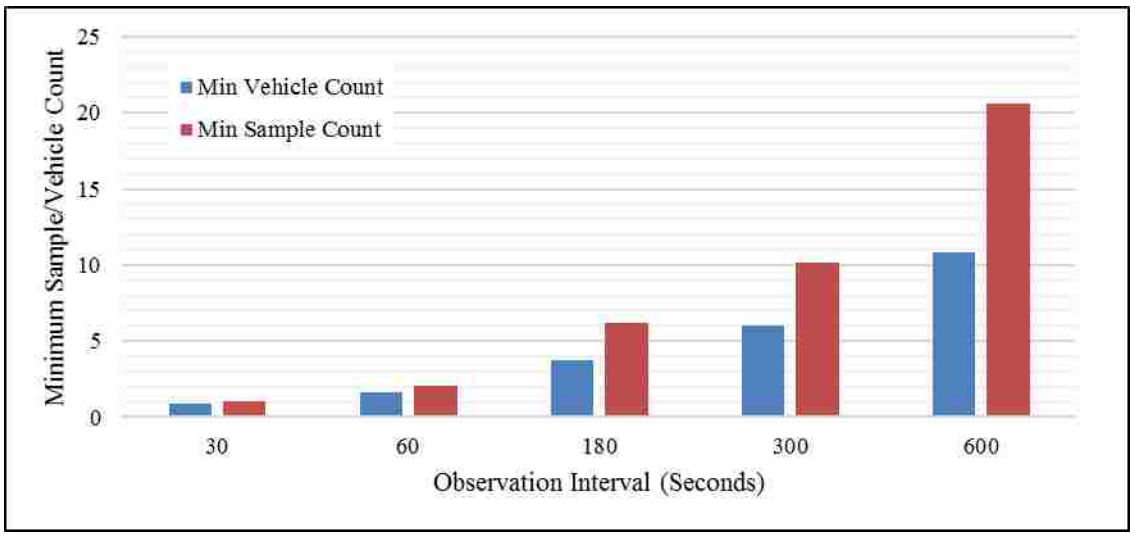

Figure 6-9 Minimum Average Sample Size and Vehicle Count vs. Observatio Interval (Top 50% of ρ Bins)

6.2.4 *Results: Scenario 2*

Having selected a minimum observation interval according to the chosen set of criteria, the next objective is to determine whether there is any substantive benefit to acquiring an additional probe vehicle data set. The additional data provider considered here represents vehicles with identical behavior and sampling interval to one of the existing data providers. As a result, the fourth vehicle category from the base scenario is now split into two different vehicles. Thus, the vehicle population represented by the fourth vehicle category now represents only 10% of all vehicles, but the within-subpopulation penetration rate is doubled to 0.054. The remaining vehicles from that subpopulation are now represented as an additional vehicle category representing 10% of all vehicles and with a penetration rate of 0.10. The revised probe vehicle population is described as shown below in Table 6-4. Note that adding this new provider has the result of increasing the overall penetration rate of all probe vehicles from approximately 0.0744 to approximately 0.0844, a significant increase that will increase the sample size by more than 13%.

Table 6-4: Expanded Dataset Scenario

Parameter	**Vehicle Subpopulation**						
	1	2	3	4	5	6	7
Vehicle Class	CV	TV	PC	CV	PC	PC	CV
Sampling Interval (sec/sample)	30	60	5	10	30	60	10
Subpopulation Fraction (all)	0.05	0.02	0.15	0.10	0.30	0.28	0.10
Subpopulation Fraction (probes)	0.457	0.365	0.034	0.054	0.062	0.054	0.10
Effective Penetration Rate	0.0228	0.0073	0.0051	0.0054	0.0186	0.0151	0.01
Speed Adjustment (mph)	-2	-2.5	0	-1.81	0.51	1.28	-1.81
Speed Standard	3	3.5	4	3.6	3	4.2	3.6

Figure 6-10 shows the minimum average completeness for the base and expanded scenarios, using the observation interval selected in the previous subsection. Figure 6-11 compares the

completeness for the top 50% of ρ quantiles, and Figure 6-12 shows the comparison of vehicle counts.

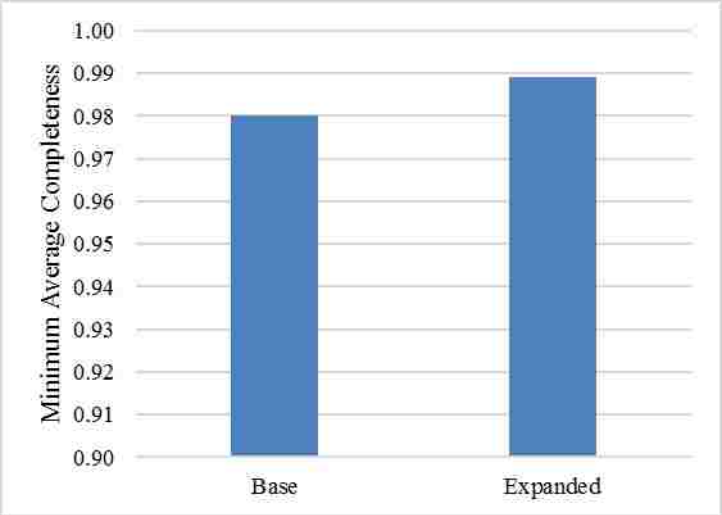

Figure 6-10: Minimum Average Completeness for the Base and Expanded Scenarios

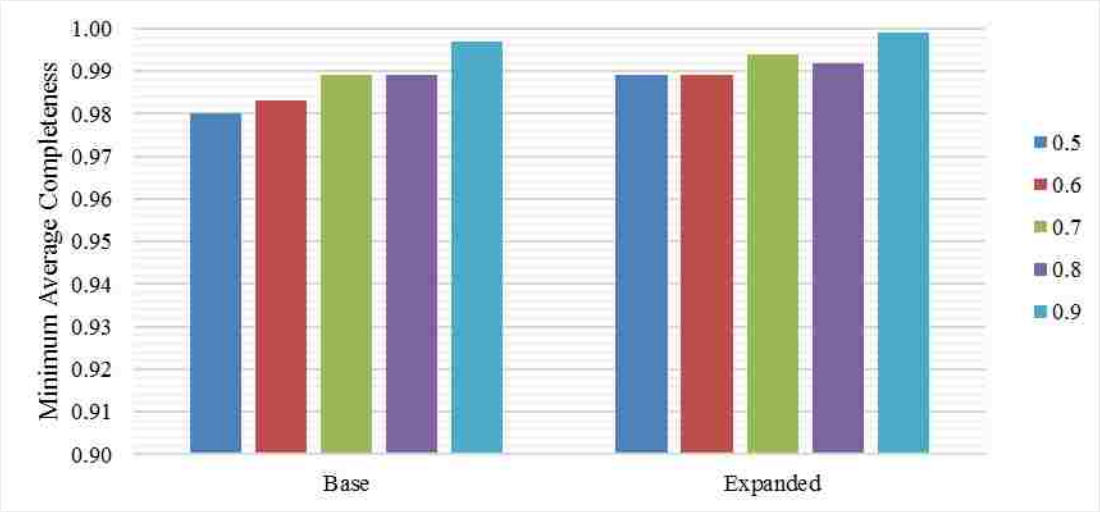

Figure 6-11: Minimum Average Completeness for the Top 50% of ρ Bins

Having determined (not unexpectedly) that both scenarios will satisfy the established criteria, an interesting question would be whether or not the expended dataset will allow us to reduce the observation interval from 180 to 60 seconds. Figure 6-12, Figure 6-13 show comparisons of the minimum average vehicle count and completeness, respectively, for the top 7 ρ bins. It is clear from this comparison that, while the expanded dataset does increase the vehicle count and

completeness significantly, the minimum criteria cannot be satisfied at an observation interval of 60 seconds. The minimum average vehicle count only reaches 3 at the top ρ bin, and the minimum average completeness falls below 85% on the 7th ρ bin.

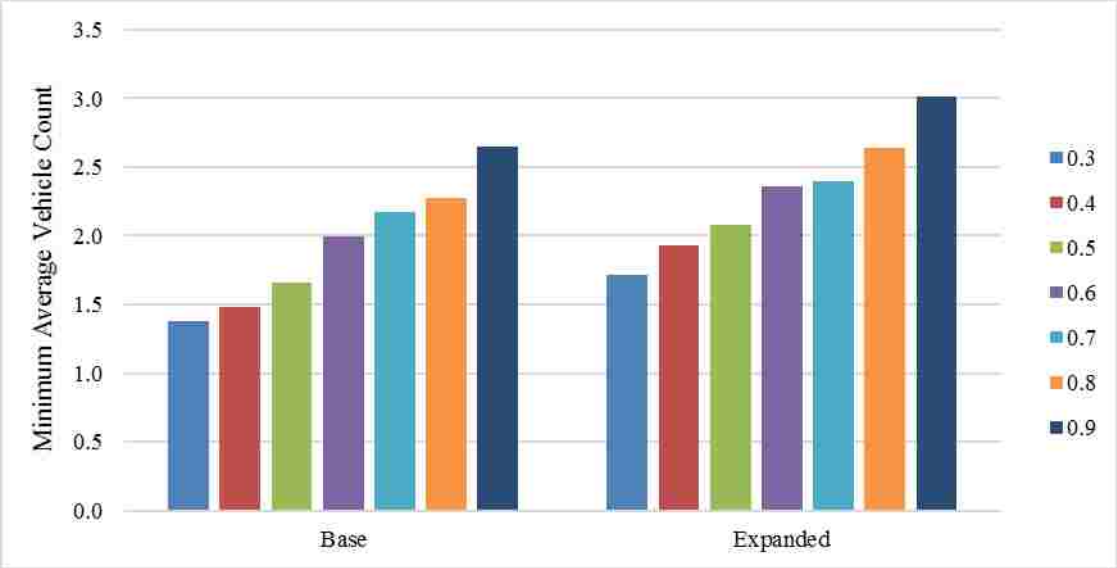

Figure 6-12: Minimum Average Vehicle Count for the Top 50% of ρ Bins (60-Second Observation Interval)

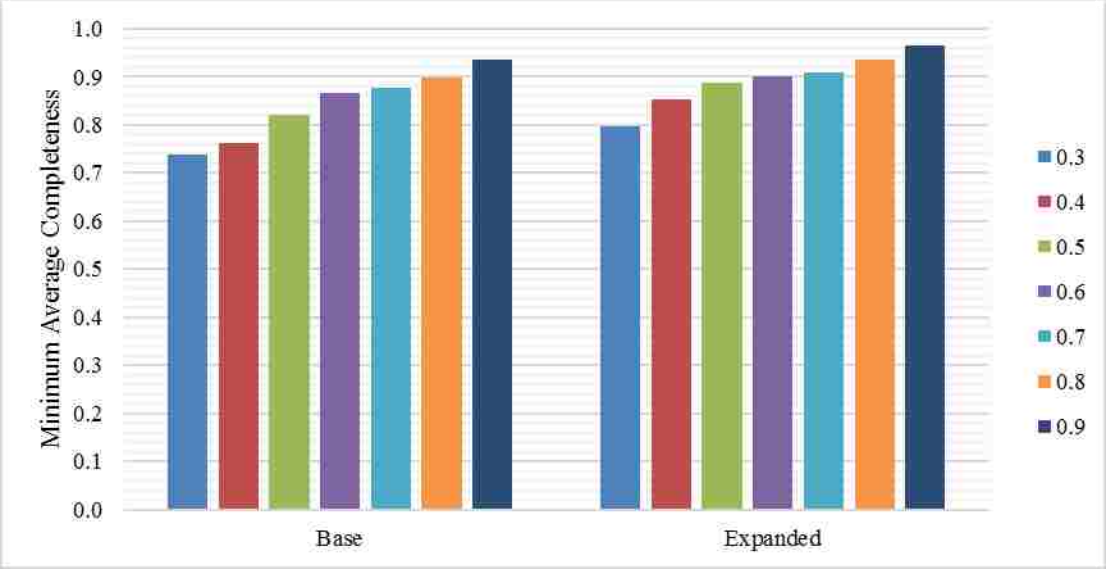

Figure 6-13: Minimum Completeness for Top 70% of ρ Bins (60-Second Observation Interval)

As a further investigation of the potential benefit of the expanded dataset, here the standard deviation of the observed speed is considered. Though the estimation procedure presented in this work ignores the impacts of GPS noise and other sources of error beyond the sampling approach and parameters, it should give a reasonable estimate of the relative contribution of the sampling parameters to speed estimation error. Referring to Figure 6-14, it is clear that the expanded dataset reduces the observed speed standard deviation only slightly. Though the scenario is hypothetical, this illustrates the fact that increasing the sample size will decrease the observed variance, especially when significant heterogeneity in vehicle speeds is present. Consider that, in this hypothetical scenario, the true speed variance is assumed to be constant in the estimation of observed speed variance for all traffic conditions. Because of this, the reduction in the observed standard deviation (associated with both the increasing ρ and higher penetration rate) is strictly attributable to the increases sample size.

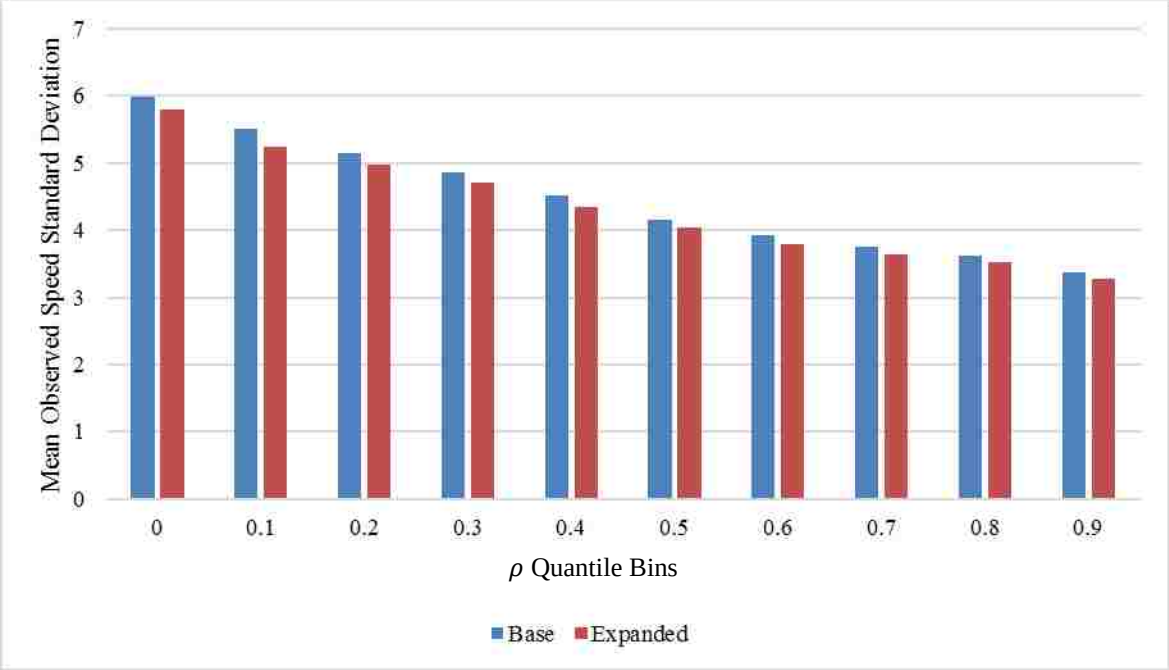

Figure 6-14: Mean Observed Speed Standard Deviation vs. ρ

6.2.5 *Discussion*

Generally speaking, probe vehicle data that has been used in practice comes from commercial sources, and so the decisions represented in this section are not typically made by the consumers of this data. In that sense, the method described here may seem to be of little use in practice. However, decisions regarding the aggregation method and parameters, the economics of data acquisition, and other factors which have a significant impact on the quality of the resulting data must be made by someone. If this topic is ignored, and these important decisions continue to be made internally by data resellers, then data will generally be consumed as is with little information on which to assess shifts in data quality and coverage, statistical anomalies, and unexpected analytical results. Thus, the methodology described in this section will become more useful as consumers become more involved in the decision making process. Many of the practical details of data collection such as GPS error, communications malfunction and failure, and other issues are ignored here. For this reason, the method is intended as an approximation which requires very little information about the traffic and sampling mechanism. If more precise results are required for a particular area, and more supporting information regarding traffic and sampling conditions is available, microscopic simulation may be more appropriate.

6.3 Bias Correction

Without knowing the true distribution of vehicle types and driving profiles on the roadway, addressing bias in probe vehicle data is a challenging issue. In the previous sections, we have focused on the case where the true conditions can be known (or estimated), and estimated the bias that can be expected in the observation process. However, as noted previously, the bias that will be represented in the final aggregated speed measures is largely determined by the method used to estimate aggregate speed from individual vehicle updates. In general, there is a higher likelihood

that no probe vehicles will traverse a segment during an observation interval during lower volume time periods. The method(s) used to compute aggregate speed measures will have little impact on this source of bias. However, these methods will have a significant impact on the bias that arises within the observed data due to the sampling mechanism.

Here the issue is approached by assuming that only some basic information about the relative fraction of vehicle subpopulations is known, and develop a set of methods to correct the sampling bias. The practicality of this method is largely driven by its computational complexity and the necessity of scaling to many thousands of roadway links. Because of this, this work focuses on methods that can be applied with minimal computational work, such as would be required to train and apply complex statistical models. Such an approach would be especially important if the methods are to be applied in near real-time.

6.3.1 *Methodology*

If it can be assumed that all vehicle subpopulations (with respect to the mean speed distribution) are represented to some extent on a road segment during a given time interval, it should be possible to estimate the unobserved true traffic and sampling parameters from the observed data using some form of weighting scheme. Developing this weighting scheme is first step in this section. If this cannot be assumed, then other methods will be required to insure that the observed data is representative of the true traffic state. For this purpose, subpopulation imputation or hole-filling scheme is developed. Finally, an alternative approach to speed estimation is proposed which reduces the need for statistical weighting and imputation, and generally provides more accurate results than the point speed estimation methods described previously.

6.3.1.1 Inverse Probability Weighting

Restricting attention to a single mean vehicle speed distribution, the set of vehicles that appears on a road segment during an observation interval will correspond to a random sample from this speed distribution. Assuming that this collection of vehicles is a reasonably representative sample of the probe vehicle population consisting of *n* probe vehicles, the result will be a vector $V = \{v_1, v_2, \ldots v_n\}$ of vehicle speeds. Each of these vehicles and corresponding speeds has some probability of appearing on a road segment, as well as some probability of being observed given it appears on the road segment, which is a function of the speed itself (or travel time) and the sampling frequency of the associated probe vehicle. Thus, though the vehicles appearing on the segment may be a representative sample of the probe vehicle population, the *observed* sample may disproportionately favor shorter sampling intervals and slower vehicles relative to the true distribution of these parameters. A straight-forward correction can be made to the samples to reduce the sampling bias. The approach is adapted from survey and missing data literature, in which survey responses are weighted according to the inverse probability of response (Gary 2007) to improve estimation of population parameters. In the case of survey data, the probability of response or "response propensity" is often estimated using a regression model, trained on both responding and non-responding individuals with sociodemographic or other variables as predictors.

In the case of probe vehicle data, the probability of response can be estimated using Equation 4.27, and the samples can be weighted by the inverse of $P(nspv > 0|TT, sr)$. This approach can be understood intuitively, knowing that the probability of a given realization of a random variable (e.g. observed mean speed or sampling interval) is defined by the product of a) the value occurring and b) the probability of the value being observed, given it has occurred. For example, there is some probability of a probe vehicle having a sampling interval of 30 seconds

appearing on a road segment during an observation interval. The probability of that vehicle being observed is a function of travel time and the sampling interval itself. By weighting the observed samples by the inverse probability of being observed, the estimated weighted mean will be a better approximation of the true expected value. The weighted speed calculation is completed as shown below in Equation 6.1, where $\tilde{\mu}$ is the estimated mean speed for a single observation interval.

$$\tilde{\mu} = \left(\sum_j P\left(nspv_j > 0 \middle| TT(v_j), sr_j\right)\right) \sum_i \frac{v_i}{P(nspv_i > 0 | TT(v_i), sr_i)} \quad (6.1)$$

6.3.1.2 Population Weighting

If the probe vehicle population can be considered a representative sample of the overall vehicle population, then this weighting step will give reasonably good results if the penetration rate is sufficiently high. In the case of multiple subpopulations, each with different penetration rates, then the probe vehicle population does not constitute a representative sample of the overall vehicle population. In this case, some information about the true overall vehicle population can be used to further adjust the subpopulation speeds. This is analogous to population weighting in survey literature. To do this, it is necessary to define vehicle classes such that a) the class of a vehicle can be identified through the GPS updates they provide and b) the driving behavior in terms of average speed relative to prevailing traffic speed within a subpopulation is expected to be relatively homogeneous. That is, the average speed on a link is a single distribution for each subpopulation and time period. For example, if the available probe vehicle data comes from a combination of taxis and on-board navigation from a certain brand of vehicles, each of the two providers would constitute a separate subpopulation.

However, it may be the case that a vehicle population is present which, through some combination of low penetration rate and/or infrequent sampling, is often missing completely during an observation interval. Furthermore, there are likely to be observation intervals during

which no vehicles are observed, and these are statistically more likely to be faster moving, less congested time periods compared to those for which data is available. To address these challenges, some form of imputation scheme is needed to insure that all subpopulations can be represented in the final observed speed. Here, initial estimates are made using linear interpolation, and a Gaussian smoothing kernel is applied to the interpolated values. A state space time series model would seem to be more appropriate, and in many real-world cases (for example, a real-time application requiring both imputation and prediction) such a model would be preferable. However, at low penetration rates the underlying data is sparse enough to make training such a model problematic. An example of the proposed imputation scheme is shown in Figure 6-15, where the variance of the Gaussian smoothing kernel is set to 16. The final approach to bias correction in mixed probe vehicle penetration rates can be described as follows:

1. For each observation interval, compute speed for each subpopulation (possibly using the inverse probability weighting method described above or the distance weighted pairwise vehicle speed as described in Section 6.3.1.3.
2. Impute missing observation intervals for each subpopulation using some method, in this case linear interpolation with Gaussian smoothing described above.
3. Compute the mean speed for each observation interval, weighting each subpopulation according to the known or estimated population fraction represented by that subpopulation.

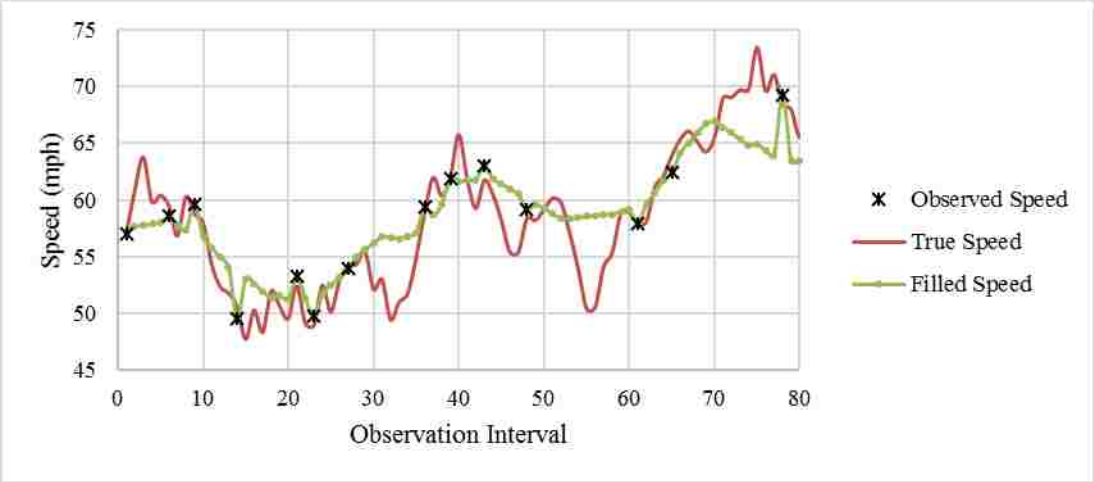

Figure 6-15: Illustration of the Interpolation + Gaussian Smoothing Imputation Scheme

6.3.1.3 Distance Weighted Harmonic Mean Speed

The final methodological element to be discussed here is the method used to combine vehicle state updates into a mean speed measurement for each observation interval and location of interest. Until now, the drawbacks of different methods of aggregating pointwise vehicle speeds in various ways have been discussed at length. Now, a discussion of distance-weighted pairwise speed is offered as a generally more favorable alternative. This is not a new concept, in fact the introductory sections of this document discussed a variety of methods that have been used in the past to estimate link-level speeds from vehicle GPS measurements. However, this method is introduced here to compare it with the pointwise methods and point out the strengths and weaknesses of both approaches. It is likely that some probabilistic model for speed estimation could be identified that provides similar or superior results. However, it must be assumed that a suitable method must be as accurate as possible while remaining massively scalable. The distance-weighted mean speed estimation approach prosed here fits both of these requirements and as will be shown, performs better in most cases that the pointwise methods described and simulation.

(Edie 1965) provides an accurate and interpretable definition of the mean speed on a road segment as the aggregate travel distance divided by the aggregate time spent on the segment by all vehicles during an observation period. This definition is the basis for the distance weighted harmonic mean speed calculated here, except of course that only probe vehicles are considered in the calculation. The equivalence between this definition and the distance weighted harmonic mean can be seen in Equation 6.2, where the distance factors out of the numerator and the expression simplifies to the sum of distances traveled divided by the sum of travel time. Despite this equivalence, the distance weighted mean interpretation is useful, because it indicates that distance rather than time is the basis for weighting individual vehicle speeds. In this way, we move away from the implicit time weighting that is the basis for the point sampling and vehicle mean methods described previously.

$$\widetilde{\mu_t} = \left[\frac{\sum_{i=1}^{n} dist_{i,t}\frac{t_i}{dist_{i,t}}}{\sum_{j=1}^{n} dist_{j,t}}\right]^{-1} = \left[\frac{\sum_{i=1}^{n} t_i}{\sum_{j=1}^{n} dist_{j,t}}\right]^{-1} = \frac{\sum_{j=1}^{n} dist_{j,t}}{\sum_{i=1}^{n} t_i} \qquad (6.2)$$

Where:

$\widetilde{\mu_t}$ = mean speed estimate for observation interval *t*

$dist_{i,t}$ = distance traveled on the link by vehicle *i* during observation interval *t*

v_i = mean speed of vehicle *i* during observation interval *t*

Generally speaking, pairwise speed estimation based on travel distance and time is more accurate and precise than pointwise GPS speed. This fact and any related analysis are ignored in this work. Instead, the focus is on the relationship between sampling parameters, traffic state, sampling bias and observation error unrelated to GPS accuracy.

To calculate the distance weighted harmonic mean speed, the first step is to estimate the approximate trajectory of a point. This done when a location update arrives by first computing the mean speed between the points, and then allocating this mean speed to current and past observation intervals and road links based on the travel time / speed relationship. This is illustrated in Figure 6-16, which shows two subsequent location updates occurring on two separate road segments. If the observation interval length is 60 seconds, then each update occurs in a separate observation interval as well, one starting at 13:04:00 and the other starting at 13:05:00. The calculated mean travel speed is allocated to the final 2360 feet of road link A, and to the first 800 feet of road link B. The time at which the vehicle left Link A is unknown, but can be estimated using the calculated mean speed. The result is shown in Table 6-5, where the distance is allocated between road segments and observation intervals. First, the transfer time between the start, end, and all intermediate links is estimated from the mean speed. Then, the distance traversed on each link is split based on the break points of the observation intervals.

Figure 6-16: Illustration of Distance Speed Calculation

Table 6-5: Distance Allocation between all Road Links and Observation Intervals

Observation Interval	1	2	
Road Link	A	A	B
Time (seconds)	39	5.85	15.15
Distance (feet)	2059	309	800
Speed (mph)	36		

It may be clear to the reader that this method will introduce some error when the speeds on adjacent road links differ significantly. For example, if the speed on road link B in Figure 6-16 is significantly lower than on road link A, the speed value assigned to road link A will to some extent reflect the slower speed traveled on Link B between updates. In effect, this smooths the speeds between adjacent links. However, this approach will reduce the rate of missing data (because an update need not fall on a road link in order for the link to be "observed") and generally improve accuracy in all but the most extreme cases.

There are, of course, other methods that could be used to combine vehicle location updates into aggregate link speeds. For example, a time weighting scheme could be used rather than distance weighting. However, the objective is to characterize as accurately as possible the mean speed over the length of a road segment from point pairs, most of which do not span the full extent of the segment. A time weighting scheme would increase the weight for slower moving vehicles, even for a point pair that represents the exact same section of a road as a faster moving vehicle. In fact, a time weighting scheme is very similar to the point-wise mean speed approach, which essentially weights vehicles' speed according to how long they spend on the segment, assuming identical sampling frequencies.

6.3.2 *Test Scenarios*

The primary mechanisms that contribute to bias and inaccuracy will be driven by the sampling parameters that are present in the probe vehicle population. Thus, to illustrate the utility of the proposed bias correction method, three hypothetical scenarios are devised, each of which associated with particular mechanism contributing to bias. Each scenario has been simulated using PTV VISSIM, and the results presented and discussed. The roadway geometry and simulation

parameters are identical to the model used in the Validation section, only the traffic, vehicle, and sampling parameters are changed.

6.3.2.1 Scenario 1

In the first scenario, the probe vehicle population is a representative of the true traffic population. Thus, the primary source of bias is the fact that slower moving and more frequently sampled vehicles are disproportionately represented. Under this scenario, it will be possible to simply use inverse probability weighting for each vehicle, with probability weights based on the likelihood of that vehicle being observed. To assess this scenario, a simulation model was built in PTV VISSIM with three distinct vehicle subpopulations and a combined total vehicle volume of 3500 vehicles / hour. All vehicle subpopulations are equal in volume and penetration rate, they only differ in their desired speed distribution and sampling interval distributions. The desired speed and sampling interval distributions are shown in Figure 6-17 and Figure 6-18, respectively. The geometry of the simulated roadway is identical to that shown in the validation section (Section Chapter 5:).

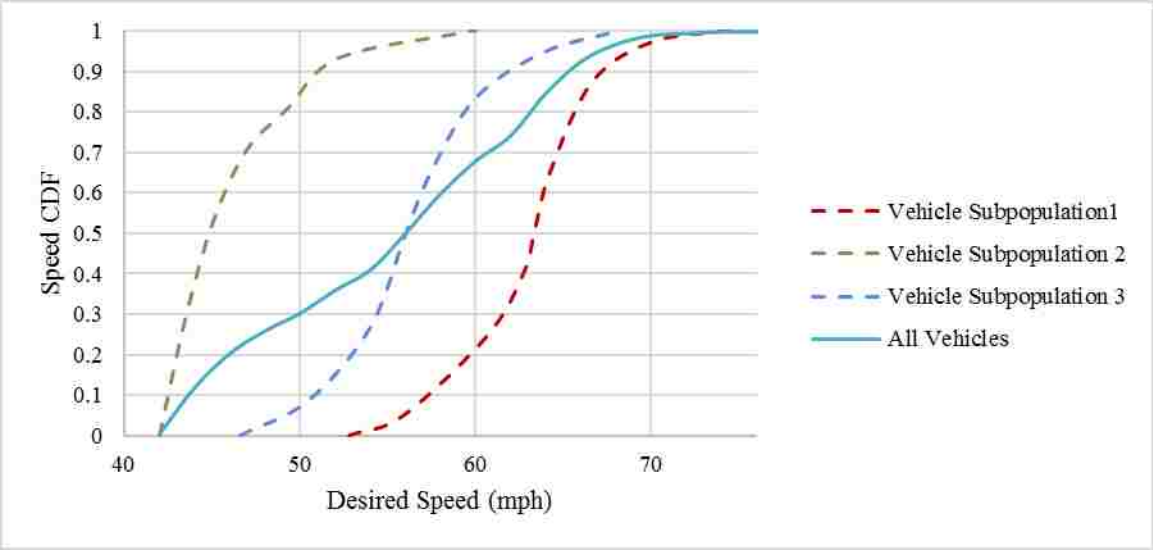

Figure 6-17: Desired Speed CDFs for Three Vehicle Subpopulations

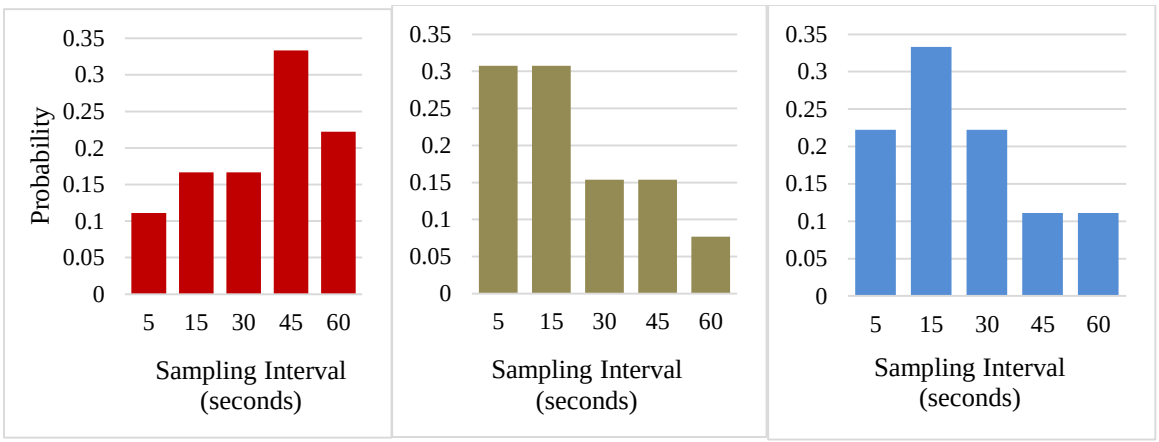

Figure 6-18: Sampling Interval Distribution for Subpopulations 1-3 (from left to right)

It can be expected in this scenario that, given the penetration rates are equal for all subpopulations, the bias present will be relatively small. Furthermore, as noted previously, the inverse probability weighting methods will be increasingly beneficial as the penetration rate of all vehicles increases. Thus, a range of penetration rates are investigated to show this relationship.

Figure 6-19 and Figure 6-20 show the mean absolute percent error and mean bias for all four speed estimation methods: Vehicle mean speed, Inverse probability weighted vehicle mean speed, distance weighted pairwise mean speed, and sample mean speed. The inverse probability weighting scheme reduces bias as expected, but does so at a small penalty in mean squared error. It is important to note that the expected observation bias in this scenario is small, because the probe vehicle population is representative of the full vehicle population. A larger disparity in the sampling rates, a highly heterogeneous vehicle/driver population, or differences in penetration rates between different subpopulations could all contribute to more dramatic bias. In such cases, the bias would be a larger contributor to mean squared error, which would favor the inverse probability weighting scheme over the vehicle mean speed in terms of mean squared error. That said, the distance weighted pairwise speed method outperforms all other methods at every penetration rate, and reduces the bias to near zero even at relatively low penetration rates.

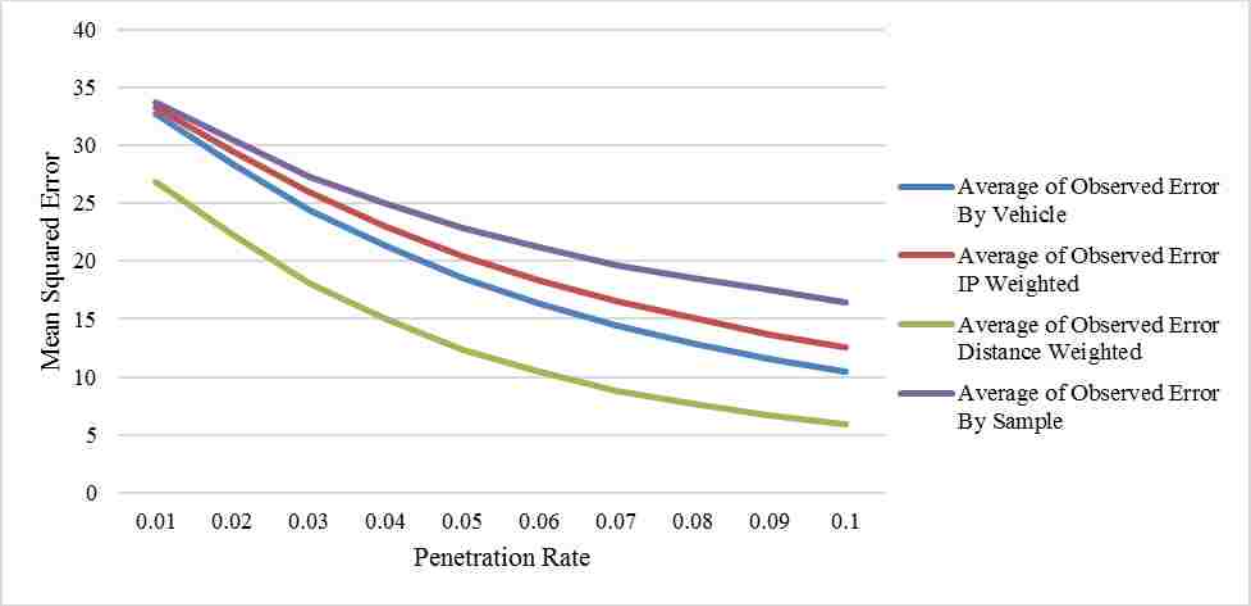

Figure 6-19: Mean Squared Error vs. Penetration Rate for all Speed Estimation Methods

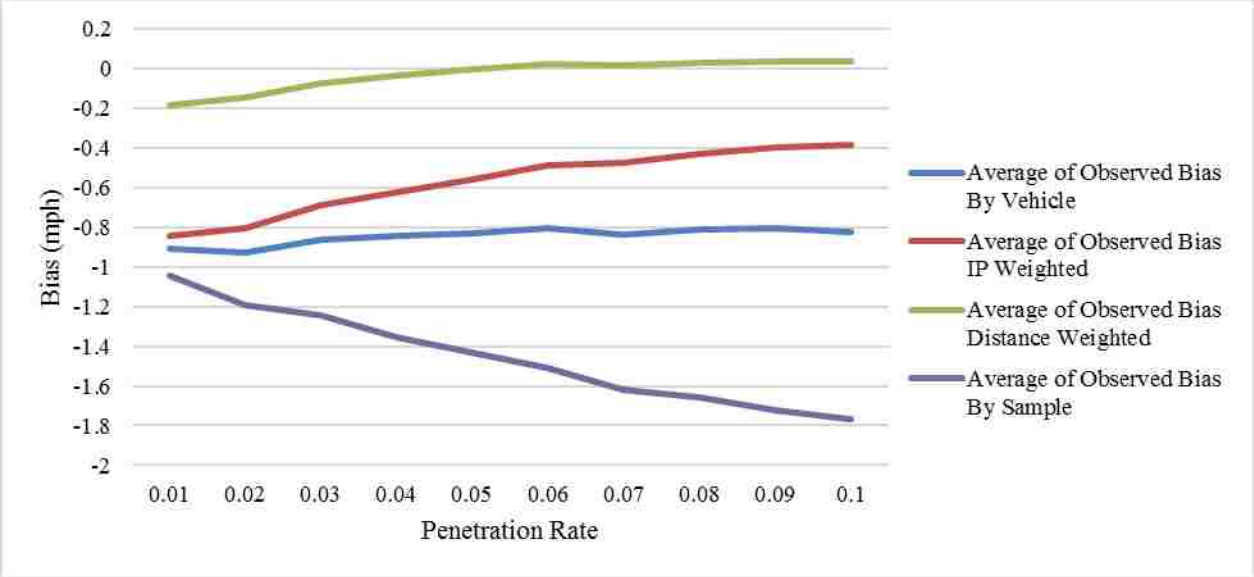

Figure 6-20: Bias vs. Penetration Rate for All Speed Estimation Methods

Figure 6-21 through Figure 6-26 show the Mean Squared Error and Bias for penetration rates 0.01, 0.05, and 0.1. It is again clear from these plots that the inverse probability weighting scheme systematically reduces bias, but increases mean squared error somewhat for all but the lowest penetration rates. The distance weighted speed method performs the best overall, reducing

bias and error over all penetration rates. Note however that the distance weighted speed method produces a positive bias on road segment 7 and a small negative bias on the two adjacent segments. This is because segment 7 has a slowdown section (as noted previously), and this speed estimation method tends to conflate adjacent segments to some degree as described in Section 6.3.1.3. That is, a certain portion of travel that took place on segment 6 was assigned to segment 7, and vice versa. Because vehicles travel faster on segment 6, this produces a slight negative bias on segment 6 and positive bias on segment 7.

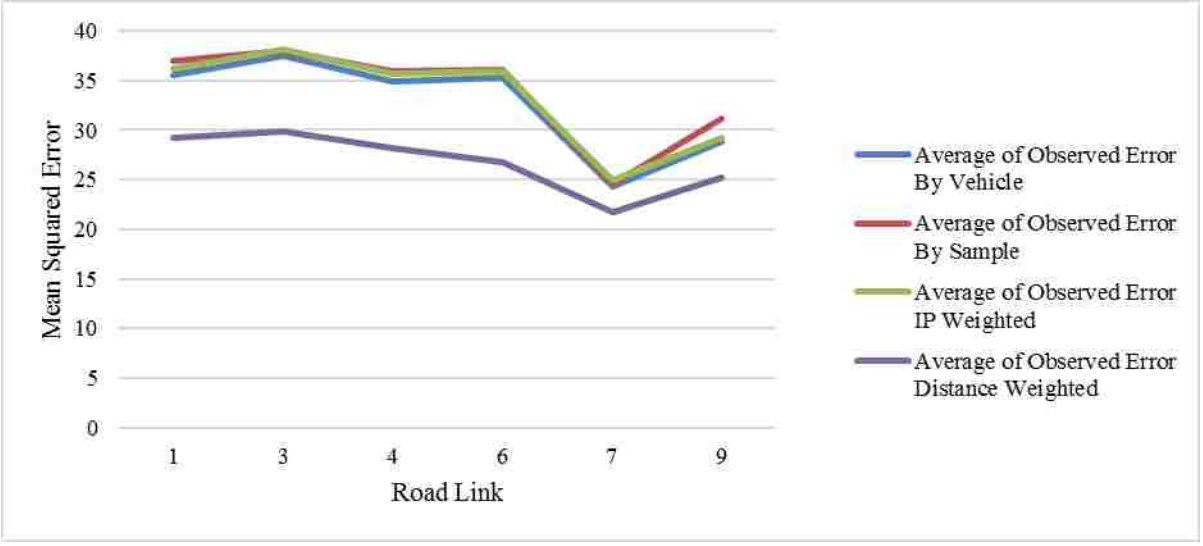

Figure 6-21: Observed Mean Squared Error For Penetration Rate of 0.01

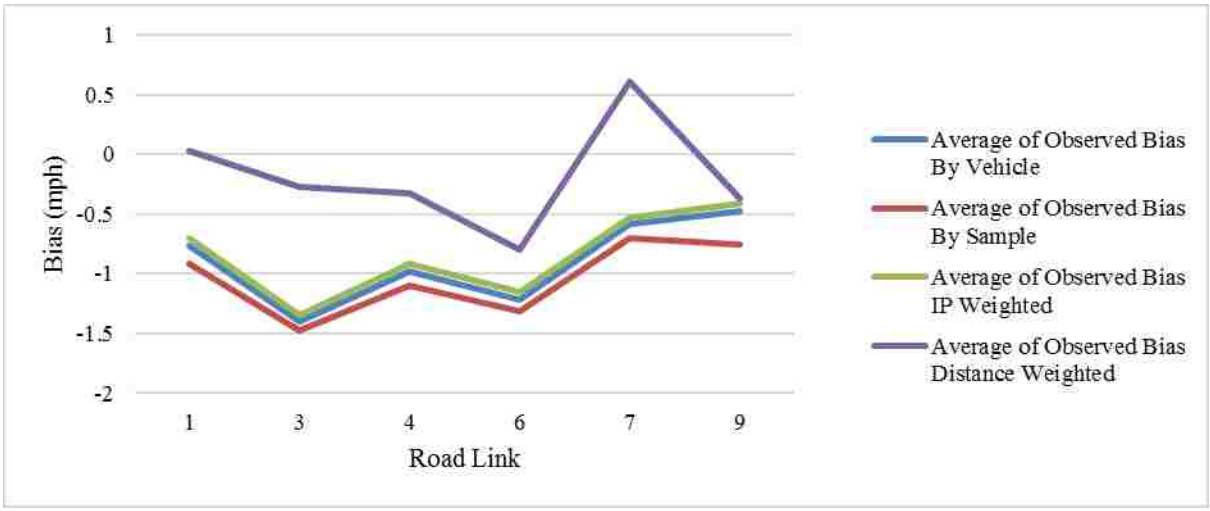

Figure 6-22: Observed Bias for Penetration Rate of 0.01

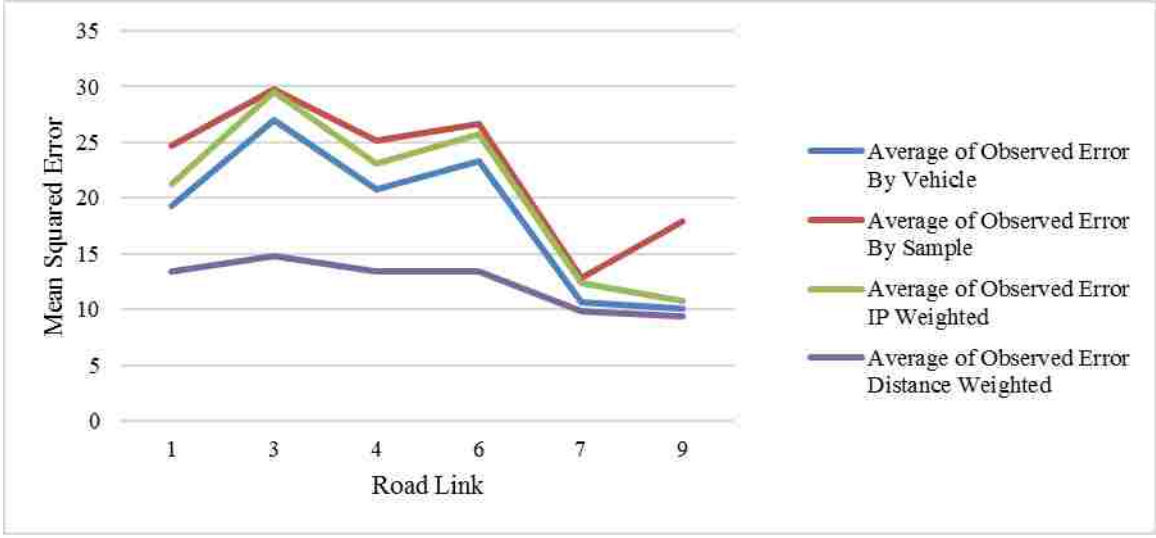

Figure 6-23: Observed Mean Squared Error for Penetration Rate of 0.05

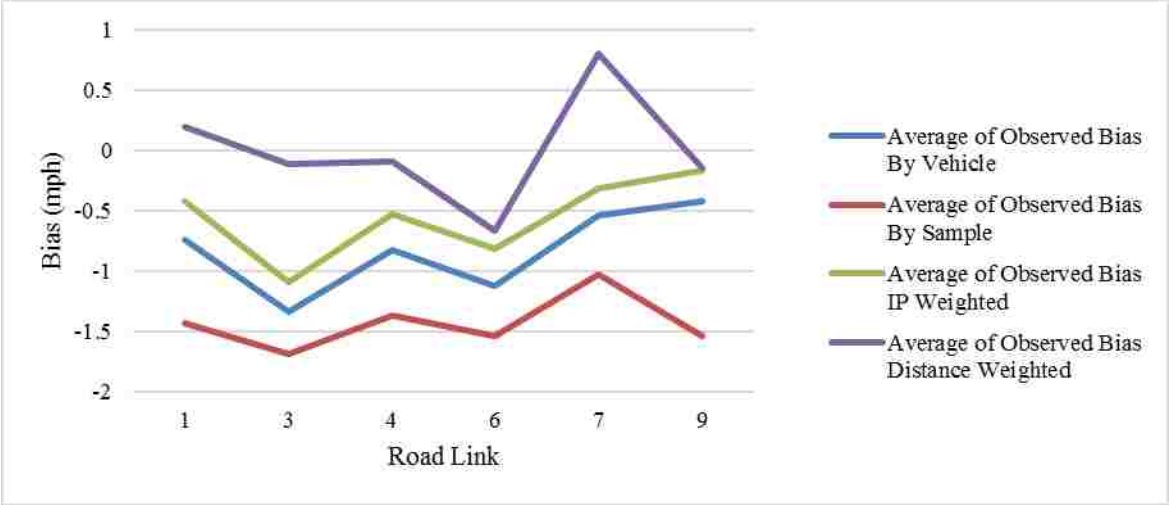

Figure 6-24: Observed Bias for Penetration Rate of 0.05

As the penetration rate increases, the accuracy of the point-based methods improves dramatically. Still, the bias of the sample average method continues to degrade, and it consistently performs the worst overall. It is also clear that the difference between the three point-based methods increases as the penetration rate increases. To explain, consider that all three methods will be equivalent when only a single vehicle is observed during an observation interval. As the number of observed vehicles per observation interval increases, the difference between mean vehicle speed, mean sample speed, and weighted mean vehicle speed will increase.

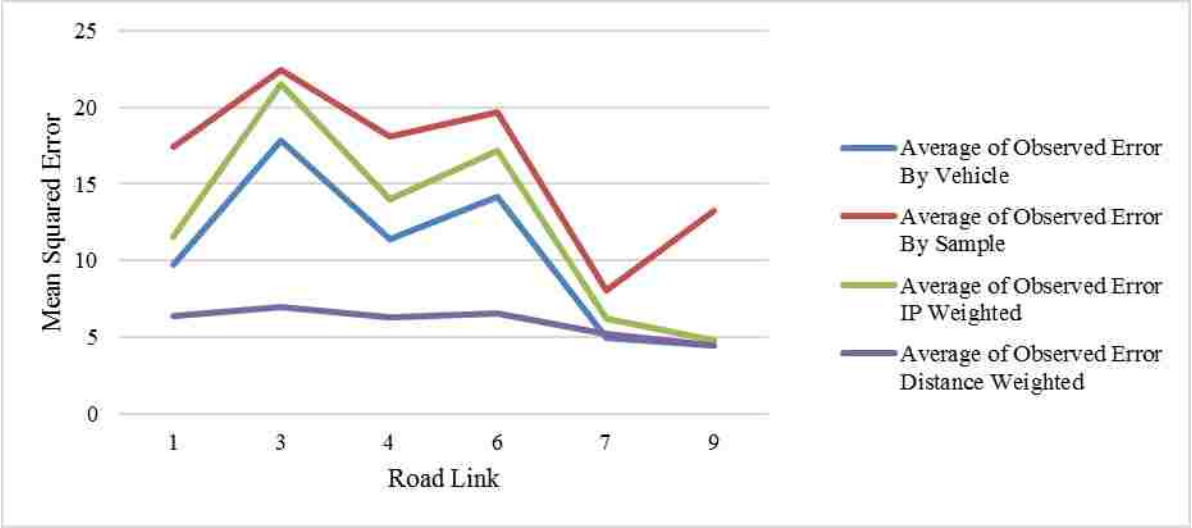

Figure 6-25: Observed Mean Squared Error for Penetration Rate of 0.1

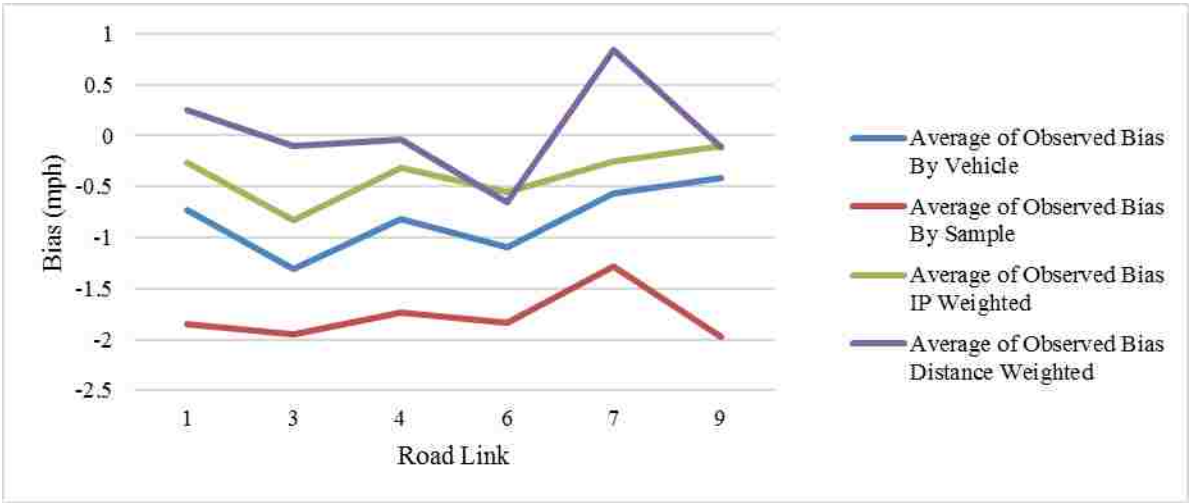

Figure 6-26: Observed Bias for Penetration Rate of 0.1

6.3.2.2 Scenario 2

In this scenario, the penetration rates of the contributing vehicle subpopulations is allowed to vary, such that the probe population is no longer a representative sample of the overall vehicle population. As a result, in addition to the adjustment method described in the previous section, a second weighting step is applied to address the resulting bias. This method assumes that some

estimate of the relative volumes of each subpopulation can be estimated, and applies this information to adjust the overall mean speed to reflect the relative share of each subpopulation present in the traffic stream. As noted previous, certain subpopulations may not be represented at all in many observation intervals, and there may be observation intervals which are missing completely. Thus, the previously described imputation scheme is applied to insure that all subpopulations are appropriately represented in all observation intervals. This addresses two forms of bias that are observed in probe vehicle data, 1) that the observed vehicle population may differ significantly from the true vehicle population and 2) that slower moving, more congested time periods are less likely to be missing completely, all else being equal.

The scenario used to test this approach is identical to the previous scenario, except that the penetration rate is allowed to vary between subpopulations. This is done by defining an overall penetration rate over all probe vehicles and a relative share for each subpopulation. This way, the effectiveness of the proposed approach can be assessed over a range of penetration rates for a fixed mix of different subpopulations. An example is given here in Table 6-6, for which the volume of all vehicles is equal between the three subpopulations, but the penetration rates of contributing vehicles changes. This table should be interpreted as showing that just over 71% of all vehicles comes from subpopulation 2, which is equivalent to a penetration rate of approximately 4.3% when the overall penetration rate is 2.0%. Note that the total penetration rate is the simple mean of penetration rates in Table 6-6**Error! Reference source not found.**, but this is only true because t he total traffic volumes for the three subpopulations are equal. Otherwise, the total penetration rate would be a weighted average of penetration rates, weighted by the respective total traffic volumes of each subpopulation.

Table 6-6: Relative Share and Example Penetration Rates for Scenario 2

	Relative Share	Penetration Rate
Subpopulation 1	0.143	0.0086
Subpopulation 2	0.714	0.0429
Subpopulation 3	0.143	0.0086
Total	1.0	0.02

The distance weighted pairwise mean speed estimation produced the best results in the first scenario, so the focus here is exclusively on comparing the adjusted vs. unadjusted speed computed using this method. All the estimated speed values in the following results are based on the distance weighted pairwise mean speed. Figure 6-27 shows part of the simulation results for the penetration rates give in Table 6-6, the sampling interval distributions given in Figure 6-18, and an observation interval of 60 seconds. Note that the adjusted results are generally less noisy and closer to the true speed values. The adjusted values are also complete for all observation intervals, which is not the case for the raw data. Furthermore, in the raw values there is a clear bias toward lower speeds, which again does not appear to be present in the adjusted values.

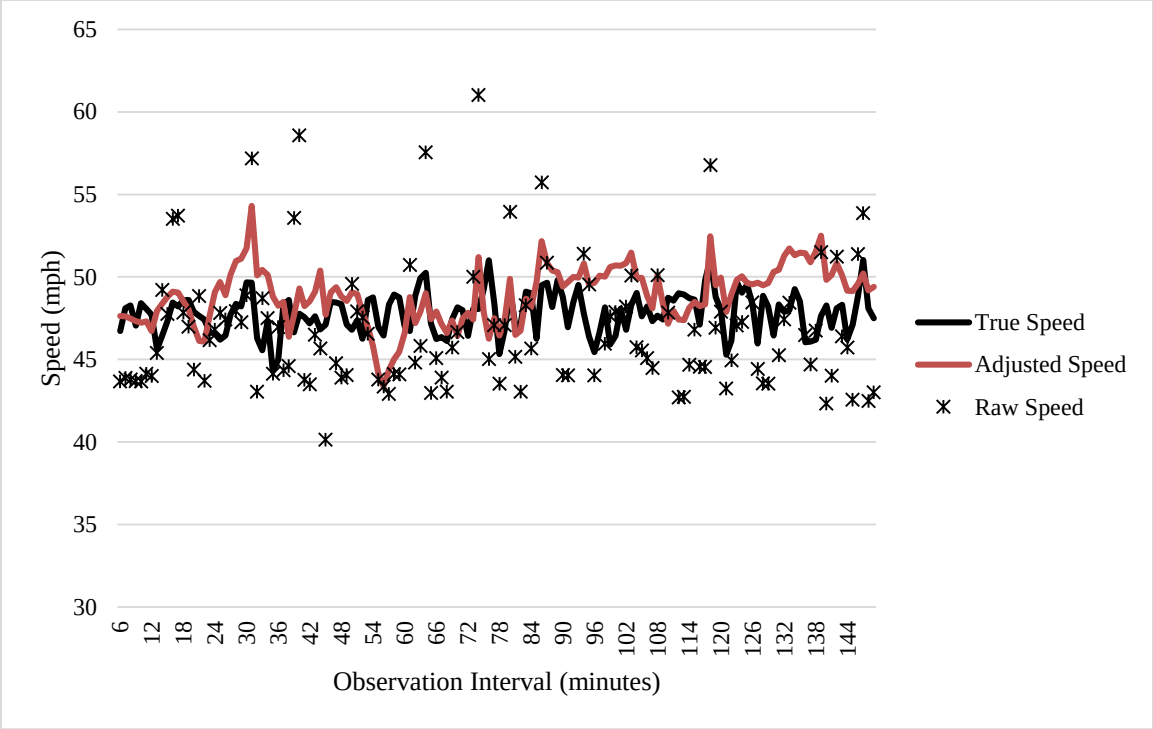

Figure 6-27: Example of Adjusted Speed Results (penetration rate = 0.02)

Figure 6-28 and Figure 6-29 show the mean absolute percent error for the raw and adjusted data, respectively, over three different penetration rates (0.01, 0.05, and 0.10). As previously noted, the relative share of each vehicle subpopulation is help constant, only the overall penetration rate is varied. It is clear from these figures that the adjusted data provides more accurate results over all road segments and penetration rates, but the benefit is less pronounced for the higher penetration rates. The reason for this is that the bias attributable to vehicles being over-represented due to being slow or more frequently sampled is not significant here, as the distance-weighted pairwise mean speed calculation method is used. Thus, the two primary causes of error are observation noise due to low sample size and missing vehicle subpopulations due to low penetration rates. Thus, as the overall penetration rate increases, it is more likely that all subpopulations will be present during a given observation interval, thereby reducing the bias attributable to these causes.

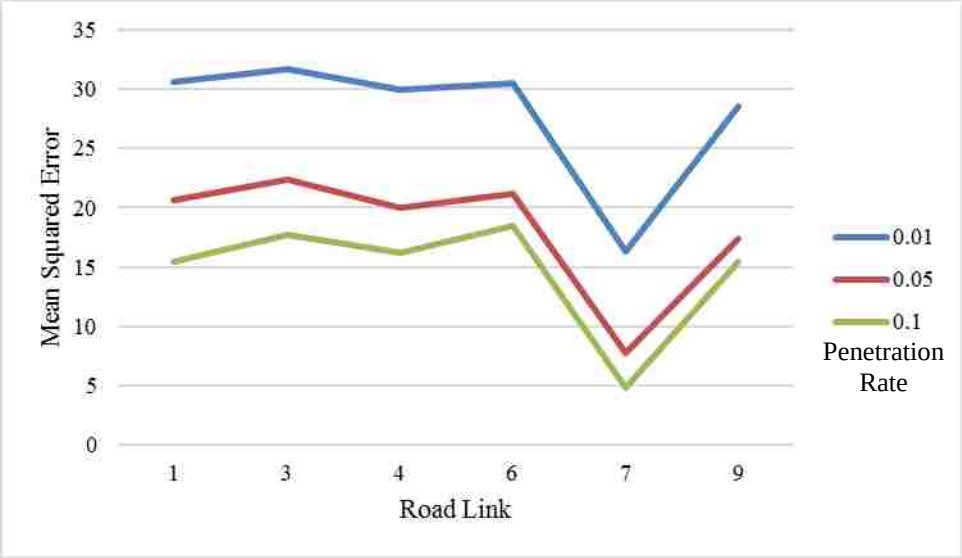

Figure 6-28: Mean Squared Error for Raw Observed Data

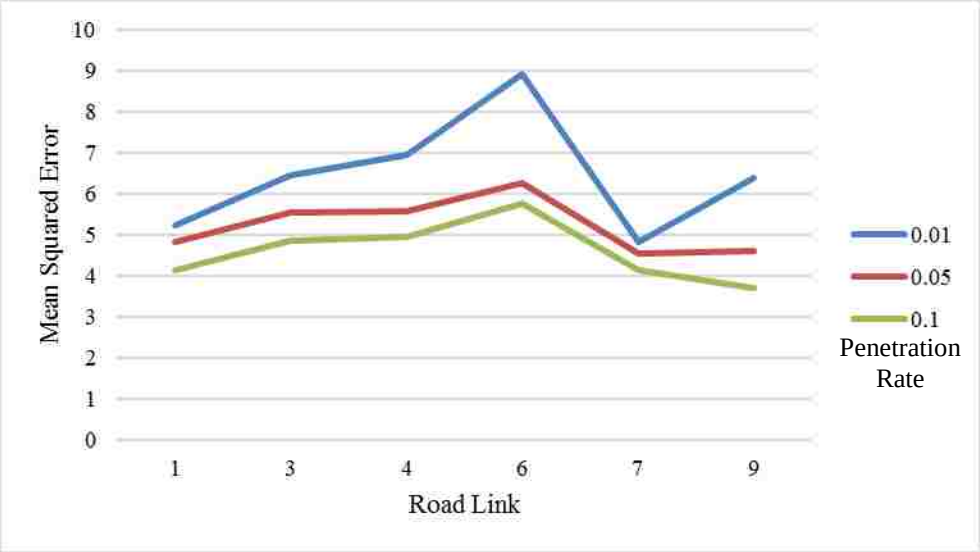

Figure 6-29: Mean Squared Error for Adjusted Data

The traffic conditions represented here are uncongested and relatively steady over time. One can expect that in more variable traffic conditions, the more congested time periods will be the most complete, which will contribute to even greater bias in the raw data. To show how the relationship between the true, raw observed, and adjusted observed speed compare under changing traffic conditions, another round of simulation was completed. In this case, the traffic volume was

increased from 3500 vehicles / hour to 8200 vehicles / hour at around 60 minutes into the simulation. This is equivalent 2050 vehicles / lane / hour, which is near the capacity of the simulated roadway.

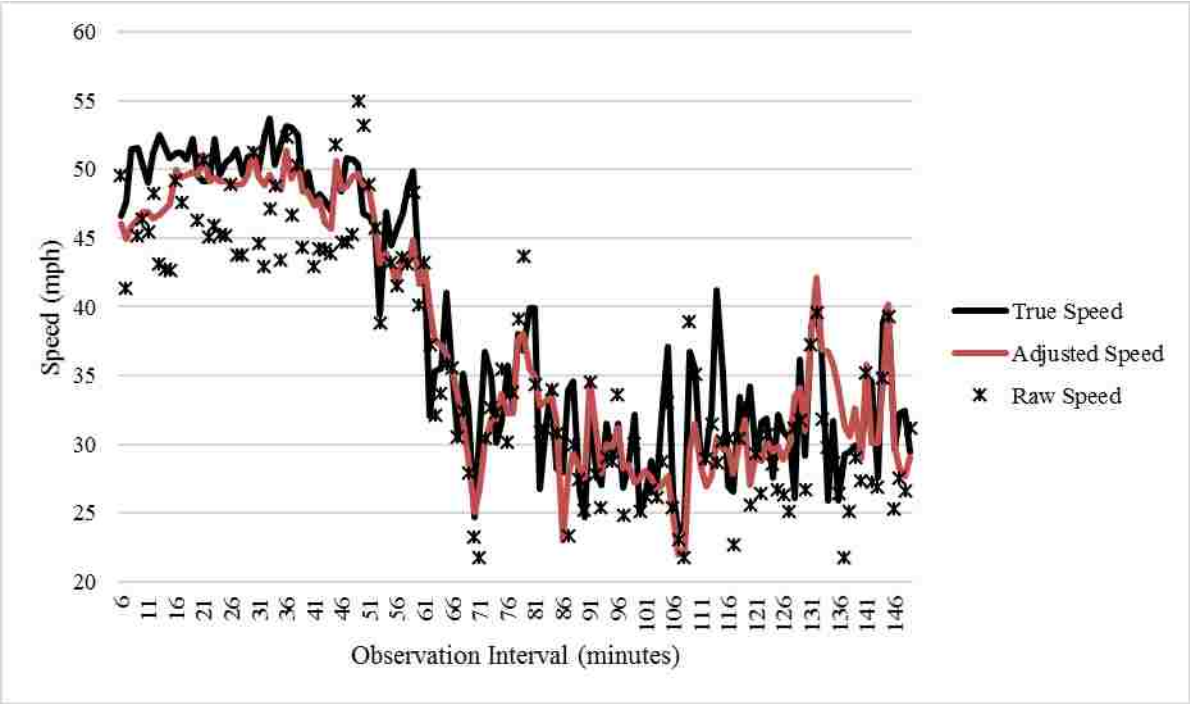

Figure 6-30: Example Adjusted Speed Results for Congested Scenario (penetration rate = 0.02)

Figure 6-31 through Figure 6-36 show the mean squared speed estimation error and bias for dynamic volume case described above. Because the difference between the raw and adjusted results is not as substantial as in the previous case, they are plotted together to allow better comparison. Though the atdvantages of the imputation and adjustment method are not as substantial in this case, the method does provide significant and consistent benefit.

Figure 6-31: Mean Squared Error for Penetration Rate of 0.01

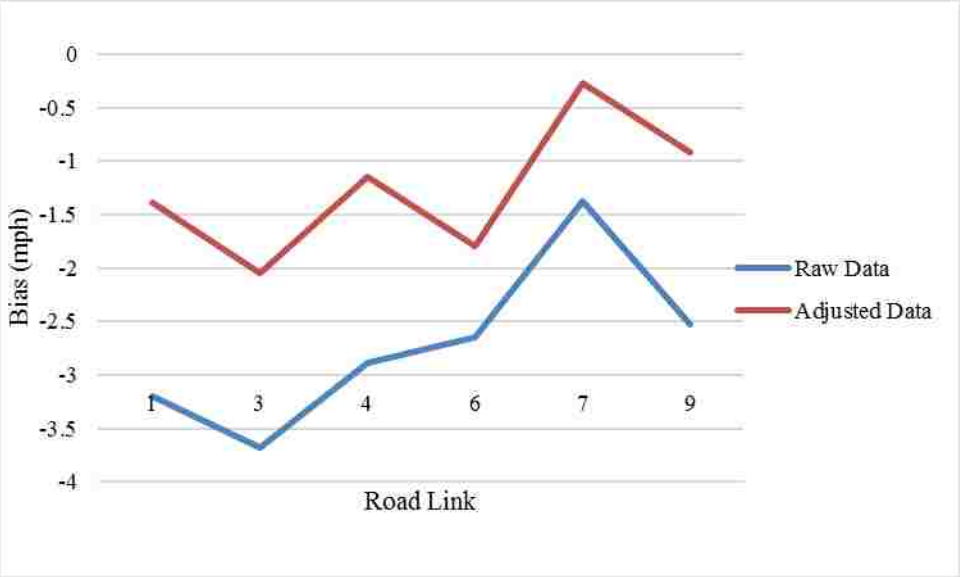

Figure 6-32: Bias for Penetration Rate of 0.01

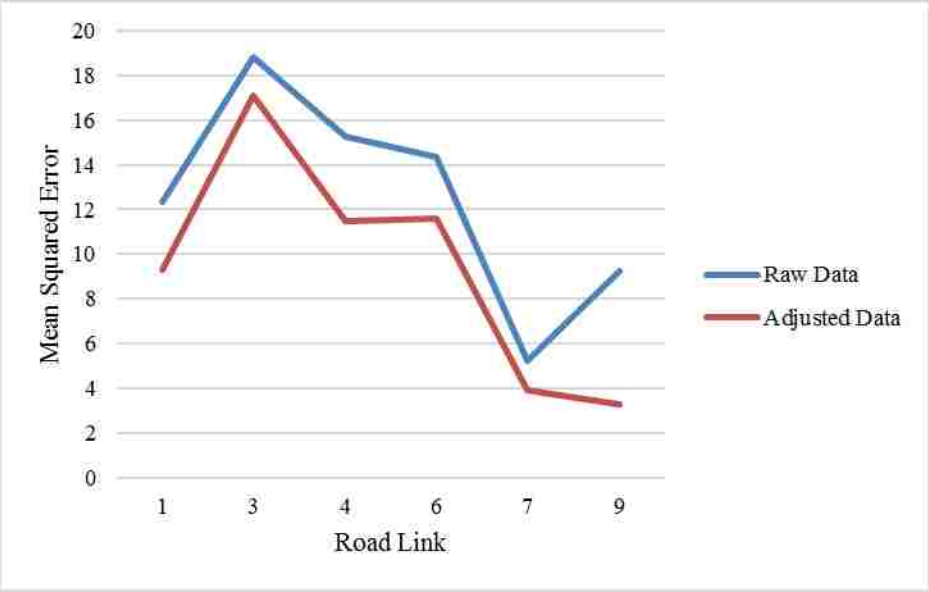

Figure 6-33: Mean Squared Error for Penetration Rate of 0.05

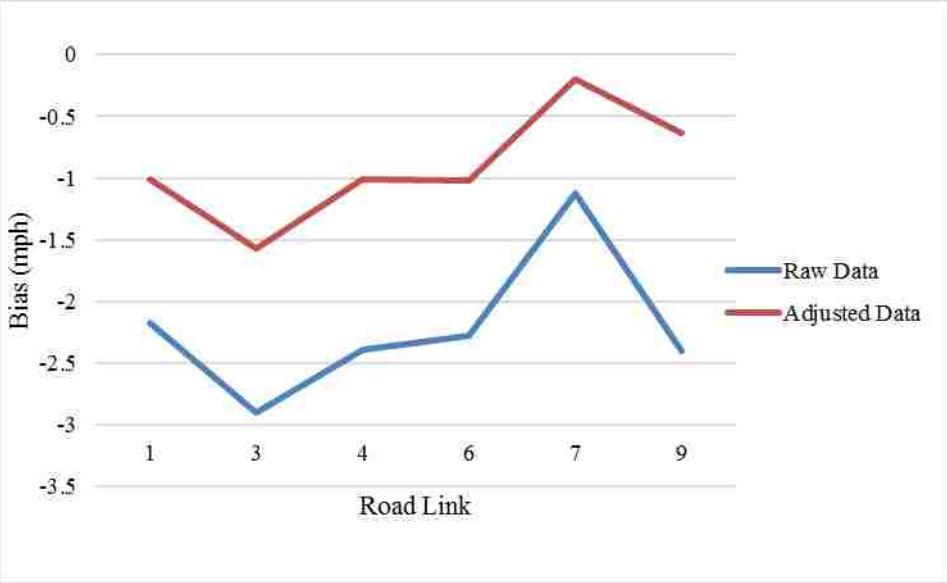

Figure 6-34: Bias for Penetration Rate of 0.05

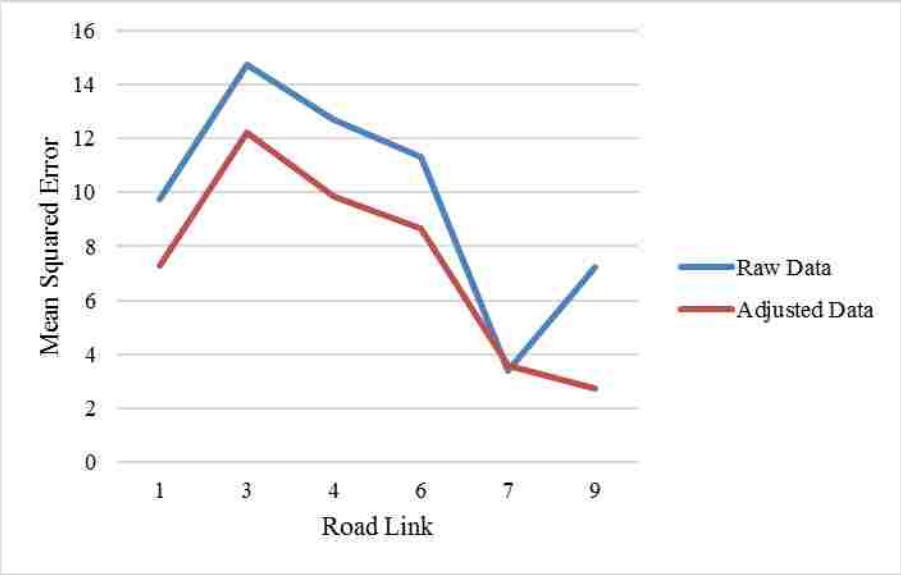

Figure 6-35: Mean Squared Error for Penetration Rate of 0.1

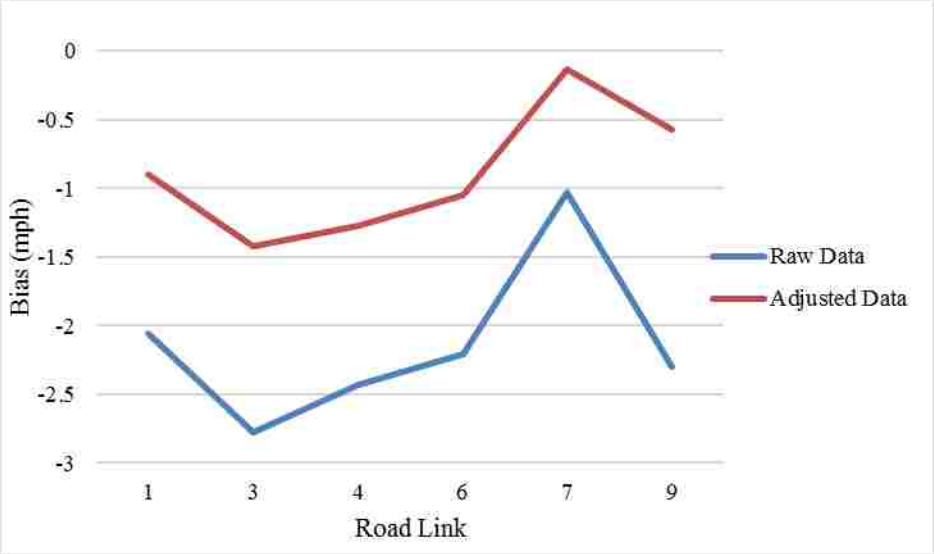

Figure 6-36: Bias for Penetration Rate of 0.1

Figure 6-37 below shows the mean squared error over all tested penetration rates. Note that, while the benefit of the speed adjustment does decline with increasing penetration rate, again the decline is not as pronounced as in the constant volume, uncongested scenario. This can be explained by the fact that the majority of these simulation time was spent at more than twice the volume as the constant volume scenario, when the majority of the unadjusted data was complete even at the lower

penetration rates. Furthermore, there is some penalty to the interpolation / smoothing operation at higher volume rates. This is because, as congestion increases, there is less difference between the vehicle population in terms of mean speed, which reduces the benefit of filling in missing subpopulations In addition to filling in missing values, the imputation methodology smooths over some of the fluctuations in traffic speed, which may further reduce the utility of the proposed method.

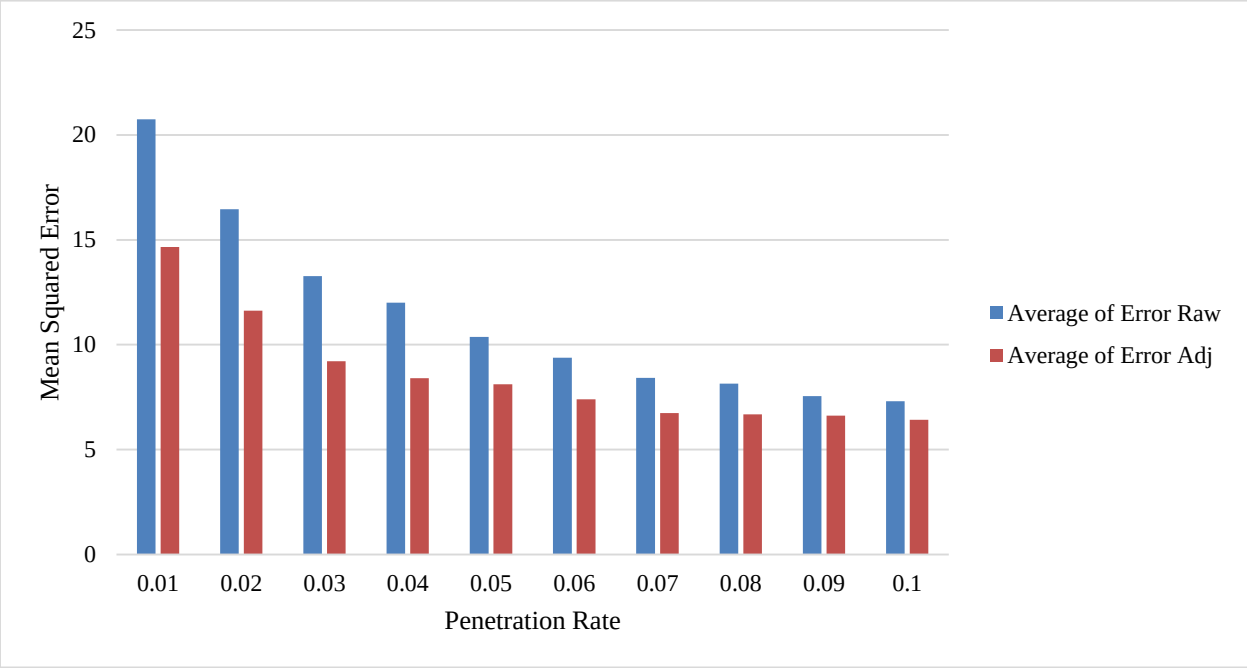

Figure 6-37: Mean Squared Error for the Raw and Adjusted Speed Values

6.3.3 *Discussion*

The methods described in this section are not highly sophisticated, but they do present a practical and highly scalable approach to improving the quality of probe vehicle-based traffic data. It was shown that, even ignoring issues of GPS speed estimation accuracy, the distance-weighted speed estimation method outperforms pointwise speed estimation in all tested conditions. It was also shown that, by imputing missing speed values at the subpopulation level, both bias and accuracy

can be improved in a variety of traffic conditions. This is an important point, and one that underscores the importance of having insight into the sampling and data collection process.

The two scenarios presented here illustrate the fact that there are several causal mechanisms related to bias in probe vehicle data. First, there is the bias caused by the fact that slower moving vehicles are over sampled relative to faster moving vehicles. The second is that differences in sampling frequency can lead to certain vehicle subpopulations being over represented in point-based speed estimation methods. Both of these mechanisms are addressed fairly well by the inverse probability weighting scheme, albeit at a small increase in absolute error. Further more, these two mechanisms have little impact on the distance weighted speed estimation method, which demonstrates no systematic bias in scenario 1. The third source of bias is differing penetration rates, which again could lead to certain subpopulations being overrepresented. This is dealt with effectively by applying some imputation method on the data for individual subpopulations, and combining subpopulation speeds according the known or estimated population fractions represented by each subpopulation. These population fractions could be estimated by developing scaling factors based on known traffic populations at fixed locations where mechanical sensors are present, and then applying the scaling factors to adjust measured speed on a larger scale. From Figure 6-27 it is clear that applying imputation on pre-aggregated link-level speed data (ignoring the differences between subpopulations) will not always address the bias that arises in the sampling process. On the other hand, when traffic is congested as shown in , there is less difference between the subpopulations and hence less benefit to applying imputation at a subpopulation level. That said, it is clear that the proposed combination of distance-weighted pairwise mean speed estimation, imputation at the subpopulation level, and population weighting provides some benefit in all tested scenarios.

Chapter 7: Concluding Remarks

This dissertation has introduced and applied a conceptual and methodological framework for analyzing sample size, sampling bias, and missingness in probe vehicle data. Furthermore, it has introduced a set of application scenarios that provide a solid basis for the proposed framework as a powerful tool for addressing data quality issues unique to this source of data. Specifically, the following work has been completed, documented, and discussed in this dissertation:

- Introduced a quantitative framework for describing the sample size, missing data rate, and sampling bias in probe vehicle data under mixed traffic and sampling conditions, given the true traffic state and sampling parameters
- Introduced analytical and Monte Carlo sampling methods to estimate the joint distribution of sampling rate, missing data, and observed speed under mixed traffic and sampling conditions
- Validated the proposed framework in a variety of conditions using microscopic simulation
- Applied the principles of the proposed framework on real-world probe vehicle data to estimate data completeness on a variety of road segment geometries under varying traffic conditions
- Introduced planning-level approach to estimate the sample size and completeness for developing new probe vehicle data collection experiments
- Introduced a set of methods for addressing bias and improving the quality of probe vehicle data

This research makes a number of assumptions about what is known, and what can be known about a unique data collection process. The extent to which this work is useful in a given application is in part based on how well these assumptions align with the scenario of interest. This

work is intended to be forward looking in that individuals applying these methods are expected to have more control over the data collection process than the current state of practice. Even with these assumptions, this work build substantially on the current state of research on this topic (e.g. M. Ferman, Blumenfeld, and Dai 2005).

The results of the work that has been completed (including work that is not described in this document) suggest that the concepts described here apply to a wide range of traffic conditions. However, it is important to note that only a small fraction of all possible scenarios is presented here, and they are by no means representatives of all conditions that may be found in the real world. A particular mix of driving behavior, traffic, and sampling parameters may produce much more severe bias issues than are represented here. Or, conversely, different mechanisms that contribute to bias may cancel each other out, producing little or no bias in the measured data. For example, if a faster moving vehicle subpopulation had a slightly higher penetration rate, this may offset the fact that slower moving vehicles are more likely to be represented in the dataset. This underscores the importance of transparency in the data collection process, and the need for public agencies to be involved in or at least aware of the data processing and quality control methods applied.

There are a variety of applications for probe vehicle data that are not dealt with in this work, such as vehicle trajectories for signal timing analysis. Such applications have their own coverage, quality, and granularity requirements that may or may not be met by existing probe vehicle data sources. While this work deals with a specific set of quality issues unique to a particular use case, it is important to keep in mind that a variety of other applications could benefit from greater transparency on the part of the probe vehicle data providers. For example, in the signal timing and intersection performance analysis example, consumers (i.e. transportation

agencies) could benefit from a better understanding of the vehicle subpopulations represented in the probe population. If the probe population heavily favors a particular driver demographic or travel pattern, this could bias the results in a variety of ways.

Future extensions of this work include assessing the relationship between pair-wise distance weighted speed estimation accuracy, sampling interval length, and road segment length. Consider that, as the sampling interval increases relative to the travel time across individual road segments, any variation in speed due to traffic conditions, traffic control devices, and other factors will be effectively smoothed out. In addition, more work remains to be completed on the impact of traffic signals, access points, and other traffic complexities common in an urban environment. The analysis described here focusses on controlled access, non-signalized highways where stops and platooning are not present, though many of the same sampling-related issues will be present on signalized arterials and other facilities.

REFERENCES

Adan, I, and J Resing. 2001. “Queueing Theory.” http://wwwhome.math.utwente.nl/~scheinhardtwrw/queueingdictaat.pdf.

Allison, PD. 2003. “Missing Data Techniques for Structural Equation Modeling.” *Journal of Abnormal Psychology*. http://psycnet.apa.org/psycinfo/2003-10098-003.

Bends, Stine. 2017. “The Danish Road Directorate Traffic Management Centre.” Copenhagen, Denmark. http://www.nvfnorden.org/library/Files/Stine Bendsen_Trafikcentret præsentation.pdf.

Berk, Richard A. 1983. “An Introduction to Sample Selection Bias in Sociological Data.” *American Sociological Review* 48 (3). American Sociological Association:386. https://doi.org/10.2307/2095230.

Bitar, N. 2016. “Big Data Analytics in Transportation Networks Using the NPMRDS.” https://www.researchgate.net/profile/Naim_Bitar/publication/305994432_Big_Data_Analytics_in_Transportation_Networks_Using_the_NPMRDS/links/57a92b4608aef20758cd124c.pdf.

Boyce, David E., James Hicks, and A. Sen. 1991. “In-Vehicle Navigation Requirements for Monitoring Link Travel Times in a Dynamic Route Guidance System.” *Operations Review* 8 (1).

Brilon, Werner, Justin Geistefeldt, and Matthias Regler. 2005. “Reliability of Freeway Traffic Flow: A Stochastic Concept of Capacity.” In *Proceedings of the 16th International Symposium on Transportation and Traffic Theory*, 125–44. College Park, Maryland. http://www.ruhr-uni-bochum.de/verkehrswesen/download/literatur/ISTTT16_Brilon_Geistefeldt_Regler_final_citation.pdf.

Bucknell, Christopher, and Juan C. Herrera. 2014. “A Trade-off Analysis between Penetration Rate and Sampling Frequency of Mobile Sensors in Traffic State Estimation.” *Transportation Research Part C: Emerging Technologies* 46 (September). Pergamon:132–50. https://doi.org/10.1016/J.TRC.2014.05.007.

Cambridge Systematics, and Texas Transportation Institute. 2015. “NPMRDS Missing Data and Outlier Analysis.” https://www.regulations.gov/document?D=FHWA-2013-0054-0103.

Cameron, A Colin, and Pravin K Trivedi. 2012. *Regression Analysis of Count Data Second Edition*.

Cetin, Mecit, George List, and Yingjie Zhou. 2005. “Factors Affecting Minimum Number of Probes Required for Reliable Estimation of Travel Time.” *Transportation Research Record: Journal of the Transportation Research Board* 1917 (January):37–44.

https://doi.org/10.3141/1917-05.

Chen, Dewang, Long Chen, and Jing Liu. 2013. “Road Link Traffic Speed Pattern Mining in Probe Vehicle Data via Soft Computing Techniques.” *Applied Soft Computing* 13 (9). Elsevier:3894–3902. https://doi.org/10.1016/J.ASOC.2013.04.020.

Crackberry. 2007. “Real Time Traffic Information with Google Maps.” Crackberry.com. 2007. https://crackberry.com/real-time-traffic-information-google-maps.

Crowder, Senica. 2012. “RFP: Vehicle Probe Data and Analysis for Research on National Performance Measurement, System Planning and Transportation.” Federal highway Administration (FHWA) Office of Acquisition Management. https://govtribe.com/project/vehicle-probe-data-and-analysis-for-research-on-national-performance-measurement-system-planning-and-transportation.

D’Este, Glen M., Rocco Zito, and Michael A. P. Taylor. 1999. “Using GPS to Measure Traffic System Performance.” *Computer-Aided Civil and Infrastructure Engineering* 14 (4). Wiley/Blackwell (10.1111):255–65. https://doi.org/10.1111/0885-9507.00146.

Dion, Francois, Ralph Robinson, and Jun-Seok Oh. 2011. “Evaluation of Usability of IntelliDrive Probe Vehicle Data for Transportation Systems Performance Analysis.” *Journal of Transportation Engineering* 137 (3):174–83. https://doi.org/10.1061/(ASCE)TE.1943-5436.0000199.

Driscoll, Don A., Annabel L. Smith, Samantha Blight, and John Maindonald. 2012. “Reptile Responses to Fire and the Risk of Post-Disturbance Sampling Bias.” *Biodiversity and Conservation* 21 (6). Springer Netherlands:1607–25. https://doi.org/10.1007/s10531-012-0267-5.

Duret, Aurélien, and Yufei Yuan. 2017. “Traffic State Estimation Based on Eulerian and Lagrangian Observations in a Mesoscopic Modeling Framework.” *Transportation Research Part B: Methodological* 101 (July). Pergamon:51–71. https://doi.org/10.1016/J.TRB.2017.02.008.

Edie, L. 1965. “Discussion on Traffic Stream Measurements and Definitions.” In *Proceedings of the Second International Symposium on the Theory of Traffic Flow*, 139–54. Paris, France.

Eisenman, Shane B., Emiliano Miluzzo, Nicholas D. Lane, Ronald A. Peterson, Gahng-Seop Ahn, and Andrew T. Campbell. 2009. “BikeNet: A Mobile Sensing System for Cyclist Experience Mapping.” *ACM Transactions on Sensor Networks* 6 (1). ACM:1–39. https://doi.org/10.1145/1653760.1653766.

EUEIP. 2018. “European ITS Platform.” 2018. https://www.its-platform.eu/.

Famoye, Felix, and Weiren Wang. 2004. “Censored Generalized Poisson Regression Model.” *Computational Statistics & Data Analysis* 46 (3). North-Holland:547–60. https://doi.org/10.1016/J.CSDA.2003.08.007.

Ferman, Martin A., Dennis E. Blumenfeld, and Xiaowen Dai. 2003. "A Simple Analytical Model of a Probe-Based Traffic Information System." In *Proceedings of the 2003 IEEE International Conference on Intelligent Transportation Systems*, 263–68. Shanghai, China: IEEE. https://doi.org/10.1109/ITSC.2003.1251960.

Ferman, Martin, Dennis Blumenfeld, and Xiaowen Dai. 2005. "An Analytical Evaluation of a Real-Time Traffic Information System Using Probe Vehicles." *Journal of Intelligent Transportation Systems* 9 (1):23–34. https://doi.org/10.1080/15472450590912547.

Ferrari, Laura, and Marco Mamei. 2013. "Identifying and Understanding Urban Sport Areas Using Nokia Sports Tracker." *Pervasive and Mobile Computing* 9 (5). Elsevier:616–28. https://doi.org/10.1016/J.PMCJ.2012.10.006.

FHWA. 2018. "National Performance Management Research Data Set (NPMRDS) Information." Operations Performance Measurement Program. 2018. https://ops.fhwa.dot.gov/perf_measurement/index.htm.

FHWA Office of Operations. 2013. "Introduction to the National Performance Management Research Data Set (NPMRDS)."

Gary, Pike R. 2007. "Adjusting for Nonresponse in Surveys." In *Higher Education: Handbook of Theory and Research*, 411–49. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-1-4020-5666-6_8.

Glasser, M. 1964. "Linear Regression Analysis with Missing Observations among the Independent Variables." *Journal of the American Statistical Association*. http://www.tandfonline.com/doi/abs/10.1080/01621459.1964.10480730.

Gong, L, and W Fan. 2017. "Applying Travel-Time Reliability Measures in Identifying and Ranking Recurrent Freeway Bottlenecks at the Network Level." *Journal of Transportation Engineering Part A* 143 (8).

Graham, JW. 2009. "Missing Data Analysis: Making It Work in the Real World." *Annual Review of Psychology*. http://www.annualreviews.org/doi/abs/10.1146/annurev.psych.58.110405.085530.

Greene, Kate. 2008. "Tracking Traffic with Cell Phones." MIT Technology Review. 2008. https://www.technologyreview.com/s/411169/tracking-traffic-with-cell-phones/.

Greene, William H. 2005. "Censored Data and Truncated Distributions." *SSRN Electronic Journal*, May. https://doi.org/10.2139/ssrn.825845.

Hallenbeck, ME, and E McCormack. 2015. "Developing a System for Computing and Reporting MAP-21 and Other Freight Performance Measures." http://wadot.wa.gov/NR/rdonlyres/4869900F-9E88-4B2E-968B-EF2CA3B3D1FD/107207/8441.pdf.

Heckman, JJ. 1977. "Sample Selection Bias as a Specification Error (with an Application to the

Estimation of Labor Supply Functions).” http://www.nber.org/papers/w0172.

Henrickson, Kristian, and Yinhai Wang. 2016. “Implausible Ignorability: An Analysis of Factors Influencing Probe Vehicle Data Completeness Background: Probe Vehicle Data.” In *NATMEC*. Miami, FL. http://onlinepubs.trb.org/onlinepubs/conferences/2016/NATMEC/HenricksonPPT.pdf.

Herrera, Juan C., Daniel B. Work, Ryan Herring, Xuegang (Jeff) Ban, Quinn Jacobson, and Alexandre M. Bayen. 2010. “Evaluation of Traffic Data Obtained via GPS-Enabled Mobile Phones: The Mobile Century Field Experiment.” *Transportation Research Part C: Emerging Technologies* 18 (4). Pergamon:568–83. https://doi.org/10.1016/J.TRC.2009.10.006.

Hofleitner, Aude, Ryan Herring, Alexandre Bayen, Yufei Han, Fabien Moutarde, and Arnaud De La Fortelle. 2012. “Large Scale Estimation of Arterial Traffic and Structural Analysis of Traffic Patterns Using Probe Vehicles.” In *Transportation Research Board 91st Annual Meeting*. Washington, United States. https://hal.inria.fr/hal-00741497/.

Hosuri, Shaghayegh Rostami. 2017. “Congestion Quantification Using the National Performance Management Research Dataset.” University of Alabama at Birmingham. https://search.proquest.com/docview/1914681659?pq-origsite=gscholar.

I-95 Corridor Coalition. 2018. “Vehicle Probe Project.” 2018. http://i95coalition.org/projects/vehicle-probe-project/.

INRIX. 2006. “Frost & Sullivan Study Recognizes INRIX as the Leading Provider of Real-Time Traffic Information.” INRIX Press Reseases. 2006. http://inrix.com/press-releases/2583/.

Iteris Inc. 2018. “Utah DOT Adds HERE Traffic Data to Iteris Performance Measurement Platform.” Business Wire. 2018. https://www.businesswire.com/news/home/20180424005647/en/.

Jenelius, Erik, and Haris N Koutsopoulos. 2013. “Travel Time Estimation for Urban Road Networks Using Low Frequency Probe Vehicle Data.” *Transportation Research Part B: Methodological* 53:64–81.

———. 2014. “Probe Vehicle Data Sampled by Time or Space: Consistent Travel Time Allocation and Estimation.” https://people.kth.se/~jenelius/JK_2014.pdf.

Jestico, Ben, Trisalyn Nelson, and Meghan Winters. 2016. “Mapping Ridership Using Crowdsourced Cycling Data.” *Journal of Transport Geography* 52 (April). Pergamon:90–97. https://doi.org/10.1016/J.JTRANGEO.2016.03.006.

Jones, MP. 1996. “Indicator and Stratification Methods for Missing Explanatory Variables in Multiple Linear Regression.” *Journal of the American Statistical Association*. http://www.tandfonline.com/doi/abs/10.1080/01621459.1996.10476680.

Kerner, B. S., and H. Rehborn. 1997. “Experimental Properties of Phase Transitions in Traffic

Flow." *Physical Review Letters* 79 (20). American Physical Society:4030–33. https://doi.org/10.1103/PhysRevLett.79.4030.

Kim, Jiwon, Hani Mahmassani, and Jing Dong. 2010. "Likelihood and Duration of Flow Breakdown." *Transportation Research Record: Journal of the Transportation Research Board* 2188 (December). Transportation Research Board of the National Academies:19–28. https://doi.org/10.3141/2188-03.

Kim, Seoungbum, and Benjamin Coifman. 2014. "Comparing INRIX Speed Data against Concurrent Loop Detector Stations over Several Months." *Transportation Research Part C: Emerging Technologies* 49 (December). Pergamon:59–72. https://doi.org/10.1016/J.TRC.2014.10.002.

Kong, Qing-Jie, Qiankun Zhao, Chao Wei, and Yuncai Liu. 2013. "Efficient Traffic State Estimation for Large-Scale Urban Road Networks." *IEEE Transactions on Intelligent Transportation Systems* 14 (1):398–407. https://doi.org/10.1109/TITS.2012.2218237.

Lattimer, Charles, and Gene Glotzbach. 2012. "Evaluation of Third Party Travel Time Data." In *ITS America 22nd Annual Meeting & Exposition*. National Harbor, MD. https://trid.trb.org/view/1215889.

Little, John D. C. 1961. "A Proof for the Queuing Formula: L = λW." *Operations Research* 9 (3). INFORMS:383–87. https://doi.org/10.1287/opre.9.3.383.

Little, RJA, and DB Rubin. 2002. *Statistical Analysis with Missing Data, Second Edition*. https://books.google.com/books?hl=en&lr=&id=AyVeBAAAQBAJ&oi=fnd&pg=PT8&ots=uxY19FrRhA&sig=L11lyoCi4wfM-fVGijhWEVfTu14.

Liu, Kai, Meng-Ying Cui, Peng Cao, and Jiang-Bo Wang. 2016. "Iterative Bayesian Estimation of Travel Times on Urban Arterials: Fusing Loop Detector and Probe Vehicle Data." Edited by Tieqiao Tang. *PLOS ONE* 11 (6). Public Library of Science:e0158123. https://doi.org/10.1371/journal.pone.0158123.

Long Cheu, Ruey, Chi Xie, and Der-Horng Lee. 2002. "Probe Vehicle Population and Sample Size for Arterial Speed Estimation." *Computer-Aided Civil and Infrastructure Engineering* 17 (1). Blackwell Publishers Inc:53–60. https://doi.org/10.1111/1467-8667.00252.

Mansournia, Mohammad Ali, and Douglas G Altman. 2016. "Inverse Probability Weighting." *BMJ (Clinical Research Ed.)* 352 (January). British Medical Journal Publishing Group:i189. https://doi.org/10.1136/BMJ.I189.

Matsukidaira, Junta, and Katsuhiro Nishinari. 2003. "Euler-Lagrange Correspondence of Cellular Automaton for Traffic-Flow Models." *Physical Review Letters* 90 (8). American Physical Society:88701. https://doi.org/10.1103/PhysRevLett.90.088701.

McCutcheon, Allan L. 2011. "Sampling Bias." In *Encyclopedia of Survey Research Methods*, edited by Paul J. Lavrakas, 785. 2455 Teller Road, Thousand Oaks California 91320 United States of America: Sage Publications, Inc. https://doi.org/10.4135/9781412963947.n509.

Mcnamara, Margaret, Howell Li, Stephen Remias, Lucy Richardson, Edward Cox, Deborah Horton, and Darcy M Bullock. 2015. "Using Real-Time Probe Vehicle Data to Manage Unplanned Detour Routes Probe Data Dashboard." *Institute of Transportation Engineerings, ITE Journal* 85 (12):32–37. http://library.ite.org/pub/b772541b-04ab-0f6f-a863-34c8217c947e.

Mjolsness, Eric. 2004. "Labeled Graph Notations for Graphical Models." Irvine, CA. https://www.researchgate.net/profile/Eric_Mjolsness/publication/249868091_Labeled_graph_notations_for_graphical_models_Extended_Report/links/5436d3900cf2bf1f1f2d4125/Labeled-graph-notations-for-graphical-models-Extended-Report.pdf.

Nagatani, Takashi. 1995. "Bunching of Cars in Asymmetric Exclusion Models for Freeway Traffic." *Physical Review E* 51 (2). American Physical Society:922–28. https://doi.org/10.1103/PhysRevE.51.922.

Nanthawichit, Chumchoke, Takashi Nakatsuji, and Hironori Suzuki. 2003. "Application of Probe-Vehicle Data for Real-Time Traffic-State Estimation and Short-Term Travel-Time Prediction on a Freeway." *Transportation Research Record, Journal of the Transportation Research Board* 1855:49–59. https://doi.org/10.3141/1855-06.

Patire, Anthony D., Matthew Wright, Boris Prodhomme, and Alexandre M. Bayen. 2015. "How Much GPS Data Do We Need?" *Transportation Research Part C: Emerging Technologies* 58 (September). Pergamon:325–42. https://doi.org/10.1016/J.TRC.2015.02.011.

Qing Ou, R. L. Bertini, J. W. C. van Lint, and S. P. Hoogendoorn. 2011. "A Theoretical Framework for Traffic Speed Estimation by Fusing Low-Resolution Probe Vehicle Data." *IEEE Transactions on Intelligent Transportation Systems* 12 (3):747–56. https://doi.org/10.1109/TITS.2011.2157688.

Raghunathan, Trivellore E. 2004. "What Do We Do With Missing Data? Some Options for Analysis of Incomplete Data." *Annu. Rev. Public Health* 25:99–117. https://doi.org/10.1146/annurev.publhealth.25.102802.124410.

Romanillos, Gustavo, Martin Zaltz Austwick, Dick Ettema, and Joost De Kruijf. 2016. "Big Data and Cycling." *Transport Reviews* 36 (1). Routledge:114–33. https://doi.org/10.1080/01441647.2015.1084067.

Rubin, DB. 1976. "Inference and Missing Data." *Biometrika*. http://www.jstor.org/stable/2335739.

Schafer, JL. 1997. *Analysis of Incomplete Multivariate Data*. https://www.google.com/books?hl=en&lr=&id=3TFWRjn1f-oC&oi=fnd&pg=PR13&dq=Analysis+of+incomplete+multivariate+data&ots=2pLNsCBcl4&sig=Dr_hgv1azDXRzQ4VPIM8Xyaqc6M.

Schafer, JL, and JW Graham. 2002. "Missing Data: Our View of the State of the Art." *Psychological Methods*. http://psycnet.apa.org/journals/met/7/2/147/.

Schrank, David, and Tim Lomax. 2009. “2009 URBAN MOBILITY REPORT.” http://mobility.tamu.edu.

Schrank, David, Tim Lomax, and Shawn Turner. 2010. “TTI’s 2010 Urban Mobility Report.” http://mobility.tamu.edu.

Science, Web of. 2018. “Core Collection Basic Search.” 2018. apps.webofknowledge.com.

Seaman, Shaun R, and Ian R White. 2013. “Review of Inverse Probability Weighting for Dealing with Missing Data.” *Statistical Methods in Medical Research* 22 (3). SAGE PublicationsSage UK: London, England:278–95. https://doi.org/10.1177/0962280210395740.

Srinivasan, Karthik, and Paul Jovanis. 1996. “Determination of Number of Probe Vehicles Required for Reliable Travel Time Measurement in Urban Network.” *Transportation Research Record: Journal of the Transportation Research Board* 1537 (January). Transportation Research Board of the National Academies:15–22. https://doi.org/10.3141/1537-03.

Sterne, JAC, IR White, JB Carlin, M Spratt, and P Royston. 2009. “Multiple Imputation for Missing Data in Epidemiological and Clinical Research: Potential and Pitfalls.” *Bmj*. http://www.bmj.com/content/338/bmj.b2393.full.pdf.

Strauss, Jillian, Luis F. Miranda-Moreno, and Patrick Morency. 2015. “Mapping Cyclist Activity and Injury Risk in a Network Combining Smartphone GPS Data and Bicycle Counts.” *Accident Analysis & Prevention* 83 (October). Pergamon:132–42. https://doi.org/10.1016/J.AAP.2015.07.014.

Tilley, Aaron Alan. 2012. “The World’s Traffic Info Provider.” Seattle Business Magazine. 2012. http://seattlebusinessmag.com/article/worlds-traffic-info-provider.

Treiber, Martin, and Arne Kesting. 2013. “Trajectory and Floating-Car Data.” In *Traffic Flow Dynamics*, 7–12. Berlin: Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-32460-4.

Turner, S.M., and D.J. Holdener. 1995. “Probe Vehicle Sample Sizes for Real-Time Information: The Houston Experience.” In *Pacific Rim TransTech Conference. 1995 Vehicle Navigation and Information Systems Conference Proceedings. 6th International VNIS. A Ride into the Future*, 3–10. Seattle, WA: IEEE. https://doi.org/10.1109/VNIS.1995.518810.

Turner, Shawn M, William L Eisele, Robert J Benz, and Douglas J. Holdener. 1998. “Travel Time Data Collection Handbook, FHWA-PL-98-035.” https://www.fhwa.dot.gov/ohim/tvtw/natmec/00020.pdf.

Victor, Ronald G, Robert W Haley, DuWayne L Willett, Ronald M Peshock, Patrice C Vaeth, David Leonard, Mujeeb Basit, et al. 2004. “The Dallas Heart Study: A Population-Based Probability Sample for the Multidisciplinary Study of Ethnic Differences in Cardiovascular Health.” *The American Journal of Cardiology* 93 (12). Elsevier:1473–80.

https://doi.org/10.1016/j.amjcard.2004.02.058.

Wang, David. 2007. “Stuck in Traffic?” Official Google Blog. 2007. https://googleblog.blogspot.ca/2007/02/stuck-in-traffic.html.

Wang, Zhongxiang, Masoud Hamedi, Elham Sharifi, and Stanley Young. 2018. “A Cross-Vendor and Cross-State Analysis of the GPS-Probe Data Latency.” In *Transportation Research Board Annual Meeting*. Washington DC: Transportation Research Board. http://amonline.trb.org/2017trb-1.3983622/t005-1.4000488/200-1.4001272/18-05026-1.3997223/18-05026-1.4001275#tab_0=1.

White, IR, J Higgins, and AM Wood. 2008. “Allowing for Uncertainty due to Missing Data in meta-analysis—Part 1: Two-stage Methods.” *Statistics in Medicine*. http://onlinelibrary.wiley.com/doi/10.1002/sim.3008/full.

Winship, Christopher, and Robert D. Mare. 1992. “Models for Sample Selection Bias.” *Annual Review of Sociology* 18 (1):327–50. https://doi.org/10.1146/annurev.so.18.080192.001551.

Wolstan, Jeremy. 2007. “NAVTEQ Corporate, Traffic, & VII.” 2007. http://www.cmap.illinois.gov/documents/10180/213403/Navteq+Traffic.pdf/3a85f7dd-385a-44b3-aa33-4104bec48d47.

Wooldridge, Je¤rey M. 2010. *Econometric Analysis of Cross Section and Panel Data*. 2nded. Cambridge, Massachusetts: MIT Press. http://www.edu.gber.ge/uploads/files_85_1.pdf.

Work, Daniel B., and Alexandre M. Bayen. 2008. “Impacts of the Mobile Internet on Transportation Cyberphysical Systems: Traffic Monitoring Using Smartphones.” In *National Workshop for Research on High-Confidence Transportation Cyber-Physical Systems: Automotive, Aviation, & Rail*, 18–20. http://bayen.eecs.berkeley.edu/sites/default/files/conferences/cps2.pdf.

Xia, Chao, Courtney Cochrane, Joseph DeGuire, Gaoyang Fan, Emma Holmes, Melissa McGuirl, Patrick Murphy, et al. 2017. “Assimilating Eulerian and Lagrangian Data in Traffic-Flow Models.” *Physica D: Nonlinear Phenomena* 346 (May). North-Holland:59–72. https://doi.org/10.1016/J.PHYSD.2017.02.004.

Xiaowen Dai, Martin A. Ferman, and Robert P. Roesser. 2003. “A Simulation Evaluation of a Real-Time Traffic Information System Using Probe Vehicles.” In *Proceedings of the 2003 IEEE International Conference on Intelligent Transportation Systems*, 475–80. Shanghai, China: IEEE. https://doi.org/10.1109/ITSC.2003.1251999.

Yee, T. W., and C. J. Wild. 1996. “Vector Generalized Additive Models.” *Journal of the Royal Statistical Society. Series B (Methodological)*. WileyRoyal Statistical Society. https://doi.org/10.2307/2345888.

Yee, Thomas W. 2015. “Vector Generalized Linear and Additive Models: With an Implementation in R.” New York, USA: Springer.

Yuan, Yufei, J. W. C. van Lint, R. Eddie Wilson, Femke van Wageningen-Kessels, and Serge P. Hoogendoorn. 2012. “Real-Time Lagrangian Traffic State Estimator for Freeways.” *IEEE Transactions on Intelligent Transportation Systems* 13 (1):59–70. https://doi.org/10.1109/TITS.2011.2178837.

Zhang, Jia-Dong, Jin Xu, and Stephen Shaoyi Liao. 2013. “Aggregating and Sampling Methods for Processing GPS Data Streams for Traffic State Estimation.” *IEEE Transactions on Intelligent Transportation Systems* 14 (4):1629–41. https://doi.org/10.1109/TITS.2013.2264753.

Zheng, Fangfang, and Henk Van Zuylen. 2013. “Urban Link Travel Time Estimation Based on Sparse Probe Vehicle Data.” *Transportation Research Part C: Emerging Technologies* 31 (June). Pergamon:145–57. https://doi.org/10.1016/J.TRC.2012.04.007.