2019

# EVALUATE PROBE SPEED DATA QUALITY TO IMPROVE TRANSPORTATION MODELING

Fahmida Rahman
*University of Kentucky*, fra228@uky.edu

Digital Object Identifier: https://doi.org/10.13023/etd.2019.137

Fahmida Rahman, Student

Dr. Mei Chen, Major Professor

Dr. Timothy Taylor, Director of Graduate Studies

# EVALUATE PROBE SPEED DATA QUALITY TO IMPROVE TRANSPORTATION MODELING

THESIS

A thesis submitted in partial fulfillment of the
Requirements for the degree of Master of Science in Civil Engineering
in the College of Engineering
at the University of Kentucky

By

Fahmida Rahman

Lexington, Kentucky

Director: Dr. Mei Chen,

Associate Professor of Civil Engineering

Lexington, Kentucky

2019

# ABSTRACT OF THESIS

## EVALUATE PROBE SPEED DATA QUALITY TO IMPROVE TRANSPORTATION MODELING

Probe speed data are widely used to calculate performance measures for quantifying state-wide traffic conditions. Estimation of the accurate performance measures requires adequate speed data observations. However, probe vehicles reporting the speed data may not be available all the time on each road segment. Agencies need to develop a good understanding of the adequacy of these reported data before using them in different transportation applications. This study attempts to systematically assess the quality of the probe data by proposing a method, which determines the minimum sample rate for checking data adequacy. The minimum sample rate is defined as the minimum required speed data for a segment ensuring the speed estimates within a defined error range. The proposed method adopts a bootstrapping approach to determine the minimum sample rate within a pre-defined acceptance level. After applying the method to the speed data, the results from the analysis show a minimum sample rate of 10% for Kentucky's roads. This cut-off value for Kentucky's roads helps to identify the segments where the availability is greater than the minimum sample rate. This study also shows two applications of the minimum sample rates resulted from the bootstrapping. Firstly, the results are utilized to identify the geometric and operational factors that contribute to the minimum sample rate of a facility. Using random forests regression model as a tool, functional class, section length, and speed limit are found to be the significant variables for uninterrupted facility. Contrarily, for interrupted facility, signal density, section length, speed limit, and intersection density are the significant variables. Lastly, the speed data associated with the segments are applied to improve Free Flow Speed estimation by the traditional model.

KEYWORDS: Minimum Sample Rate, Bootstrapping, Probe Data Quality, Random Forests, Free Flow Speed.

Fahmida Rahman

April 25th, 2019

# EVALUATE PROBE SPEED DATA QUALITY TO IMPROVE TRANSPORTATION MODELING

By
Fahmida Rahman

Dr. Mei Chen
Director of Thesis

Dr. Timothy Taylor
Director of Graduate Studies

April 25th , 2019

DEDICATION

*To my parents, brother, teachers, and friends*

# ACKNOWLEDGMENTS

First, I would like to express my gratitude to the Almighty who has given me the wisdom, patience, strength and has made me successful in the completion of this research work. His consistent kindness over me keeps me focused and confident to do this research.

I am immensely grateful to my research advisor, Dr. Mei Chen, who helped me with her constant guidance and supported me throughout the course of this research. Her constructive feedbacks at various phases of this research assisted me in shaping up the deliverables and gaining insights.

I would like to thank my committee members Dr. Reginald Souleyrette, Dr. Gregory Erhardt and Dr. Mei Chen for their evaluation and instructive reviews on my document. Their feedback helped me to improve the document substantially.

Sincere thanks go to my co-workers, especially Xu Zhang who supported me from the very beginning of my Master's program here at the University of Kentucky. He helped me thoroughly with the conceptual understanding of my research. Furthermore, Jacob Brashear, Alex Mucci, Sneha Roy, Hasibul Islam were there to amend me with the challenging working environment here. Their perseverance and dedication towards work encouraged me making every possible attempts to achieve the goal. In addition, I am thankful to the departmental faculties and staffs for their time and effort in any situation of assistance.

Lastly, I am grateful to my parents and younger brother for believing in me to pursue my goals. I would like to also thank all of my friends in Bangladesh and here for their love and inspiration.

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# Chapter 1 Introduction

## 1.1 Background

Travel speed is a critical piece of information for many applications such as congestion management, air quality conformity analysis, and travel demand model calibration and validation. Speed data are used to estimate current traffic conditions. After knowing the conditions, travelers and agencies can make better decisions about how to use and manage the transportation network. Moreover, the Moving Ahead for Progress in the 21st Century (MAP-21) Act states that travel speed is an essential input while calculating measures to assess the nation's highway performance.

Traditional speed data collection methods such as loop detectors, radar guns, and floating cars require significant efforts and resources to achieve the desired accuracy. These are also limited to a corridor or a location device. Currently, transportation agencies capture traffic data primarily from fixed sensors that are relatively expensive to install and maintain. However, with the recent growth of communications technologies, Global Positioning System (GPS), and the mobile internet, an increasing amount of real-time location information is collected. This information is distributed by private companies and marketed for retail to public agencies such as state Departments of Transportation (DOTs). As a result, with the advances in GPS and communication technologies, speed data have become increasingly available through private data vendors.

Interestingly, traffic conditions are being monitored by using probe vehicles which utilize GPS technologies. Data generated by probe vehicles enable direct measurement of link travel times and travel speeds. These measurements can be subsequently used to estimate travel times across a road segment. The travel time information is essential for advanced traveler information system (ATIS) applications and real-time route guidance.

Probe data are obtained by aggregating compound technologies consisting of GPS, map matching and digital road maps. Using these technologies, the position of one's own vehicle and the corresponding time of that position are obtained calculating the latitude, longitude, elevation, and the time by receiving signal data from four GPS satellites. For different precision levels (e.g. one second, five seconds or five minutes), this GPS based probe vehicle reports the coordinate, travel speed, travel time, etc. The higher the frequency of reporting the position by the probe vehicle, the more accuracy can be expected.

There are advantages of using probe data. Probe vehicles cover a greater area at a lower cost and do not require expensive and maintenance-intensive equipment. Previously, there were issues with using probe data in terms of accuracy (e.g. reporting travel time or travel speed) compared to embedded sensors. To check probe data fidelity, a number of researches performed validation tests on these data supporting their applicability for operational purposes (*1, 2*). Consequently, probe-based speed data are being widely accepted for transportation applications, such as determining travel time reliability (TTR) metrics, congestion measures, improving transportation models, etc. (*3, 4*).

While using probe data for transportation modeling purposes, researchers face issues. One of the issues is the availability of probe data. For example, in a high-volume road, there is good temporal coverage of probe vehicles over the day. The frequency of reporting data is also high in this case. From the data, one can aggregate the speed or travel time data for 15 minutes or an hour. These aggregated data capture the temporal speed variations over a day. Conversely, a small number of probe vehicles traverse in a low volume road reporting data at a low frequency. Besides, some roads may not have night time data or some may not have off-peak period data during the day time. Thus, 15 minutes or hourly aggregation of these data may not capture the temporal speed variations properly due to missing data. If anyone evaluates the operational performance of a road network, which includes both high and low volume segments, it may not give reliable performance measures for all the segments. Hence, an adequate number of speed observations are required to be ensured for operational and modeling purposes.

The question is: how many data corresponding to a certain interval from a year would be deemed as adequate. It can be referred as sample size estimating the adequate respondents from the total number of the target population to be used.

Since inadequate sample size may lead to poor results, sample size determination is an important step before using data for transportation applications. Moreover, performance measures can be accurately determined with adequate speed data. Therefore, the challenge is determining the minimum probe speed data required for a segment to ensure that the speed estimates are within a permissible error range. This is defined as the minimum sample rate, which reports the minimum number of 5-minute epochs with probe speed data as a percentage (%) of all 5-minute epochs in a year.

Evaluation of probe data quality determines if the data from a road segment are adequate or not. The quality is checked by comparing the minimum sample rate with the data availability of a segment. If the data availability fulfills the minimum sample rate requirement, it indicates that the data are adequate. These data can be directly used to measure Free Flow Speed (FFS), hence, congestion measures. Moreover, the quality check also helps to decide whether an alternative procedure or model is required or not to obtain congestion measures for the segments with inadequate data. The data, where the availability is less than the required minimum sample rate, are defined as inadequate.

To the author's best knowledge, the question about the required minimum sample rate of probe speed data has not yet been properly addressed by research. This study proposes a method to estimate the minimum sample rate for different facility types, which is required to evaluate probe data quality for transportation modeling and operational studies.

## 1.2 Research Statement and Objectives

Due to limited coverage of probe speed data on some road segments, it is important to know how much data would be considered enough for transportation applications. The evaluation of the data quality can be done by estimating the minimum sample rate and comparing it with the probe data temporal coverage. If the temporal coverage of a segment satisfies the minimum sample rate requirement, the data can be trusted, and the segment will be referred as a segment with adequate data. This study addresses this probe data quality issue by setting the objectives stated below. The objectives are:

1. To develop a method for determining the minimum sample rate of speed data in order to evaluate probe data quality.
2. To identify and rank the significant factors that affect the minimum sample rate of interrupted and uninterrupted facilities by developing regression models.
3. To recommend facility specific regression models that correlate the significant factors with minimum sample rate. These factors will help the practitioners to have an idea on the data collection efforts.
4. To improve FFS estimation by calibrating the traditional model utilizing the adequate speed data.

The overall framework of this research, including its applications, can be pictured as in Figure 1. The first step is to propose a methodological framework of determining minimum sample rate for all facilities in Kentucky. The method follows a bootstrapping sampling procedure on probe speed data provided by the Kentucky Transportation Cabinet (KYTC) at the link level. The sampling procedure results in the minimum sample rate for each road segment. After that, the segments with adequate data will be identified by comparing data availability with the minimum sample rate. The next step will be applying the minimum sample rate of these segments to determine the factors that affect the minimum sample rate of a defined road facility. These factors are basically Highway Performance Monitoring System (HPMS) attributes which are listed as:

- Physical Attributes: Number of lanes, lane width, type of median, shoulder width, section length, etc.
- Accessibility Attributes: Density of signalized intersections, and density of access points.
- Mobility Attributes: Speed, and Annual Average Daily Traffic (AADT).

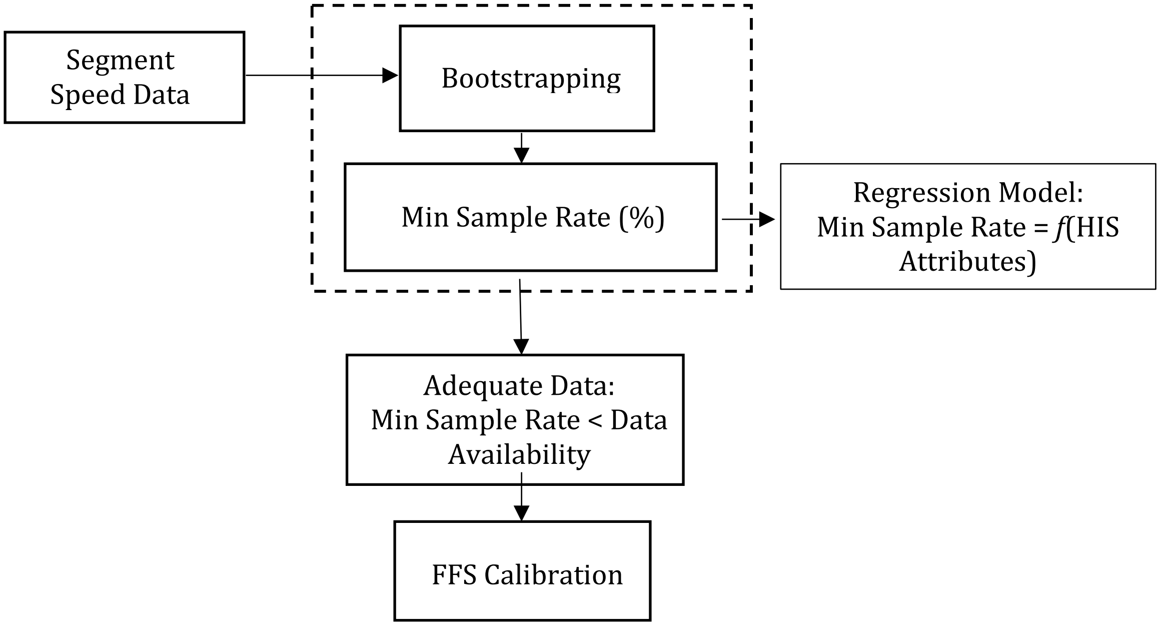


**Figure 1 Research Framework**

After identifying the factors, this study recommends a facility (for uninterrupted and interrupted) specific regression model to determine the minimum sample rate for a completely new road segment. This study also shows another application of the adequate speed data, which involves improving a FFS model. For this purpose, the FFS model is calibrated utilizing the speed data of these segments. Using the calibrated parameters in the model, the FFS of a road segment can be calculated with further accuracy.

## 1.3 Thesis Organization

This document consists of five chapters. Below are the contents of the chapters in brief.

- Chapter One: An overview of the research problem and research goals.
- Chapter Two: Literature reviews/ prior studies on the relevant field.
- Chapter Three: Data sources and processing, the methodological framework for minimum sample rate, and analysis results.

- Chapter Four: Application of minimum sample rate and probe speed data to find out the significant factors using regression models, results from regression models and comparison among the models, and FFS model improvement.
- Chapter Five: Discussions on the findings, a summary of the research, limitations, and future work.

This introductory chapter gives an overview of the research goals and the outline for the structure of this document. The next chapter presents the literature that provides an understanding of why there is a necessity for this study.

# Chapter 2 Literature Review

To begin, the author conducted a review on the current state-of-practice for checking the probe data fidelity before generating performance measures. Reliable performance measures require the adoption of different methods to evaluate probe data quality, which includes estimating probe vehicle sample size or minimum sample size of speed data. Therefore, this review helps to understand the rationale behind the adoption of different proxies for the estimation of the sample size. Furthermore, it provides a direction for this study focusing on the minimum sample rate as a measure of probe data quality evaluation. Although the existing researches on estimating minimum sample size for probe speed data are not significant, a review on the current practices brings forth the value of this study in providing a basis for future industry and research implementation.

In the following sections, a detailed background on probe data quality assessment to test its fidelity is presented. After that, studies that worked on estimating probe vehicle sample size and probe data sample size are documented. All these probe data quality assessments and sample size-based studies imply that performance measures require a valuation of data quality and adequacy. Thus, researches are going on to determine either minimum probe vehicle or minimum probe data requirements for assessing data quality. Despite focusing on the determination of minimum sample rate for speed data, this study also reviewed existing works regarding some statistical techniques on probe vehicle sample size.

## 2.1 Assessment of Probe Data Fidelity

Prevailing efforts on assessing probe speed data have been focused on validating the travel time measures based on this data by comparing with those derived from other data sources. A number of studies checked the fidelity of these data for freeways (*5*), whereas some other researches worked on the accuracy of the data for arterials (*1, 2, 6*). For freeways, Haghani et al. (*5*) compared INRIX data with Bluetooth traffic monitoring (BTM) data. During the validation process, they determined that road segments greater than one mile provided the most accurate speed measurements. Their analysis based on the speed distribution over the day concluded that speed data provided by INRIX is of good quality.

Eshragh et al. (*1*) indirectly checked the fidelity of probe data for the arterials based on roadway attributes. They developed a linear regression model which showed a strong correlation between the road attributes and probe data accuracy like average absolute speed error (AASE). They found that AADT, average signal density, and average access point density are correlated to probe data accuracy. They validated the regression model and concluded that the model can be used to predict the accuracy of the probe data indirectly for the arterial corridors.

Juster et al. (*2*) tested the probe data quality for arterials from case studies. They compared probe acquired data with BTM, where BTM was a ground truth data. Using hourly scatterplots for both sources, they tried to observe how well or poorly probe data capture the hourly travel time distribution. In addition, they calculated performance measures (e.g. travel time index, planning time index, etc.) using probe data and BTM data and compared the performance of probe data with respect to BTM. Finally, they suggested that probe data are suitable for arterials corridors that have an AADT greater than 40,000 vehicles per day, at least two lanes in each direction, and a signal density of one or less per mile.

The same authors from the studies (*1*) and (*2*) further researched on outsourced vehicle probe data of arterials. They adopted the same method from these two studies. Using the AASE regression model and BTM vs probe data hourly distribution plots, they recommended using probe data on arterials with signal densities (measured in signals per mile) up to one. However, they mentioned that probe data should be further investigated for signal densities between one and two.

Patire et al. (*7*) assessed probe data quality after fusing it with loop detector data. They used two data quality measures: sampling rate and penetration. According to the authors, "Sampling rate is the average rate at which any device reports its position and velocity. Any data set will have a distribution of devices with a range of sampling rates (typically between 0.5 and 60 reports per minute) whereas penetration rate is the flow fraction of vehicles (unique devices) reporting to the probe data set as compared to the total flow of vehicles along a road". After investigating these two measures, they preferred a high penetration rate of probes over a high average sample rate. It indicates that quality probe data can be achieved by having less frequent data from a larger number of unique probe vehicles, than having more frequent data from a smaller number of unique probes.

Probe data are also used as a substitution of the modeled travel time or speed data. Florida Department of Transportation (FDOT) supported HERE data, which is a source of probe data, as a replacement for modeled travel time/speed data (*8*). Before making a conclusion on HERE data, FDOT checked the quality of the data. The quality check included Turkey method that ranked all travel times for a road section and treated any value greater than the 75th percentile plus 1.5 times the interquartile distance, or less than the 25th percentile minus 1.5 times the inter-quartile distance as an outlier. Secondly, they checked if two consecutive travel times change more than 40% or not. Thirdly, they removed a travel time data if it is more than one standard deviation above or below the moving average of the 10 previous entries.

Washington State DOT (*9*) compared INRIX, Sensys, Traffic Cast, Blip System, and BlueTOAD data to investigate their quality. They evaluated their accuracy based on travel time distribution plots, Mean Absolute Deviation, Mean Percent Error, Mean Absolute Percent Error, and Root Mean Squared Error. They found that if accuracy drops below a critical limit (Mean Absolute Percent Error =25%), it is wise to avoid that data source. Moreover, if the sample count and penetration are much lower, the travel time for that data cannot be representative. After observing these error measures, sample counts and penetrations rates for all the data sources, they suggested that INRIX data from probe vehicle have wide availability and more accuracy (*9*).

Furthermore, KYTC (*10*) looked at the temporal coverage of the probe data to evaluate its quality. As the percentage of 15-minute epochs with probe data decreases, confidence in the data diminishes. A minimum threshold of 1% temporal coverage (measured by the percentage of 15-minute epochs with probe data) was considered acceptable. If the probe data satisfy this threshold, they are suitable for measuring and tracking the performance of roadways across several years (*10*). Such data can help DOTs and Metropolitan Planning Organization (MPOs) identify bottlenecks in the network, prioritize improvement strategies, and assess the effectiveness of projects. Therefore, with the assessment of probe data and acceptance given by the DOTs, the fidelity of this data source has been verified.

To set reference speed for freeways and arterials, Jha et al. (*11*) used the travel speed data from INRIX database. INRIX provided annual average travel times for each 15-minute interval of each day of a week. While setting the reference speed, they essentially had to know the data adequacy of probe speed data for a certain interval, so that they could have a reliable reference speed. Furthermore, the Texas Transportation Institute (*12*) recommended using probe data in measuring delay, determining the level of service, and evaluating signal operations. They evaluated the quality of INRIX data before doing the analysis. For this, they compared INRIX data with Bluetooth data. They determined that the INRIX speed data sufficiently reflected the ground-truth Bluetooth speed data and were suitable for the applications. Zhang et al. (*13*) also assessed INRIX probe data quality with respect to Bluetooth data. To measure the accuracy of INRIX data, they used correlation matrix and correlation plot for each performance metrics like travel time index, planning time index, etc., and compared them for INRIX and Bluetooth data. The comparison concluded that INRIX data are suitable for calculating reliability measures of the segments with homogeneous lanes as well as for performance reporting.

The discussion above shows that the DOTs and other organizations are using probe-based data to reliably measure performance metrics for different facilities. While a few of the studies are assessing the probe data quality before using the data, the majority are not likely to explain how they test the data quality for their applications. It should be of great concern since inadequate data may not give strength to performance measures. Several studies are trying to deal with this issue and propose data quality test giving some guidelines based on either probe vehicle sampling or probe data sampling. The next section gives a brief on different methodologies regarding probe vehicle sampling.

## 2.2 Existing Literature on Probe Vehicle Sample Size

The accuracy of travel information depends on the number of instrumented probe vehicles. Several probe vehicles traversing in the traffic stream can potentially provide valuable information about current travel times or travel speed. However, too few probe vehicles can provide erroneous or misleading data, weakening the credibility of the transportation agency and eroding public confidence in the traffic management system. Due to this issue, studies are determining the sample size of the probe vehicles in a traffic stream. This is defined as the probe vehicle sample size providing the minimum number of probe vehicles required on a certain road for travel information accuracy. The researches on the probe vehicle sample size follow different statistical approaches. These can be explored to gather knowledge of the general procedure for sampling analysis. Most of these studies are undertaken with a major focus on average travel time/speed estimation within a specified acceptance level. For example:

- Generally, statistical sampling theory is used for the required probe vehicle sample size to reliably estimate link travel time/speed. It assumes that travel times/speeds on links follow a normal distribution or *t* distribution. However, formulations based on the *z*-statistic can be performed only if the sample size is greater than 30 (*14*). On the other hand, formulations based on the *t*-statistic has no closed form solution, and an iterative procedure must be applied to search the possible sample size (*14*).
- One caution in using the sample size determination formulations is that the assumptions do not always hold under interrupted traffic flow. This may be the case that link travel times appear to have multistate features (*15-18*).
- Zhou et al. (*19*) used a simulation-based approach to determine the sample size of probe vehicles with consideration of road network coverage and link average speed estimation. The estimation accuracy increased little and the efficiency decreased significantly with the increase of probe sample size when it reached a certain level, as proved by microscopic VISSIM simulation. However, these Simulation-based sampling studies lack enough calibration and validation from real-world data, which limit their applications.

- The findings from Chen et al. (*20*), Cetin et al. (*21*), and Miwa et al. (*22*), in general, concluded that road geometrics, traffic volumes, estimation accuracy, and the characteristics of the activity along the road would influence the required minimum number of floating cars.

Regardless of the above-mentioned studies, the mean and standard deviation are the main measures used for estimating the required probe vehicle sample size with reasonable precision in most of the existing studies. However, these measures contain only a portion of information conveyed by the probability distributions, which have intimate relationships with underlying traffic conditions (*23, 24*). Overall, this section gives an idea on several sampling techniques to estimate the minimum sample size for probe vehicles or floating cars. At the same time, it helps to know about the shortcomings of these techniques separately, if one wants to adopt the methods for any sampling research.

## 2.3 State-of-Practice in Minimum Sampling Rate of Speed/Travel Time Data

Earlier studies on the data quality check were mostly focused on finding the minimum sample size of the probe vehicles or floating cars. Few of the studies worked on the sample size determination of speed data. This section documents the existing few efforts to find the minimum sample size for travel time data or speed data generated from different sources (probe vehicle or other agencies). After analyzing those studies and their limitations, the motivation of this research can be justified.

For travel time studies, it is important to check the data quality before producing reliability measures. Several studies adopted different approaches for checking data quality collected from different sources. For example, some were looking at travel time distributions, or some are building parametric and non-parametric models. An overview concerning those is discussed below.

- Previous research randomly selected the time periods for their travel time data without additional justification. Due to the limited availability of traffic sensor data in the 1970s, Polus (*25*) manually collected 211 samples of travel time data, used Gamma distributions to fit the data and then estimated the reliability measure accordingly.

- As traffic sensors were installed on more roads, researchers gained access to more data, allowing Van Lint and Van Zuylen (*26*) to use data for the entire year of 2002 to build travel time distributions. Although Emam and Al-deek (*27*) used four weeks of data to fit their selected statistical distributions, Higatani et al. (*28*) utilized a full year of expressway data in their study. All these studies did not show the premise behind using four weeks or one-year worth of data.
- Kwon et al. (*29*) used a non-parametric model to fit 256 non-vacation days of data in their case study, stating that "the sample size is large enough", while Yazici et al. (*30*) utilized nearly 11 months of data (from Jan.15, 2010 to Nov. 28, 2010) to build their travel time distributions.
- Yang et al. (*31*) determined the minimum sample size required to build stable travel time distributions for freeway TTR by proposing both parametric and non-parametric method. However, their analysis was based on the presumed distribution of travel time data.
- Yang and Cooke (*32*) applied a bootstrapping approach to identify the optimal size of travel time data for measuring freeway TTR.

Sample sizes for speed studies can vary from a fraction of an hour to 24 hours a day to 365 days a year, depending on the purpose of the study. Generally, for the speed study, peak hours are included in all samples (*33*). Holiday or on the day before or after a holiday is excluded for taking traffic counts. Normally, Monday mornings and Friday evenings show high volumes.

Oppenlander (*34*) developed a procedure for sample size determination based on the average range of the observed travel speeds. The estimate of the standard deviation of travel speeds can also be used for the similar purpose. Quiroga and Bullock (*35*) developed a hybrid method for the determination of sample size. The sample size estimation is shown in Equation (1).

$$n = \left[\frac{t_\alpha \times \bar{R}}{d\varepsilon}\right] \qquad (1)$$

Where; $n$ = minimum sample size

$t_\alpha$ = normal, two tailed statistics for a confidence interval of 1-α

$\bar{R}$= average range

$d = \frac{\bar{R}}{\sigma}$

$d$= factor for estimating $\sigma$ from $\bar{R}$

$\varepsilon$ = user-selected allowable error

Li et al. (*36*) suggested a modified method to determine the sample size. Use of $Z_{\alpha/2}$ values in place of $t_{\alpha/2}$ values induced some error in sample size. As a result, they proposed a correction factor for the calculation of sample size. Revised equation is given below,

$$n = \left[\frac{Z_{\alpha/2} \times \sigma}{\varepsilon}\right]^2 + \varepsilon_n \qquad (2)$$

Where; $n$ = minimum sample size

$\sigma$= population standard deviation

$\varepsilon$ = user-selected error

$\varepsilon_n$= sample size adjustment

Barnett et al. (*37*) studied regression to mean (RTM) phenomenon and concluded that it is a ubiquitous phenomenon in repeated data. It should always be considered as a possible cause of an observed change. Use of better study design and suitable statistical methods can be used to alleviate these effects. Park and Lord (*38*) adopted graphical methods to illustrate the RTM phenomenon. They used aggregated speed data to show how to reduce RTM bias in before-and-after speed data analysis. From the numerical examples, the estimated magnitude of the mean speed change can be misleading due to the introduction of an engineering treatment and the amount of uncertainty which can be measured by the estimated standard error and confidence interval. This problem can be addressed by accounting RTM.

Varsha et al. (*33*) determined sample size for speed obtained using a video-graphic survey on Urban Arterials. Their assumption was made based on speed variability and traffic conditions. They showed that there was not much variation in speeds for a given vehicle type and location. The mean and variance in speeds, obtained from first ten-speed measurements for a vehicle type, were not statistically different from those obtained after one hour of data collection. However, for some locations, where the proportion of heavy vehicles and flow were low, statistically stable mean speeds of heavy vehicles could not be obtained even after hours of data collection.

From the above discussions on sampling size methods, the existing literature shows a gap to give a proper guideline for minimum sample rate while using probe data. Most of them assume a fixed distribution of the speed data or travel time data. Moreover, the minimum sample rate may vary based on facility type. Existing methods either work on freeways or arterials. No research is found that integrates both uninterrupted and interrupted facility to give a direction on the minimum sample rate required. There is still a need for research confirming the minimum sample rate for both facility types using a method that does not require an assumption on the data distribution.

An increasing number of agencies are embracing probe speed-based measures to quantify congestion and TTR. However, no specific study worked on determining the minimum sample rate for further utilization in improving transportation models. Thus, it is important to know what percentage of the whole year speed data will be enough for operational and transportation modeling purpose. In search of the answer to this research question, the next chapter shows the subsequent methodology with a brief on the data sources.

# Chapter 3 Determining Minimum Sample Rate

This chapter gives an overview of the data sources and the method involved in achieving the objectives of the research presented in section 1.2.

## 3.1 Data Sources and Processing

This study uses historical speed data acquired from a third-party data provider, HERE Technologies (*10*). They provided probe speed data from 2017 on all Kentucky roadways including all facility types. These speed data were attached to the links of the HERE Street Network. This network is called HERE 2017Q3 Street Network. HERE provided probe speed data for all vehicles, cars only, and trucks only. The total number of links on the most recently updated HERE street network is 1,033,842 for Kentucky State. Probe vehicles traversing through these links report 5-minute epochs speed data daily. These 5-minute intervals of data were obtained for this study's purpose.

HERE datasets come into two forms: GPS probe-based, and GPS path-based. In this study, GPS path-based is used. This path-based approach involves tracking probe trajectories, computing space mean speed and integrating it with the GPS point-based speed in a link to produce a path-processed dataset. These path-based speeds were then assigned to all links that were part of the path. As a result, links that probe vehicles traversed but that were not polled for instantaneous speeds would be included. A trial analysis of the Lexington area indicated that path-processed datasets contain about 50% more records than the probe-based data (*10*).

The HERE database contains directional speeds, probe vehicle sample counts, i.e. the number of intervals, mean, minimum, maximum, and standard deviation of the probe speeds in 5-minute intervals over a year. For this study's purpose, afternoon peaks (time span: 3 pm- 6 pm) from non-holiday weekdays were included. During the afternoon peak, the traffic demand is high, and so the probe vehicle coverage is high. The probe data used in this study provided 18.73% data for afternoon peaks, 13.3% data for morning peaks, and 14% data for midday peaks with respect to the total 5-minute epochs for the whole year. It indicates that the number of 5-minute epochs that have probe data during the afternoon peak period is maximum compared to other peak periods of the day. To determine the minimum sample rate of a road segment, a time period which provides high probe data coverage and gives confidence in the estimated sample rate is desirable. In this case, the afternoon peak period provided high data coverage. Moreover, the randomness in speed data was greater during this period. Figure 2 shows the spatial coverage of data over the state of Kentucky during afternoon.

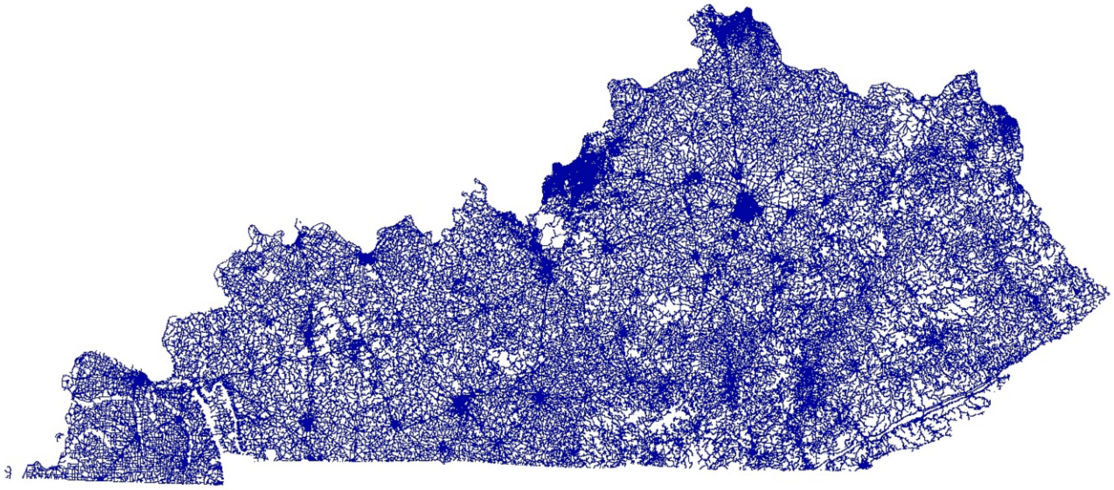


**Figure 2 HERE Link-Referenced Network of 2017**

This study also used roadway geometry, condition, and usage data from KYTC's Highway Information System (HIS). Table 1 shows the KYTC provided list of the data items in the HIS database.

**Table 1 List of Data Items**

| Items | |
|---|---|
| Pavement Type | Functional System |
| Facility Type | Peak Lanes |
| Area Type | Lane Width |
| At Grade Signal | Right Shoulder Width |
| At Grade Stop | Left Shoulder Width |
| At Grade Other | Peak Truck Percentage |
| Section Length | Daily Truck Percentage |
| Through Lanes | Interchanges |
| Median Type | Speed Limit |
| Median Width | Percent of Passing Sight Distance |
| Access Control | Truck Climbing Lane |
| Terrain Type | Turning Lanes |
| AADT | Peak Parking |
| K Factor | Green Ratio |
| D Factor | Curve |
| | Grade |

HIS network provided the data at a segment level over the state. The HIS segments were mapped with HERE links using conflation methodology. The method used HERE speed data to create performance measures for the HIS network in Kentucky. The approach followed transforming one network (e.g. HERE street network) to a point shapefile and then projected it to the other network (e.g. HIS network). Details on this method can be found in the research by Green et al. (*39*), where an automated conflation process to integrate two networks was developed. Following this method, the HERE 2017Q3 network was conflated with KYTC's HIS network. However, mismatches due to network complexities were an issue. To address the issue, a set of screening rules was developed to facilitate the quality assurance process based on functional class mapping and network connectivity (*40*).

Originally, this study performed the analysis on the HERE links. A particular HIS road segment consists of a number of HERE links. The minimum sample rates were estimated for each link and were aggregated for the whole HIS segment. The network mapping of HIS and HERE street allowed the determination of minimum sample rate at HIS segment level. The aggregation from the link level to segment level will be described in the later section 3.4.

## 3.2 Statistical Measures to Estimate Minimum Sample Size

This section discusses some statistical measures used by several studies estimating sample sizes along with their limitations.

Currently, statistical measures, such as standard error, confidence intervals, and statistical distributions are widely used to estimate the minimum sample size from a given dataset. However, those measures face issues while following strict assumptions.

Yang and Cooke (*32*) used standard error to assess the accuracy of statistical estimators like sample mean and confidence interval. To estimate freeway corridor travel time within a short time period, *n* observed travel times were recorded in their study. The estimated standard error of a mean, $\bar{s}$, based on the *n* independent observed travel speeds ($s_1, s_2, \ldots, s_n$) were calculated using Equation (3).

$$\text{Standard error} = \sqrt{\frac{\sum(s_n - \bar{s})^2}{(n-1)*n}} \qquad (3)$$

The measure of accuracy was derived based on this standard error. For this, the authors estimated the margin of error (ME) associated with a particular confidence interval. This ME was then used to determine sample size in their study. However, there are issues with estimating the sample size requirement in this way, especially for travel time measurement (*41, 42*). One of the major disadvantages is that the equation assumes a normal distribution. Moreover, no explicit equation is found for measuring standard error of the most statistical estimator (median, n$^{th}$ percentile, standard deviation, etc.).

Conventional methods of measuring accuracy are unsuitable for travel time and congestion study using speed data. Studies (*24, 43, 44*) show that freeway travel time does not follow a strict distribution. The freeway travel time distribution is sometimes left skewed or sometimes right skewed without having a unique common distribution to represent the travel time. Additionally, congestion studies do not concern only calculating the mean of available speed data, but also checking speed data adequacy over the year before using the data. Hence the question arises, are the available speed data enough to obtain reliable congestion or reliability measures? What should be the required minimum speed data to get credible measures? Therefore, a methodological framework is required to have confidence in data and to evaluate the accuracy of the estimator.

Following along with the line of existing literature, there is a need to investigate minimum sample rate without accounting for a fixed assumption on the distribution of speed data. The investigation also involves choosing an accuracy measure to examine the acceptance of the estimated sample rate corresponding to an acceptable error value.

## 3.3 Study Method to Estimate Minimum Sample Rate

After counting the issues in the existing practices, this section describes the method to determine the minimum sample rate without assuming a distribution of the speed data. The method included an algorithm that allowed to repeatedly sample a percentage of data using a bootstrapping resampling approach and ensured that the sample means were within a small margin (e.g., 5%) of the population mean.

Bootstrap resampling method is a simple procedure that involves repeatedly resampling from the available data to develop a number of plausible data sets that might have been observed under different circumstances. Each successive individual bootstrap replication has a data size that is equal to a predefined sample size. For example, assume that $t = (t_1, t_2 \dots t_{15})$; one individual replication with replacement for 80% size of this dataset could be $t* = (t_1, t_2, t_1, t_1, t_3, t_5, t_7, t_7, t_8, t_9, t_{11}, t_{14})$. Such resampling procedure is performed *m* times to create *m* replications. This sampling method was used to create replications of the original speed dataset in this research. The sampling process is graphically presented in Figure 3.

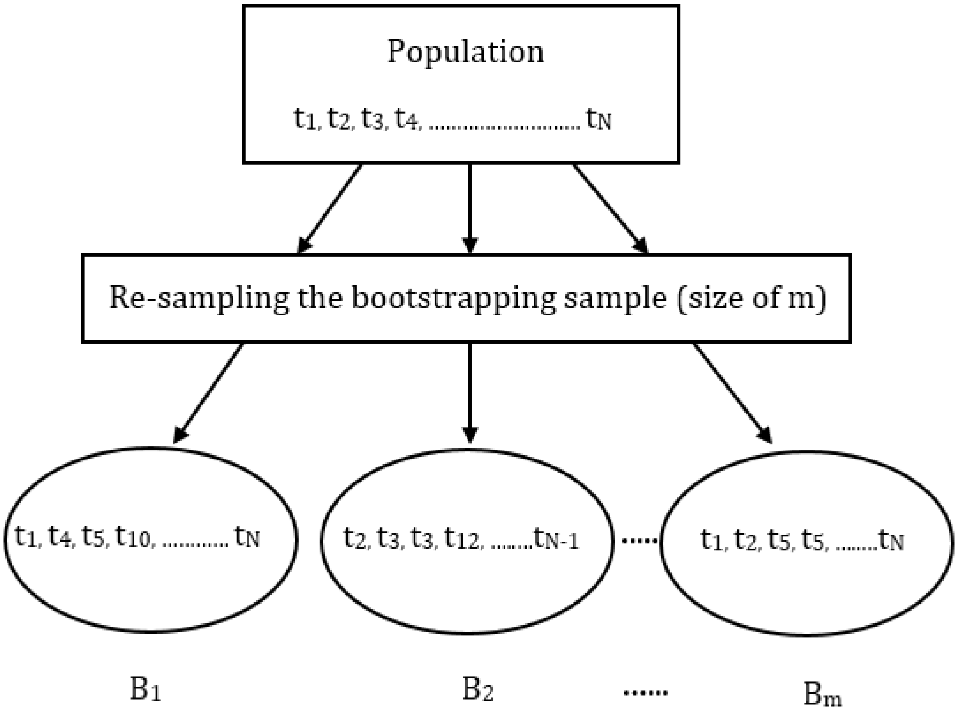

Population
t1, t2, t3, t4, ........................... tN
Re-sampling the bootstrapping sample (size of m)
t1, t4, t5, t10, ............ tN
t2, t3, t3, t12, ........tN-1
.....
t1, t2, t5, t5, ........tN
B1
B2
......
Bm

**Figure 3 Bootstrap Sampling Procedure**

The classical procedure of bootstrapping involves sampling through replacement to have multiple duplicate observations in the bootstrapped replications. The replacement helps to account for the samples that are close to the observations in a dataset. Moreover, the bootstrapping method allows for estimating the distribution of various parameters such as the sample mean (*45*). It treats a sample of data (from observations/from simulation) as a new population. Furthermore, it allows for determining multiple estimates of the parameter of interest. The most important advantage of this method is that no assumption is required about the underlying distribution of the population. In addition, uniform resampling is done from the original observations. This study implemented the method because of these advantages.

Finally, the algorithm to determine the minimum sample rate of probe speed data for a link consists of the following steps:

- Feed original speed dataset and treat this dataset as population.
- Define population size, *N*, of the original speed dataset.

- Estimate population mean using Equation (4).

$$\mu = \frac{1}{N}\sum_{k=1}^{N} y_k \quad (4)$$

Where; $y_k$= $k^{th}$ speed data

$N$ = size of the population

- Directly apply the bootstrapping to those original speed datasets with varying sample rates ($x$ %). For example: if the original dataset contains 100 data points, a sample rate of 20% should mean that 20 data points will be uniformly generated with a random selection from the dataset, by performing bootstrapping resampling to this original (population) dataset.
- Get bootstrap samples of the speed dataset using random sampling with replacement, which means $m$ times replications of $x$% sample from a population size of $N$ are produced. To select the replication number that should fulfill the purpose of this study, other existing efforts can be referred here. For example, Efron and Tibshirani (*46*) recommended that 1,000 bootstrapped replications are sufficient to estimate standard errors. Besides, 2,000 replications are sufficient to estimate confidence intervals (*46*). Since the percentile confidence interval is used as an accuracy estimate of bootstrapped samples in this study, 2,000 replications are enough to do the process.
- Calculate the parameter of interest, i.e. sample mean, of each bootstrapped sample using Equation (5).

$$\bar{y} = \frac{1}{n}\sum_{k=1}^{n} y_k \quad (5)$$

Where; $y_k$ = $\mathrm{k^{th}}$ speed data in the bootstrapped samples

$n$ = size of the bootstrapped samples i.e. $x\%\ of\ N$

- Form a new data set based on the $m$ number of sample means, $\bar{y}$ , that are already calculated from the $m$ replications of the speed data.
- Approximate the distribution of this set of sample means. In this study, a sampling distribution of the means is created. This distribution depends upon the distribution of original data.

- Use the approximate distribution to obtain percentile confidence interval (CI) of the sample means. The bootstrap method suggests that approximately 95% of the time, the population mean, $\mu$, falls between the 2.5th percentile and the 97.5th percentile of the bootstrap sample means. This is also known as the 95% CI.
- Calculate ME using 95% CI as presented in Equation (6).

  $$ME = \frac{(\bar{y}_{0.975} - \bar{y}_{0.025})}{2} \quad (6)$$

  Where; $\bar{y}_{0.975}, \bar{y}_{0.025}$ = 97.5th and 2.5th percentile of the bootstrapped sample means respectively
- Set an acceptable error rate in percentage, $\varepsilon$ , which is introduced to define how much the error can differ compared to the population mean, $\mu$. If the ratio of ME to the population mean, $\mu$, exceeds the defined error rate, $\varepsilon$, as shown in the Equation (7), the algorithm will increase the value of sample rate, $x$%. Hence, the process continues until an error converges to the defined $\varepsilon$.

  $$\frac{ME}{\mu} \times 100 > \varepsilon \quad (7)$$
- Report the corresponding sample rate, x%, as the minimum sample rate, once the ratio of ME to $\mu$ converges to $\varepsilon$.

The Framework for the minimum sample rate algorithm is shown in Figure 4.

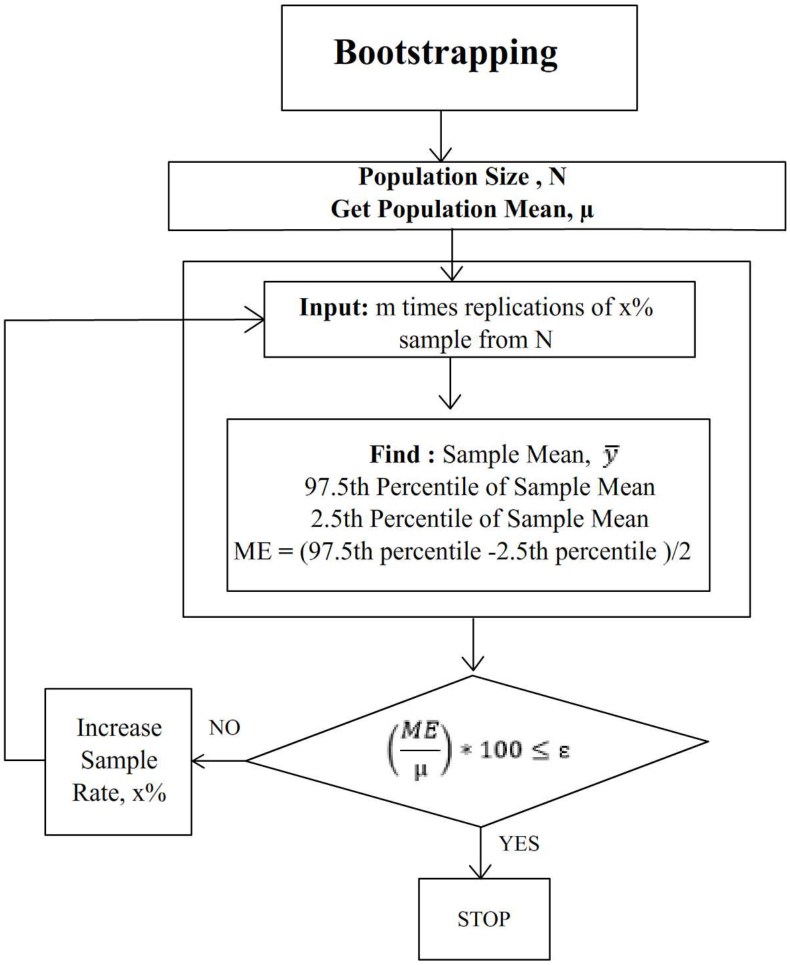
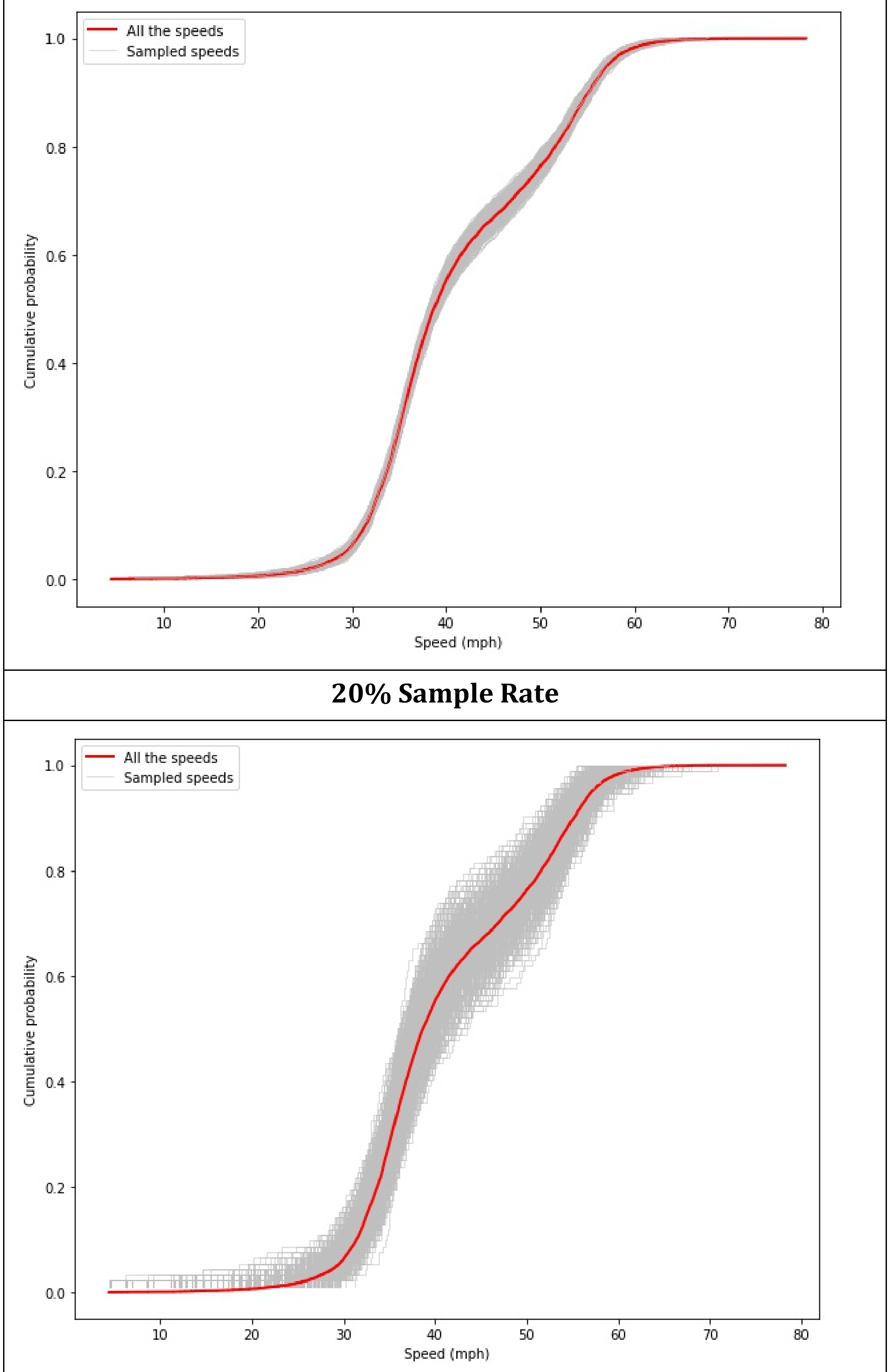

Bootstrapping
Population Size , N
Get Population Mean, μ
Input: m times replications of x% sample from N
Find : Sample Mean, $\overline{y}$
97.5th Percentile of Sample Mean
2.5th Percentile of Sample Mean
ME = (97.5th percentile -2.5th percentile )/2
$\left(\frac{ME}{\mu}\right) * 100 \leq \varepsilon$
NO
Increase Sample Rate, x%
YES
STOP

**Figure 4 Methodological Framework for Minimum Sample Rate**

An example to demonstrate a realistic reflection of the method is shown in Figure 5. It shows the Cumulative Distribution Function (CDF) curves for 2,000 bootstrap replicated datasets, where Figure 5(a) is for 20% sample rate and Figure 5(b) is for 1% sample rate. To compare these 20% and 1% replications with the original data, CDF for the original dataset is also added in the figures using a red curve. From Figure 5, the distribution of the data converges on some value with increasing sample rate. Moreover, the distribution of probable results in the tails is wider than the median. Consequently, more data are required to confirm that any parameter estimated from the tails of the distribution has converged to within the same tolerance that might be used for the mean.



**20% Sample Rate**

**1 % sample rate**

**Figure 5 Cumulative Distribution Plots for (a) 20% Sample Rate, (b) 1% Sample Rate for Speed Data**

For the 1% sample rate shown in Figure 5(b), a broader range lies for a 95% CI. On the contrary, for the 20% sample rate presented in Figure 5(a), a narrower range lies within the 95% CI. As decided before, this range is allowed for 5% times (error rate) of the true population mean for calculating minimum sample rate.

### 3.4 Minimum Sample Rate Analysis and Results

The bootstrap minimum sample rate method was applied to HERE link speed data from non-holiday weekday afternoon peaks in 2017. The high data coverage during the afternoon peak periods well represents the population characteristics, hence, the estimate of the population mean. Moreover, it was assumed that the probe speed data represents ground truth. After applying the method to the data, the results were analyzed separately based on facility types to define a threshold of minimum sample rate for the road segments in Kentucky. This section gives an overview of the analysis and the results from the method.

Bootstrapping was applied to the HERE extracted speed dataset. The sampling process started with a bootstrap sample set containing only a single speed data (e.g. 1-speed data from a population set of 100 data means 1/100 = 0.01% sample rate). The whole process continued until it yields an error rate, $\varepsilon$, of ±5% pertaining to 95% CI. The sample rate corresponding to the error rate was reported as the minimum rate for the HERE link. After obtaining minimum sample rates for all the HERE links, these were aggregated to the HIS segment levels. Using Equation (8), the length weighted average of the link estimated sample rates was calculated for the aggregation on the segment level.

$$Minimum\ Sample\ \ Rate(Segment\ Level),\ MSSL_j = \frac{\sum_1^i MS_{ij} \times L_i}{\sum_1^i L_i} \qquad (8)$$

Where;

$MS_{ij}$ = minimum sample rate for the corresponding $i^{th}$ HERE link of the $j^{th}$ segment from HIS extracts

$L_i$ =length of the $i^{th}$ HERE link

Equation (8) gives the minimum sample rate for each of the segments in the HIS network. Finally, results were analyzed separately for all the transportation facilities in Kentucky to decide a threshold for the minimum sample rate. These facilities are:

- Uninterrupted Facility which includes:
    - Freeways
    - Multilane highways
    - Rural one/two/three lanes
    - Urban one/two/three lanes
- Interrupted Facility which includes:
    - Signal controlled facilities
    - Stop sign controlled facilities

Uninterrupted facility type of Kentucky provided high data coverage on freeways and urban road segments. In the beginning, the method was applied to the freeways with a speed data coverage of more than 75%. The required minimum sample rate for these segments was determined to be approximately 8%. Next, the same analysis was conducted for the remaining freeway segments with less than 75% coverage. Likewise, most of the segments showed a value of 8%. Certainly, 8% of the speed data were considered enough to be trusted for the freeways. Therefore, the recommended minimum sample rate for freeways is **8%**.

Following the similar procedure as mentioned above, multilane highways, rural highways (one/two/three lanes) and urban roads (one/two/three lanes) were analyzed. Finally, the minimum sample rates for those facility types were determined as below.

- Multilane highways: **5%**
- Rural one/two/three lanes: **9%**
- Urban one/two/three lane: **10%**

Results showed that most of the uninterrupted segments converge to a 10% minimum sample rate. Thus, this study recommends a threshold for minimum sample rate of 10% for this facility. A 10% sample is roughly equivalent to speed sample of 3 data within the 3-hour period each day. Note that if the speed data availability is greater than the minimum sample rate, the speed data are considered adequate. Based on this statement, the segments with adequate data can be identified from all the road segments of Kentucky. The observation of uninterrupted segments with adequate data over the state of Kentucky gives an idea of which road segments can be trusted in terms of measured speed data.

The total number of uninterrupted segments, including cardinal and non-cardinal direction, is 11,082. It was noticed that 99.5% of the freeway segments satisfied the data availability greater than the required 10% sample rate in terms of mileage. However, 55.3% of the rural highways satisfied the requirement. More than half of the total rural roads fulfilled the requirement due to the presence of low volume roads. Additionally, 91.1% of the total urban roads met the requirement. As a whole, the uninterrupted segments providing data availability greater than 10% are presented graphically in Figure 6. All the uninterrupted segments with adequate data are marked in green. The segments in red are not trustworthy in terms of measured speed since these have data availability of less than 10%.

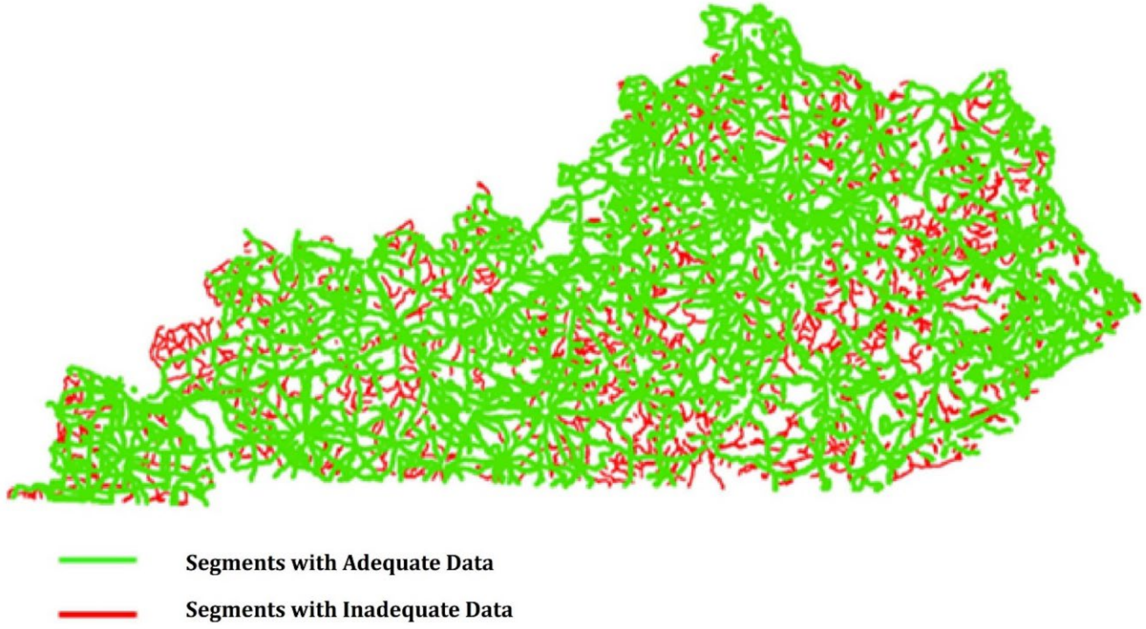


**Figure 6 Highlighted Green Routes Satisfying Minimum Sampling Rate Requirement for Uninterrupted Facilities**

The analysis also investigated the minimum sample rate for interrupted facility type using the bootstrapping approach. Since the interrupted facilities include signals and stop signs, results were analyzed for both. Finally, the minimum sample rates for these segments were determined as below.

- Signalized controlled: **10%**
- Stop sign controlled: **10%**

From the results, a threshold of 10% for minimum sample rate was recommended for the interrupted facility in Kentucky. Using this threshold value, the interrupted segments with adequate data over Kentucky were observed. The total number of interrupted segments is 11,146 including cardinal and non-cardinal direction. It was noticed that 92.6% of the total signalized segments satisfied the minimum requirement in terms of mileage. Contrarily, 44% of the total stop sign controlled segments had adequate data. A small portion of the stop sign segments fulfilled the requirement. It was due to very light traffic volume on this facility, which tended to have insufficient speed data. Compared to the uninterrupted facilities, the interrupted facilities resulted in a slightly large sample rate (10%). Reasonably, speed may be impeded by traffic control devices even when the intersections were operating at light traffic conditions. These control devices caused random variations in the speed data requiring more speed data as a minimum sample rate.

In summary, the interrupted road segments having adequate data are presented graphically in Figure 7. All the interrupted segments with adequate data are marked in cyan. Brown indicates the segments that are not trustworthy in terms of measured speed since they have data availability of less than 10%.

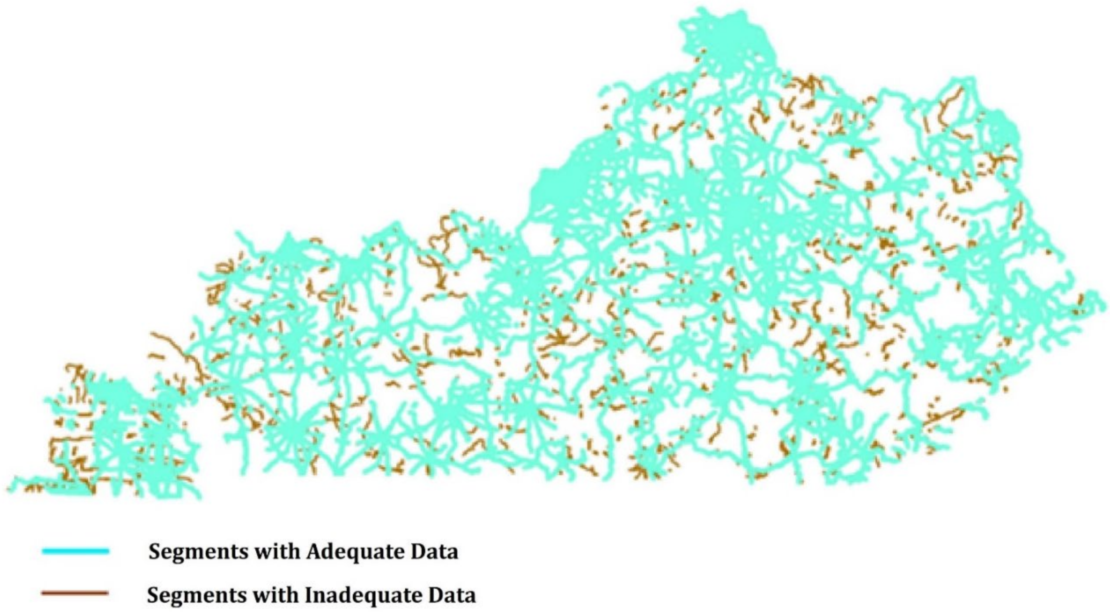


**Figure 7 Highlighted Cyan Routes Satisfying Minimum Sampling Rate Requirement for Interrupted Facilities.**

Although approximately 10% minimum sample rate is recommended both for the uninterrupted and interrupted facility after doing the analysis, Table 2 gives an overview of the minimum sample rate for individual facilities obtained from the bootstrap minimum sample rate method of this study.

**Table 2 Minimum Sample Rate Required**

| Highway Type | Minimum Sample Rate Required |
|---|---|
| Freeways | 8% |
| Multilane Highways | 5% |
| Rural One/Two/Three Lane | 9% |
| Urban One/Two/Three Lane | 10% |
| Stop Sign Controlled | 10% |
| Signalized Arterials | 10% |

This chapter shows a method to evaluate probe data quality using the bootstrapping approach. The estimated minimum sample rate is exploited to identify the segments with adequate data. The next step is to apply this and the measured speed data of each segment in investigating the factors affecting minimum sample rate as well as improving a traditional FFS model. In the next chapter, these two applications are discussed.

# Chapter 4 Applications

Previously, a method was developed to evaluate probe data quality. The method uses the bootstrapping approach to determine minimum sample rate of probe speed data for Kentucky. Later, this minimum sample rate is used to identify the segments with adequate data. This chapter shows the applications of these minimum sample rates and the speed data associated with those identified segments. Firstly, the factors that affect the minimum sample rate of uninterrupted and interrupted facility types will be identified using a regression model as a tool. The goal is to utilize the factors to have an idea about the minimum sample rate of one's own speed data before purchasing it from the data vendor. Lastly, the data, where deemed adequate, are applied to the calibration of the HERS-ST speed model.

## 4.1 Factors Affecting Minimum Sample Rate

This section attempts to identify and rank the significant factors for the minimum sample rate. This analysis intends to provide some general estimates on the data adequacy for given applications, which would be useful to agencies during the data acquisition process. A random forests regression model was developed to identify those factors along with their rankings. After that, the model, consisting of all the significant variables, was compared with two other models based on the goodness of fit.

All the steps involved with the random forests model development, identifying important variables, and comparison of the random forests model with other models are presented in the following sub-sections.

### 4.1.1 Model Description

Data mining procedure assists in learning and extracting information from data (*47*). One of the popular techniques in data mining is decision trees, which are also known as a classification and regression tree (CART). It does not require a functional form like statistical regression models. For better prediction accuracy, assemble of CART models is used. One of these assemble approaches is random forests (RF). This study uses this RF model as a tool to identify the factors influencing minimum sample rate.

RF is a non-parametric model used for exploring the non-linear relationship among the input variables. This model is made up of a number of decision trees which are built from several training samples randomly drawn from the original data with replacement. Observations selected out of the training samples are called testing data. Each tree provides a prediction result using the testing data. Finally, the prediction results from the trees are averaged. To avoid the correlation between individual trees, the RF model uses a subset of explanatory variables for splitting each node in each decision tree. The best split point is determined for each node in the tree by applying the splitting algorithm on the subset of the selected explanatory variables. The splitting algorithm produces maximum homogeneity to the successive node at a particular value of a selected variable.

An important feature of the RF model is Variable Importance (VI) to rank the explanatory variables. VI indicates the contribution of a variable to the output prediction when all other variables are present in the model. This study used Mean Decrease in Accuracy (MDA) method to measure the VI. MDA measures how much the model accuracy decreases when the testing data of each variable are permuted. If the variable is important, the model accuracy will be highly affected and decreases significantly after permutation. Then, the variables can be ranked according to the mean accuracy decrease. As the accuracy measure, mean squared error (MSE) is calculated for testing data using the following Equation (9).

$$MSE = \frac{1}{n} \sum_{\substack{i \in testing \\ data}}^{n} (y_i - \hat{y}_i)^2 \quad (9)$$

Where; $MSE$ = mean squared error using the testing data

$y_i$ = the observed value of the i<sup>th</sup> observation in the testing data

$\hat{y}_i$ = the predicted value of the i<sup>th</sup> observation in the testing data

$n$ = the number of observations in the testing data

For each explanatory variable, $MSE$ is calculated before and after permutation. The differences between before and after permutation $MSE$ are averaged over all the trees. Equation (10) shows the VI calculation of a variable based on the $MSE$ for testing data (*48*).

$$VI_j = \frac{1}{n_{tree}} \sum_{tj=1}^{n_{tree}} (EP_{tj} - E_{tj}) \quad (10)$$

Where; $n_{tree}$= the number of trees in the forest

$E_{tj}$= the $MSE$ on tree $t$ before permuting the values of variable $X_j$

$EP_{tj}$= the $MSE$ on tree $t$ after permuting the values of variable $X_j$

$VI_j$= VI for the variable $X_j$

Equation (10) implies that the larger the difference between the $MSE$ values, the more importance is given to that particular variable.

This study uses the RF model to identify the factors that have a significant influence on the minimum sample rate of a facility type. The reason for choosing RF model is that RF requires no explicit functional form, is well suited to the highly collinear data sets with a large number of explanatory variables, and does not assume a linear relationship between explanatory variables (*49-52*) and correlated explanatory variables (*52, 53*). Moreover, it can rank the explanatory variables unlike other "black box" models such as Neural Network.

### 4.1.2 Variable Importance

#### 4.1.2.1 Data and Preliminary Analysis

To identify the factors of the minimum sample rate, segments with adequate data were utilized. RF model was developed using the attributes of these segments as input variables. The attributes related to these segments were collected from KYTC's HIS database, including road geometry, accessibility, and mobility conditions, which are shown in Table 1 of section 3.1. During the analysis, it was also assumed that geometric condition remained constant over the year. Based on a preliminary analysis, the variables that would be considered in the RF model were selected. Table 3 shows the descriptive statistics of those selected variables.

Table 3 contains numerical variables where mean, standard deviation (SD), and minimum and maximum values for the numerical variables are presented. For categorical variables, only maximum and minimum categories are presented. It is noted here that the response variable is the minimum sample rate, and explanatory variables are the attributes related to the geometric conditions, accessibility, and mobility.

During the preliminary analysis, correlation plots were generated for the explanatory variables and the response variable. It was found that the relation between the explanatory variables and response variables was different for the uninterrupted facility and interrupted facility types. For example, minimum sample rate of the interrupted facility showed correlation with signal density, whereas uninterrupted facility did not. Hence, two separate RF models were developed for these facilities. Table 3 shows the variables that were considered for the two facility types separately. These variables were collected for a total of 7,117 uninterrupted segments and 7,594 interrupted segments with adequate data.

## Table 3 Descriptive Statistics

| | | | Uninterrupted Facility | | | | Interrupted Facility | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Total = 7,117 | | | | Total =7,594 | | | |
| | | | Measures | | | | Measures | | | |
| | Variables | Unit | Mean | SD | Min | Max | Mean | SD | Min | Max |
| **Response Variable** | Minimum Sample Rate | % | 0.73 | 0.65 | 0.05 | 9.41 | 1.84 | 1.13 | 0.12 | 10 |
| **Input Variables for Regression Models** | Section Length | miles | 3.63 | 3.15 | 0.01 | 21.84 | 1.02 | 0.95 | 0.01 | 15.85 |
| | Pavement Type* | | | | 1 | 8 | | | 1 | 7 |
| | Signal Density | no per miles | | | | | 2.35 | 3.43 | 0.00 | 40.82 |
| | Access Point Density | no per miles | 3.16 | 5.14 | 0.00 | 153.85 | | | | |
| | Intersection Density | no per miles | | | | | 9.45 | 8.36 | 0.15 | 166.67 |
| | Sign Density | no per miles | | | | | 1.62 | 4.66 | 0.00 | 125 |
| | Through Lanes* | no of lanes | | | 1 | 12 | | | 1 | 8 |
| | Access Control Type* | | | | 1 | 3 | | | 1 | 3 |
| | Terrain Type* | | | | 1 | 3 | | | 1 | 3 |
| | AADT | vehicles per day | 8,097 | 18,917 | 22 | 1,97,407 | 8,015 | 8,201 | 20 | 73,955 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Peak Lanes* | no of lanes | | | 1 | 6 | | | 1 | 5 |
| | Lane Width | ft | 10.57 | 1.55 | 6.00 | 32.00 | 11.09 | 1.81 | 6.00 | 32.00 |
| | Right Shoulder Width | ft | 5.14 | 3.41 | 0.00 | 18.00 | 3.34 | 3.31 | 0.00 | 14.00 |
| | Speed Limit | mph | 55 | 10 | 15 | 70 | 40 | 10 | 10 | 65 |
| | Volume to Service Flow Ratio (VSF) | | 0.18 | 0.16 | 0.00 | 1.34 | 0.34 | 0.24 | 0.00 | 2.46 |
| | Functional Class (FC)* | | | | 1 | 19 | | | 2 | 19 |

**The sign represents the categorical variable used in the regression models.*

For the uninterrupted facility, Speed Limit, Functional Class (FC), AADT, Section Length, Access Point Density, Lane Width, etc. were considered in the analysis. Note that Access Point Density is defined as the number of access points per length of a segment. These access points can be controlled or uncontrolled.

For the interrupted facility, Intersection Density, Signal Density, Sign Density, AADT, FC, Section Length, etc. were considered. Note that Intersection Density is defined as the number of junctions per length, where the junctions can be signal/stop controlled or uncontrolled.

The next section will rank and prioritized the variables mentioned above using the RF model.

#### 4.1.2.2 Model Calibration and Variable Importance

Before identifying the factors, the RF model required tuning of hyper-parameters for obtaining good prediction accuracy. From the literature (*49, 50, 54*), these hyper-parameters are:

- Number of trees in the forest ($n_{tree}$)
- Number of variables selected at each node for splitting ($m_{try}$)

Studies (*49, 55*) indicated that a large number of trees ($n_{tree}$) in a RF model would achieve more stable prediction performance. Saha et al. (*56*) tried 500, 1000, 5000, and 10,000 as the values for $n_{tree}$. This study adopted these values in order to tune $n_{tree}$. For $m_{try}$, Breiman (*49*) suggested three trials in RF regression model. According to his suggestion, the recommended trials are made as p/3, half of p/3 and twice of p/3 for $m_{try}$, where p is the total number of explanatory variables from the dataset. In this study, p = 13 for the uninterrupted facility, and 15 for the interrupted facility were considered.

The best combination of the two hyper-parameters was obtained by using Python package 'RandomizedSearchCV', which automates the whole process of searching the best combination incorporating cross-validation (CV). 'RandomizedSearchCV' built a total of 12 models from all the pairs of $n_{tree}$ and $m_{try}$ for each facility type. All of the 12 models were evaluated by CV. This study used a 10-fold CV to evaluate each model and control overfitting in the models. The 10-fold cross validation split the data into 10 stratified parts as shown in Figure 8. Each part successively was used as a testing data for estimating prediction performance. The remaining data was used as a training set. $MSE$ was calculated for each of the 10 folds and was averaged over the 10 folds (Figure 8). This 10-fold CV was performed for each of the 12 models and average $MSE$ was obtained for each model. Finally, the best combination of $n_{tree}$ and $m_{try}$ was reported from the model that estimated lowest $MSE$.

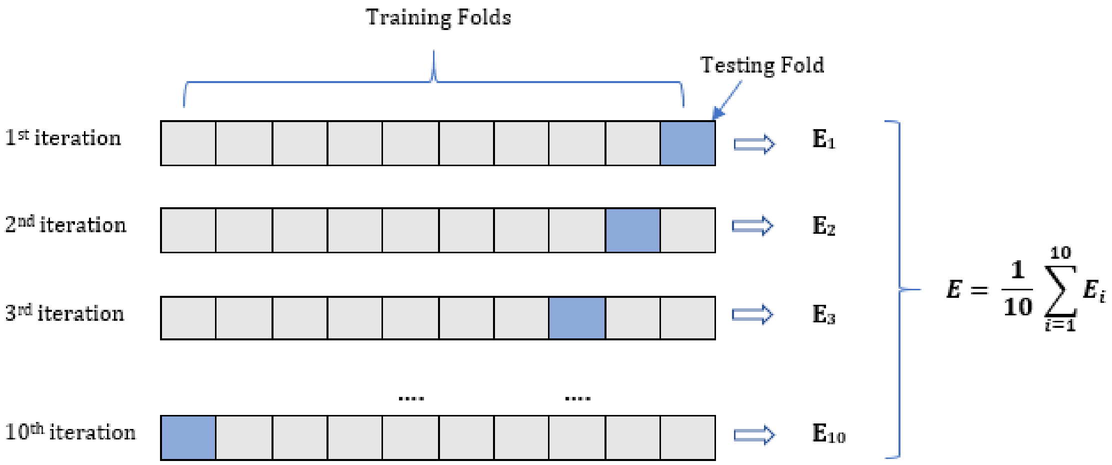


**Figure 8 10-fold Cross-Validation**

The best combination of hyper-parameters for both facilities was estimated as $n_{tree}$=10,000, and $m_{try}=\frac{1}{2} \times \frac{p}{3}$. The $n_{tree}$ value was found consistent with Saha et al. (*56*), where the authors mentioned that an assemble of 10, 000 trees is considered suitable for stable prediction from the RF model. The next step is to measure VI from the RF that was built using this combination of hyper-parameters.

To obtain VI, the average increase in $MSE$ (IncMSE) was calculated while permuting a variable. During the analysis, the variables with a VI greater than zero were kept in the RF model and others were eliminated (*55*). For example, Figure 9 shows the VI after running the RF model for interrupted facility. It presents that Through Lanes and Access Control Type have VI below zero. These variables were excluded from the model. RF model was run again keeping the variables with VI greater than zero, and this elimination process repeated until all the remaining variables in the model had a VI greater than zero.

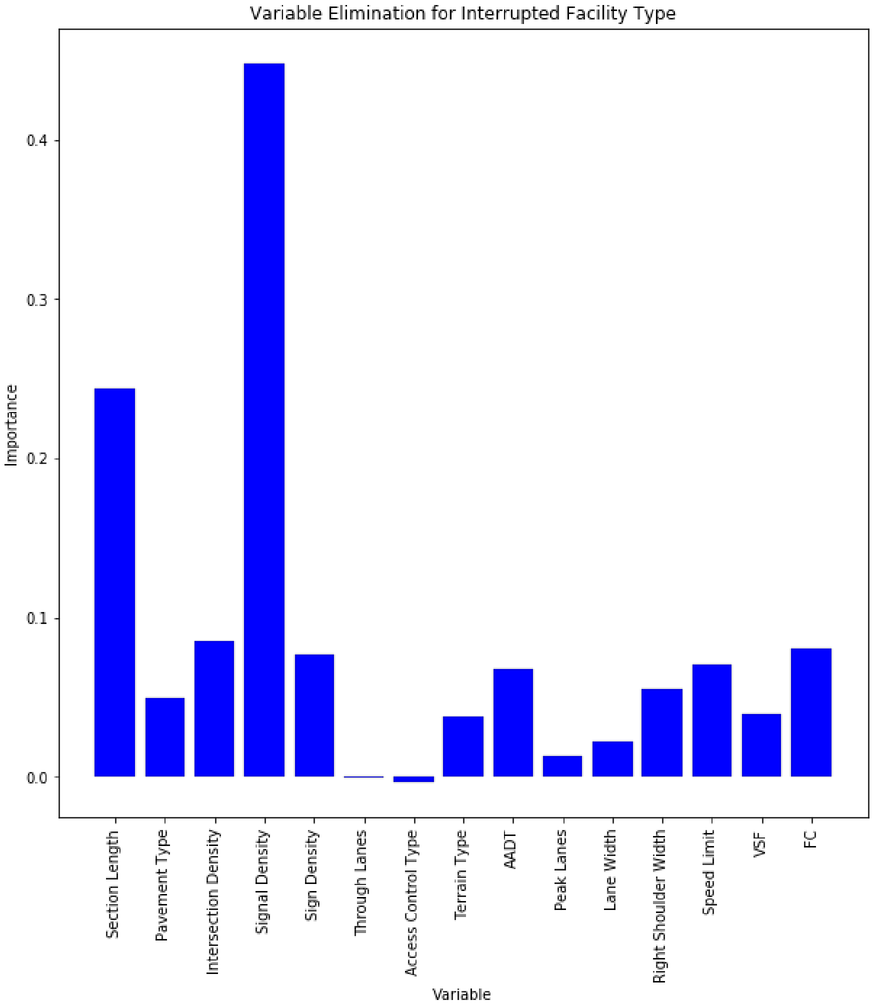


**Figure 9 Elimination Stage for the Variables of Interrupted Facility Type**

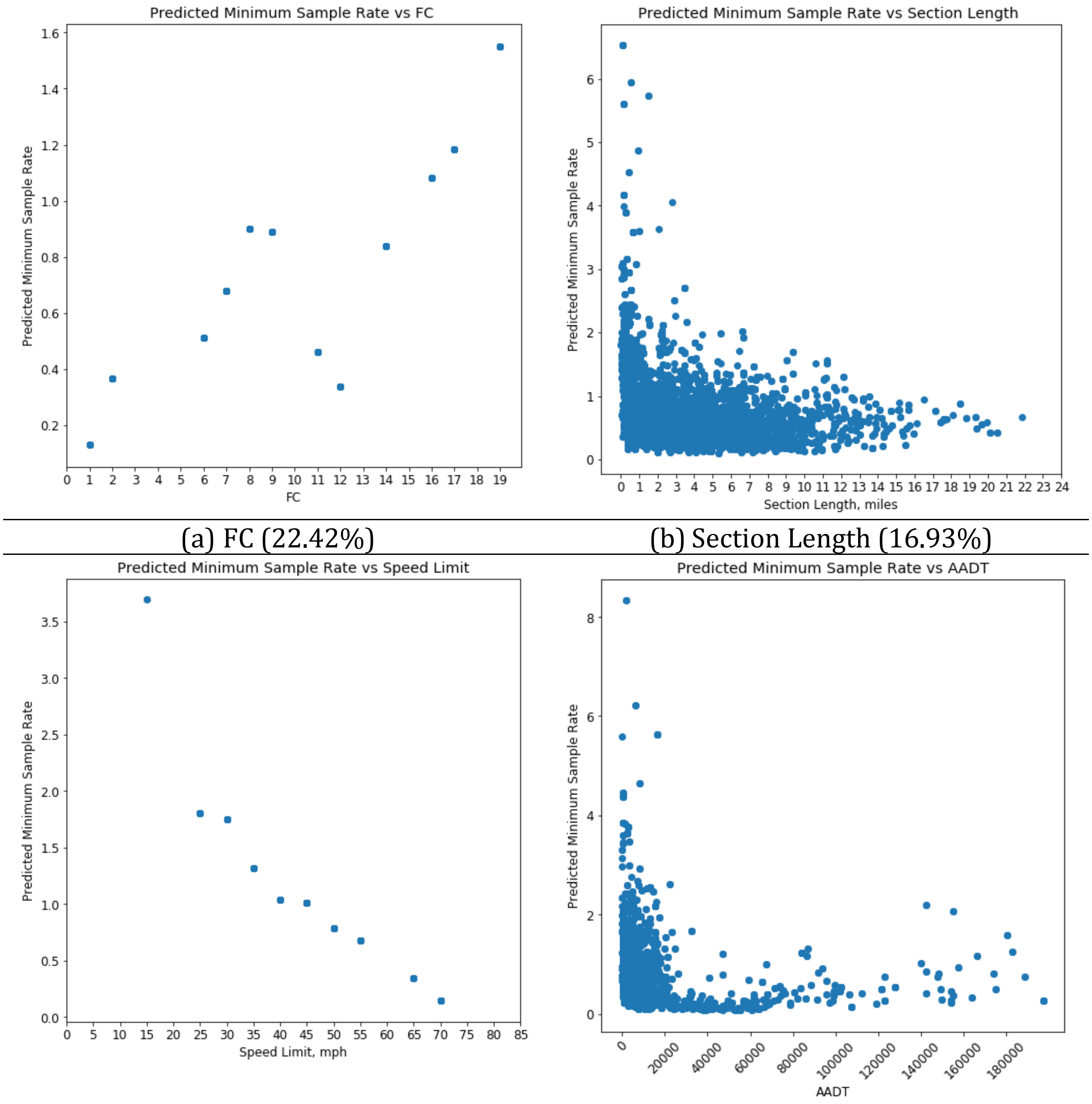
Two separate RF models were built for the uninterrupted and interrupted facility types. These models contained the significant variables based on VI. The results from VI are presented below for both facility types.

For uninterrupted facility type, the important variables are shown in Table 4. FC is the top-ranked variable. From Figure 10(a), it seems that higher FC such as FC1, FC2, FC11, and FC12 require a smaller sample rate. Conversely, lower FC roads require a larger sample rate. The second variable is Section Length. It appears in

Figure 10(b) that the longer section requires smaller sample rates compared to the shorter section. Speed Limit is the third variable according to VI. From Figure 10(c), segments with higher Speed Limit require a smaller sample rate and vice versa. Since Speed Limit varies for different FC road segments, it contributes to the minimum sample rate of a segment. AADT contributes as the fourth important variable. It seems from Figure 10(d) that segments with higher AADT, for example; interstates, require smaller sample rates. Alternatively, low AADT roads appear to need larger sample rates. Access Point Density contributes as the fifth important variable. Access points, with or without traffic control devices, add random fluctuation in the speed pattern. Hence, segments may require a larger sample size with increasing Access Point Density. Other variables like VSF, Lane Width, Peak Lanes, etc. were also found important for uninterrupted facility type.

**Table 4 Variable Ranking for Uninterrupted Facility Type**

| Variables | IncMSE | VI (%) | Rank |
|---|---:|---:|---:|
| FC | 0.360 | 22.42 | 1 |
| Section Length | 0.272 | 16.93 | 2 |
| Speed Limit | 0.178 | 11.11 | 3 |
| AADT | 0.151 | 9.39 | 4 |
| Access Point Density | 0.148 | 9.19 | 5 |
| Right Shoulder Width | 0.138 | 8.62 | 6 |
| VSF | 0.113 | 7.06 | 7 |
| Lane Width | 0.087 | 5.44 | 8 |
| Terrain Type | 0.051 | 3.15 | 9 |
| Access Control Type | 0.050 | 2.95 | 10 |
| Peak Lanes | 0.044 | 2.75 | 11 |
| Pavement Type | 0.010 | 0.68 | 12 |
| Through Lanes | 0.005 | 0.32 | 13 |



(a) FC (22.42%)

(b) Section Length (16.93%)

(c) Speed Limit (11.11%)

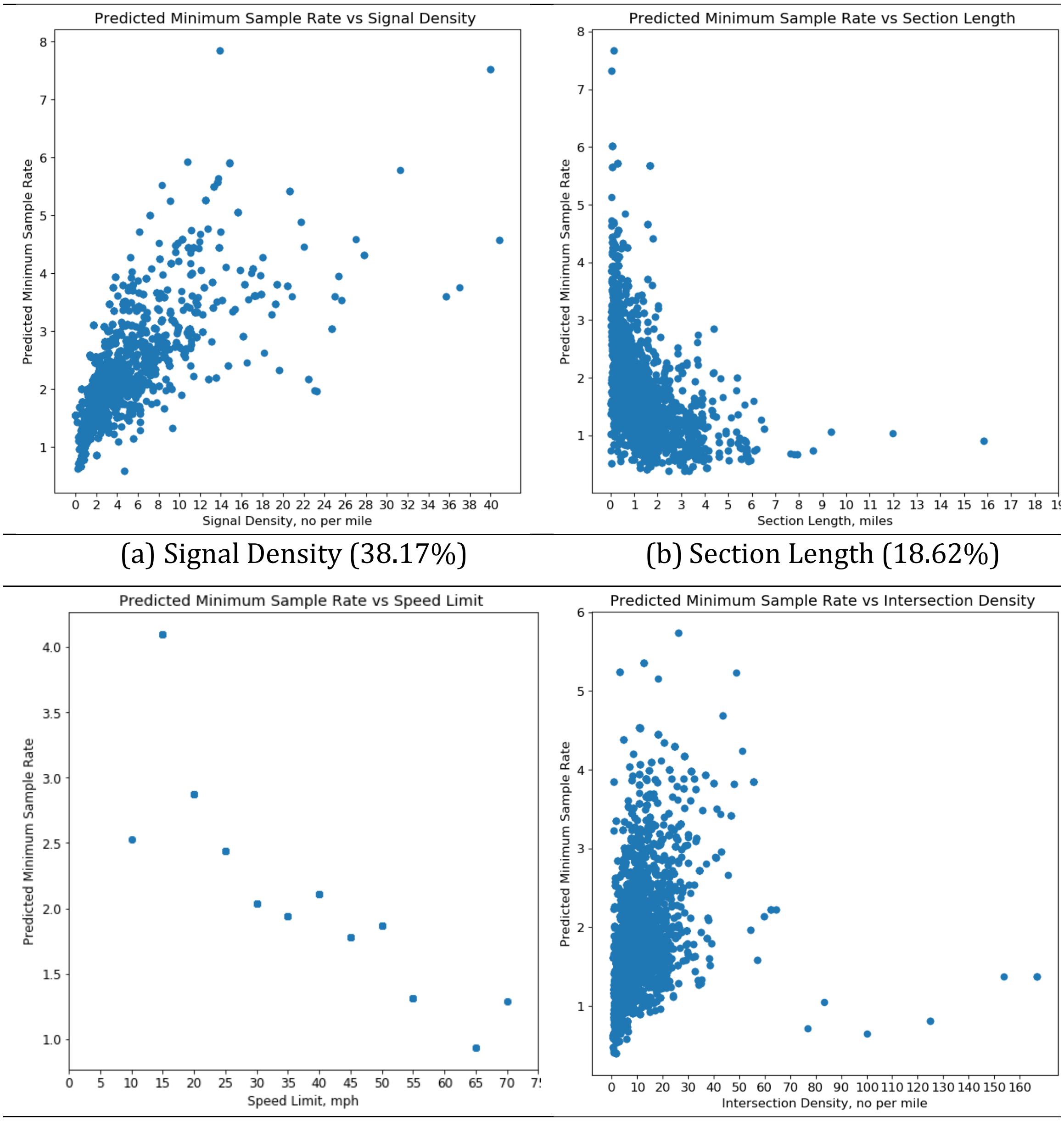
(d) AADT (9.39%)

**Figure 10 Individual Variable's Effect on Minimum Sample Rate of Uninterrupted Facility from RF model**

For interrupted facility type, the important variables from the RF model are presented in Table 5. The top-ranked variable is Signal Density for this facility. It tends to influence the minimum sample rate positively from Figure 11(a). The requirement of minimum sample rate increases with the increase in Signal Density with some deviations. This finding also agrees with the analysis results from Eshragh et al. (1), where Signal Density was one of the contributing factors affecting the accuracy of probe data. The second most important variable is Section Length. Although most of the interrupted facilities are not very long, it seems that the minimum sample rate is decreasing with an increase in Section Length according to Figure 11(b). The third variable is Speed Limit, which tends to affect the minimum sample rate negatively from Figure 11(c). Segments with higher Speed Limit seem to require fewer samples compared to the lower Speed Limit roads. The fourth variable is Intersection Density. Seemingly, an increase in Intersection Density involves higher sample rates and vice versa. Other variables such as FC, Sign Density, AADT, VSF, etc. were also found significant for interrupted facility type.

**Table 5 Variable Ranking for Interrupted Facility Type**

| Variables | InMSE | VI (%) | Rank |
|---|---:|---:|---:|
| Signal Density | 0.561 | 38.17 | 1 |
| Section Length | 0.274 | 18.62 | 2 |
| Speed Limit | 0.110 | 7.45 | 3 |
| Intersection Density | 0.091 | 6.20 | 4 |
| FC | 0.085 | 5.78 | 5 |
| Pavement Type | 0.077 | 5.23 | 6 |
| Sign Density | 0.070 | 4.71 | 7 |
| AADT | 0.057 | 3.87 | 8 |
| Lane Width | 0.034 | 2.31 | 9 |
| Right Shoulder Width | 0.032 | 2.20 | 10 |
| VSF | 0.028 | 1.89 | 11 |
| Peak Lanes | 0.027 | 1.85 | 12 |
| Terrain Type | 0.025 | 1.72 | 13 |



(a) Signal Density (38.17%)

(b) Section Length (18.62%)

(c) Speed Limit (7.45%)

(d) Intersection Density (6.20%)

**Figure 11 Individual Variable's Effect on Minimum Sample Rate of Interrupted Facility from RF model**

The RF model gave the list of significant variables for both facility types. However, this list is long since it enlisted 13 variables for both facilities. The longer the variable list, the costlier the data collection. To ease the data collection, this study decided to prioritize the variables for both facilities. The prioritization will make the variable list shorter, minimizing data collection effort and cost while confirming the accuracy of the RF model.

To prioritize variables both for the uninterrupted and interrupted facilities, two measures were used in this study for predicting error on testing data. These measures are Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE). RMSE is a measure of the differences between predicted values ($\hat{Y}_i$) of a model and the observed values ($Y_i$ ).

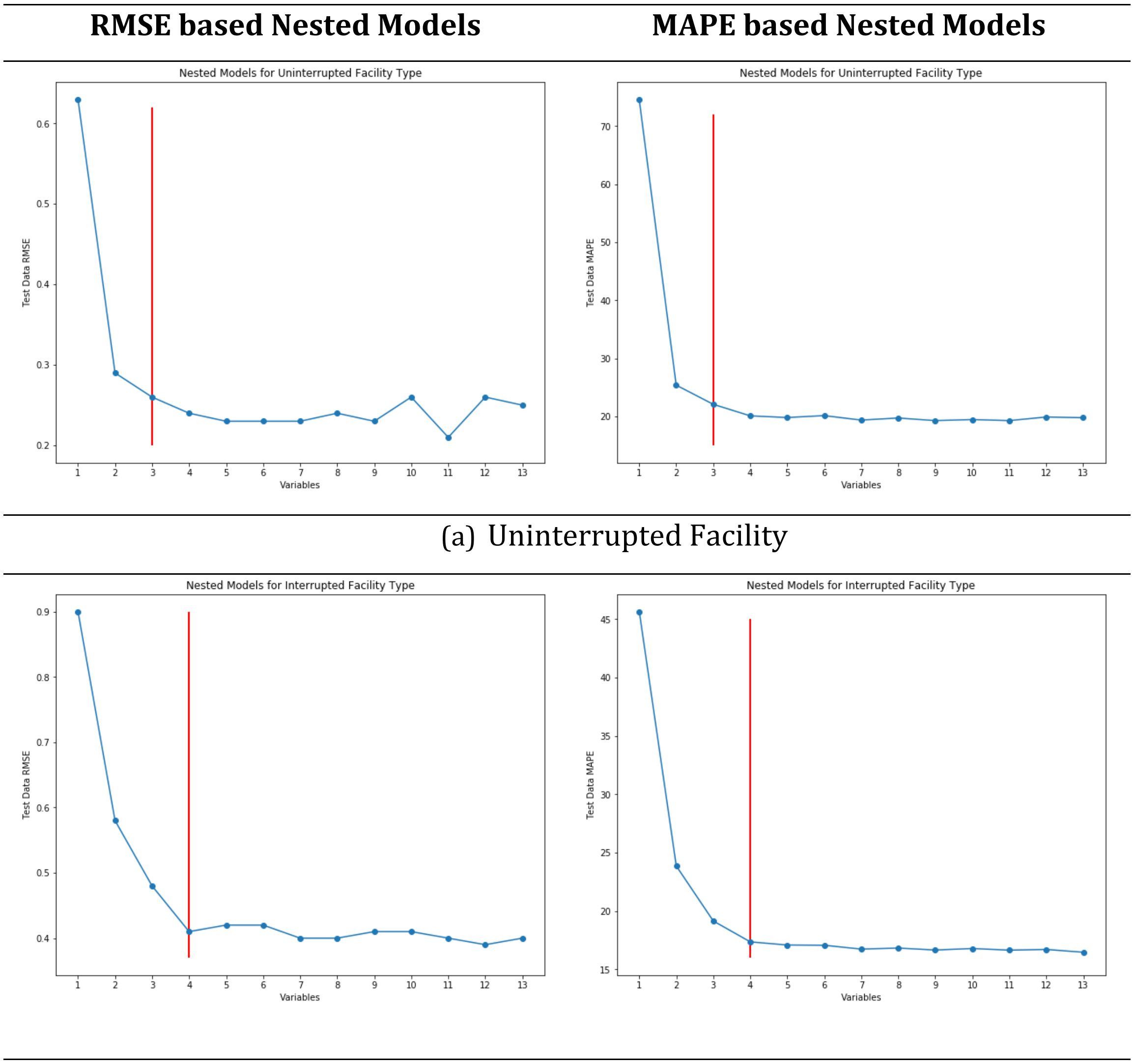
$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2} \quad (11)$$

MAPE is a measure of prediction accuracy which estimates the mean or average of the absolute percentage errors of prediction. Here, the error is defined as the difference between actual value ($Y_i$ ) and predicted value ($\hat{Y}_i$).

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\frac{\left|Y_i - \hat{Y}_i\right|}{Y_i} \quad (12)$$

These two measures helped in restricting the variable list to overcome data collection complexity. A final RF model was built using those restricted variables for both facility types. To track the decrease in these two measures for each variable, a nested collection of RF models was constructed. The nested models started from the one with top-ranked variable and ended with the one involving all important variables that were kept in the previously built RF models for each facility type. For example, 13 variables were found for the uninterrupted facility, where FC was the number one variable, Section Length was the number two variable, etc. The first nested model would contain only FC, the second nested model would contain FC and Section Length and so on. Finally, the last nested model would contain all the 13 variables. For each nested model, RMSE and MAPE were reported for testing data. The set of the variables that led to a significant decrease in RMSE and MAPE of the model was finalized.

RMSE and MAPE based nested models are shown in Figure 12. For the uninterrupted facility, there is a significant decrease in RMSE and MAPE after adding the top 3 variables as shown in Figure 12(a). The remaining variables with low ranks do not contribute significantly to the minimization of RMSE and MAPE. It is wise to exclude them from the final model. Consequently, the final model for uninterrupted facility contains 3 variables. This model is named as RF_Uninterrupted. For the interrupted facility, a combination of top 4 variables shows a significant drop in RMSE and MAPE in Figure 12 (b). Thus, these four variables are finalized, and the final model is named as RF_Interrupted for the interrupted facility.



**RMSE based Nested Models** | **MAPE based Nested Models**

(a) Uninterrupted Facility

(b) Interrupted Facility

**Figure 12 Nested RF Models for Uninterrupted and Interrupted Facilities**

To summarize the results from the above variable prioritization process, the final RF models containing the significant variables for each facility type are presented in Table 6.

**Table 6 Significant Variables for Uninterrupted and Interrupted Facilities**

| **Facility Type** | **Name of the RF Model (in this study)** | **Total Variables** | **Significant Variables** |
|---|---|---|---|
| Uninterrupted Facility | RF_Uninterrupted | 3 | FC, Section Length, and Speed Limit |
| Interrupted Facility | RF_Interrupted | 4 | Signal Density, Section Length, Speed Limit, and Intersection Density |

To observe the RF models' performance results using the listed variables in Table 6, the next sub-section discusses the comparison among the RF model, linear regression model, and neural network model.

### 4.1.3 Results

RF models for both facility types were compared with the neural network (NN) model and liner regression model in this study. To compare the models, MAPE and RMSE were used as the Measures of Effectiveness (MOEs). Smaller values of RMSE and MAPE indicate the better performance of a model.

The linear regression model was used to estimate the impacts of explanatory variables on the minimum sample rate. NN model was also used for the same purpose. For NN, Python package named *Multi-Layer Perceptron Regressor* (MLP) was used, which optimizes the squared-loss using LBFGS or stochastic gradient descent. Parameter tuning for NN was done based on 'GridSearchCV' and using 10-fold CV which follows the same mechanism as discussed for RF model earlier.

From Table 7, for both facility types, linear regression models lead to the largest error showing lower prediction performance. NN model also performs badly, which shows error measures closer to the linear regression model for both facilities. Clearly, the RF model outperformed both NN and liner regression model in terms of RMSE and MAPE.

**Table 7 Predictive Performance Evaluation Table**

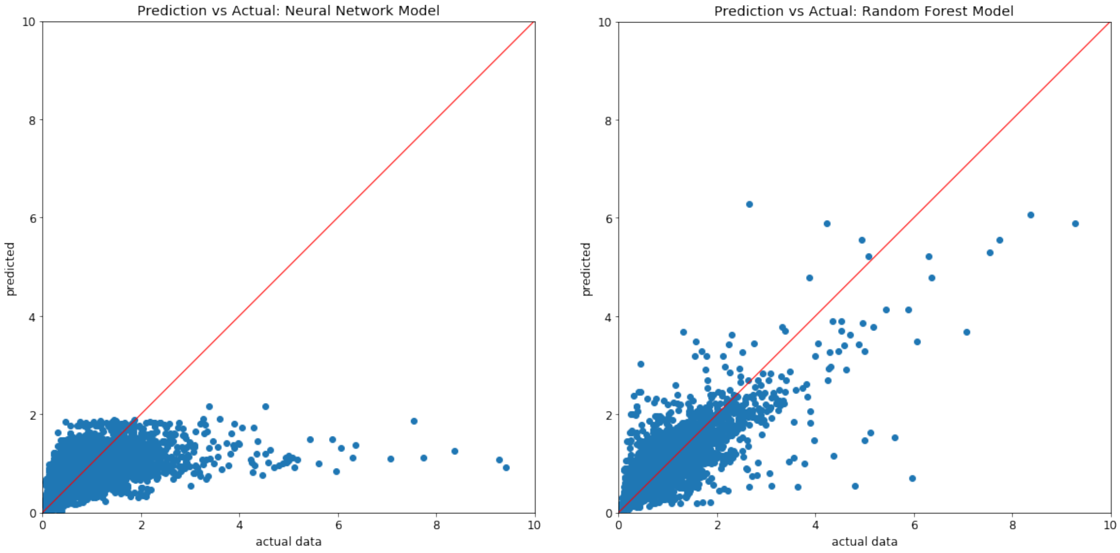
**(a) Models for Uninterrupted Facility**

| **Models for Uninterrupted** | **RMSE** | **MAPE (%)** |
|---|---|---|
| RF_Uninterrupted | 0.31 | 26.88 |
| NN Model | 0.50 | 53.26 |
| Linear Regression Model | 0.57 | 64.09 |

**(b) Models for Interrupted Facility**

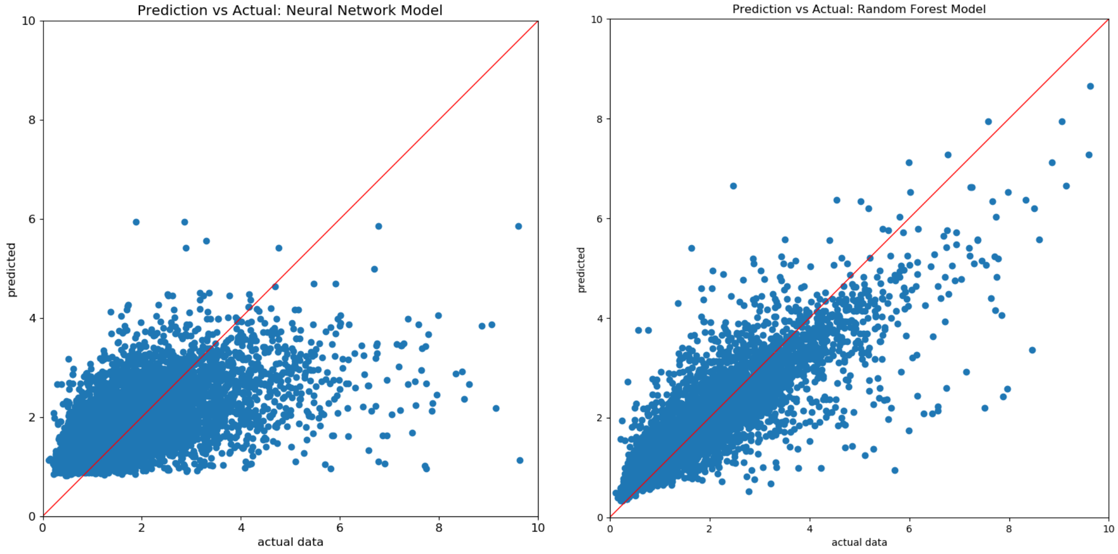
| **Models for Interrupted** | **RMSE** | **MAPE (%)** |
|---|---|---|
| RF_Interrupted | 0.59 | 22.97 |
| NN Model | 0.85 | 43.29 |
| Linear Regression Model | 0.98 | 48.55 |

Figure 13 and Figure 14 show the comparison between the predicted and observed minimum sample rate from the NN models and the RF models for both facility types. In Figure 13 and Figure 14, the more data in the diagonal line, the better the prediction performance of the models. Undoubtedly, RF models perform better for both facilities. Therefore, based on the MOEs and prediction performance, RF model is recommended for predicting minimum sample rate of a road segment.



NN RF_Uninterrupted

**Figure 13 Comparison between Model Predictions and Actual Minimum Sample Rates for Uninterrupted Facility**



NN RF _Interrupted

**Figure 14 Comparison between Model Predictions and Actual Minimum Sample Rates for Interrupted Facility**

This section presents a practical application of the bootstrap sampling results to identify the variables affecting the minimum sample rate both for uninterrupted and interrupted facility types. The RF regression model with variable ranking was used to identify the variables. After knowing the important variables, those can be used in the RF regression model for estimating the minimum sample rate of a new segment, where the facility type is given. Moreover, variable ranking gives the list of variables regarding the data acquisition for sampling analysis. Consequently, the RF model can be a substitute for the bootstrapping approach of determining the minimum sample rate. In the future, the minimum sample rate can be determined for a new segment, once the required dataset containing the important variables is collected.

## 4.2 Improving the Estimation of Free Flow Speed

This section demonstrates an application of the speed data acquired from the segments with adequate data, specifically on improving the method of estimating FFS.

Traditional models built on Highway Capacity Manual (HCM) method are usually used for average speed estimation. One of these models is the Highway Economic Requirements System-State Version (HERS-ST) model. This study mainly focuses on the FFS estimation step using HERS-ST approach. FFS is defined as the speed when traffic is light and vehicular speed is restricted by geometric condition and traffic control devices, but not by the presence of other vehicles. The general framework to calculate FFS in HERS-ST model (*57*) is shown in Figure 15.

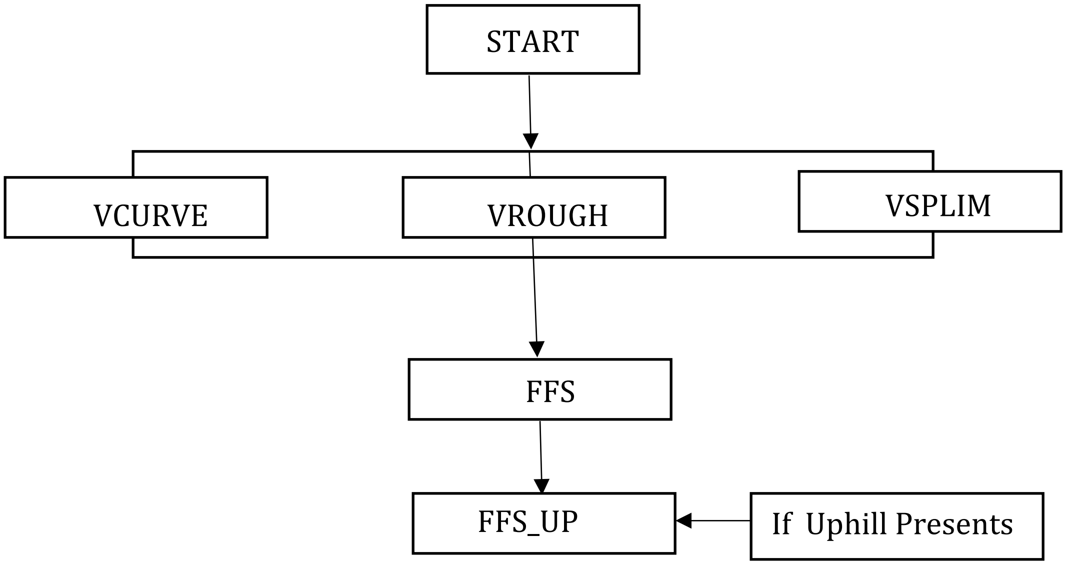


**Figure 15 Framework for HERS FFS Estimation**

From HIS data listed in Table 1 from section 3.1, measured pavement roughness (IRI or PSR), grade, and curve lengths are required to calculate FFS using HERS-ST. The FFS is determined using the following three inputs:

- The maximum allowable speed on a curve (VCURVE)
- The maximum allowable ride-severity speed (VROUGH)
- The maximum speed resulting from speed limit (VSPLIM)

Equation (13) demonstrates the FFS calculation for **Error! Bookmark not defined.**HERS-ST.

$$FFS = \frac{e^{\sigma^2/2}}{(VCURVE^{-1/\beta} + VROUGH^{-1/\beta} + VSPLIM^{-1/\beta})^{\beta}} \quad (13)$$

The recommended values of the model parameters are $\sigma = 0.1$ and $\beta = 0.1$ for all types of facility without accounting for the challenges that may arise based on facility type.

In **Error! Bookmark not defined.Error! Bookmark not defined.Error! Bookmark not defined.**Equation (13 ), VCURVE represents the effect of curves on vehicle speed. It is related to the maximum perceived friction ratio, super-elevation, and degrees of curvature. Friction ratio values are set in accordance with vehicle types. If a section has no curves, the VCURVE does not influence the FFS. The overall effect of curves in a section is the weighted average effect on different vehicle classes. The equation is listed below in miles per hour.

$$VCURVE = 292.5 \times \sqrt{(FRATIO + SP)/DC} \qquad (14)$$

Where; FRATIO = maximum perceived friction ratio

0.103 for combination trucks;

0.155 for automobiles; and

0.155 for single-unit trucks

DC = degrees of curvature

SP = super elevation

0 if DC<=1;

0.1 if DC>=10; and

$0.0318 + 0.0972 \times \ln(DC) - 0.0317 \times DC + 0.007 \times DC \times \ln(DC)$; otherwise

VROUGH represents the effect of pavement roughness on speed. HERS speed model uses pavement serviceability rating (PSR) to measure pavement roughness. VROUGH's value is determined by the following formulas:

$$VROUGH = 5 + 15 \times PSR \qquad \text{if PSR<=1.0}$$

$$VROUGH = 20 + 32.5 \times (PSR - 1.0) \qquad \text{if PSR>1.0} \qquad (15)$$

The effect of speed limits on vehicle speeds is represented by VSPLIM. The operational speed is assumed to be 9.323 mph greater than the posted speed limit for urban freeways and rural multilane highways with partial or full access control and a median which is either a positive barrier or has a width of at least 4 feet. For all other roads, it is assumed to be 6.215 mph greater than the posted speed limit.

For those segments with a positive grade, the FFS should be adjusted to account for the impact of the grade. The delay due to grade, DGRADE, is determined based on vehicle characteristics and the average grade of a section. The HERS-ST speed model first estimates the crawl speed for a section and then calculates the delay due to grade for each vehicle type. Overall, delay due to grade is weighted by each vehicle type. HERS-ST speed model uses the following equation to calculate FFSUP which represents the free-flow speed on an uphill section.

$$FFSUP = \frac{1}{\frac{1}{FFS} + \frac{DGRADE}{SLEN}} \quad (16)$$

Where; DGRADE = delay in hours

SLEN = length of the section

### 4.2.1 Limitations in Existing HERS-FFS Estimation

To investigate the performance of existing HERS-FFS model, FFS generated by the model was compared with the reference speeds calculated based on the speed data from the segments providing adequate data. Note that reference speed is a threshold speed value, below which travel is considered as delayed. In this study, facility-specific reference speeds were used which were determined during the analysis. The reference speeds are listed here.

- For freeways and multilane highways: The 85th percentile speed of all speed data in a year represents reference speed.
- For other facilities: The 85th percentile speed from weekday daytime (6 am – 8 pm) speed data is used as reference speed.

A comparison of predicted FFS from existing HERS-ST and measured reference speeds is demonstrated in Figure 16. The figures show estimated HERS-FFS with the default value of the parameters σ and β. It is obvious that modeled FFS and measured reference speed do not fit well. It indicates that existing parameters not necessarily always produce good results both for uninterrupted and interrupted facility type. However, no step has yet been taken to adjust these parameters for different facilities. Moreover, the parameters were not estimated using enough speed data in the past (*57*). Therefore, calibration of the parameters σ and β is required. The calibration process is shown in the next sub-section.

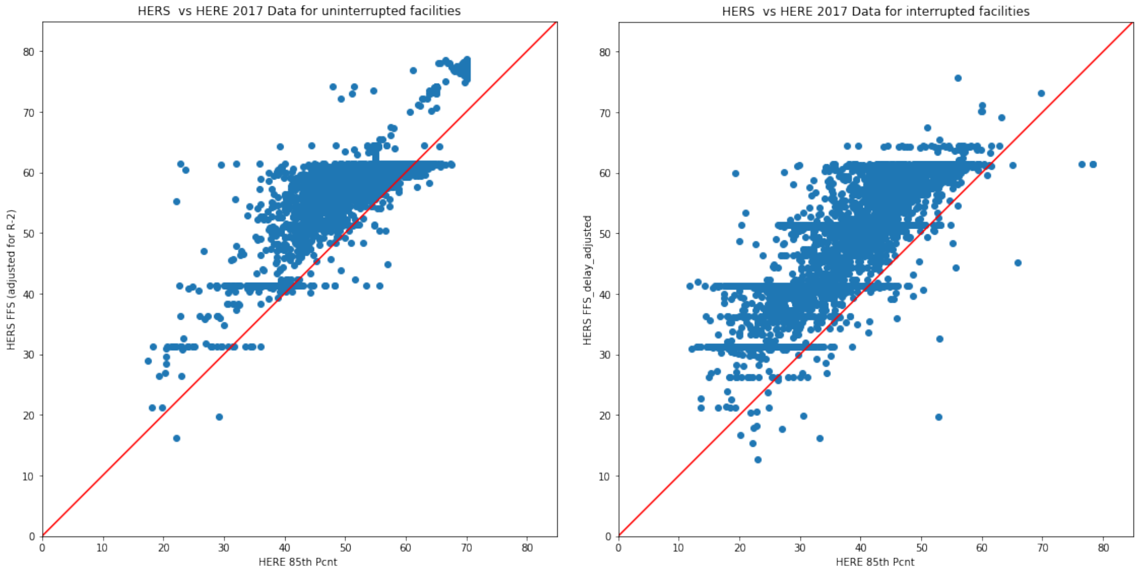


**Figure 16 Comparison of Existing HERS FFS with Measured Reference Speed**

### 4.2.2 Model Calibration

HERS-ST FFS model calibration was separately done for the interrupted and uninterrupted facilities using the measured speed data. To calibrate HERS-FFS model, reference speeds for each segment were compared with the modeled FFS.

The goal of the calibration was to find the values of $\sigma$ and $\beta$ that produce the best fit between the modeled FFS and measured reference speed. Hence, Equation 13 for FFS was calibrated by adopting a non-linear least squares fit method. This study used the Levenberg-Marquandt (LM) algorithm for non-linear least-squares optimization. The algorithm works by minimizing the squared residuals ($e$) defined for each data point as,

$$e^2 = (y - f(x))^2 \quad (17)$$

where; $y$ is the measured reference speed and $f(x)$ is the calculated FFS using HERS-FFS equation. The LM algorithm performs iteration and optimizes the solution that results in minimum residuals.

During the sensitivity analysis, it was observed that the HERS-ST model does not account for the impact of narrow lanes while estimating FFS for rural two-lane roads. HERS-FFS was significantly overestimating for those roads. Thus, it was required to adjust FFS using lane width adjustment factor based on HCM. Moreover, the model does not address the effect of traffic control devices for the interrupted facility including signal and stop sign controlled facilitates. This effect is calculated as “zero volume delay”. For the signal-controlled facility,

$$ZVDSIG = 0.0687\left(1 - e^{-NSIG/24.4}\right) \quad (18)$$

where; ZVDSIG is zero volume delay in hours per vehicle-mile traveled, while NSIG is the number of signals per mile.

For stop sign controlled facility,

$$ZVDSTP = NSTP(1.9 + 0.067FFS) \quad (19)$$

in which ZVDSTP is zero volume delay due to stop sign in hours per 1000 vehicle miles, and NSTP is the number of stop signs per mile. This is adapted from the HERS-ST speed model for stop sign controlled delay by setting the volume to zero. The adjusted FFS for the signal-controlled facility can be estimated as $1/(\frac{1}{FFS} + ZVDSIG)$ and that for the stop-controlled facility would be $1/(1/FFS + ZVDSTP/1000)$. No adjustment is needed for other facility types.

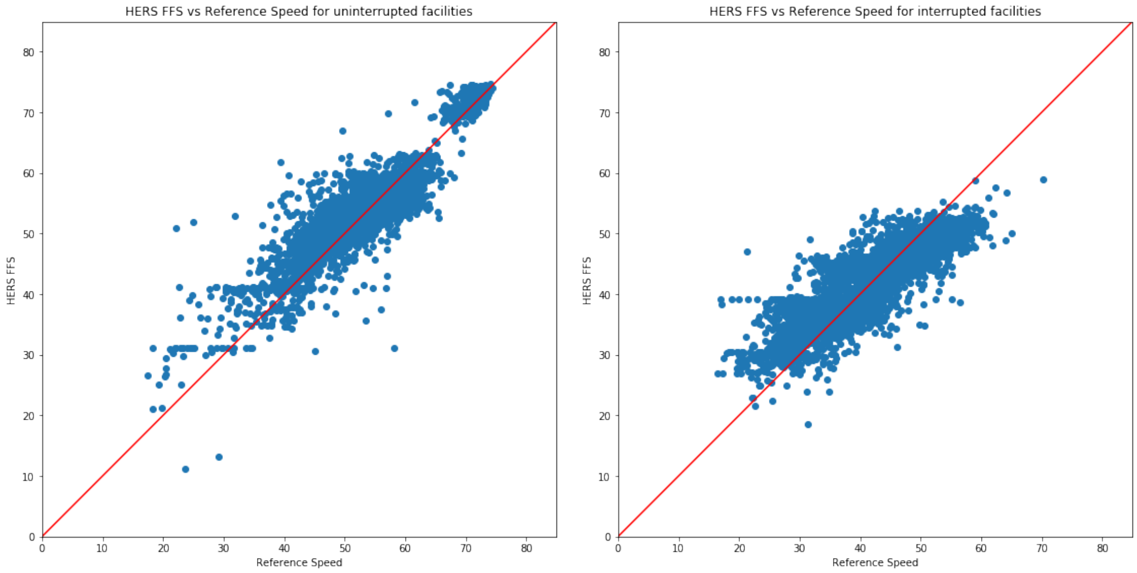
At first, the calibration was done separately for freeways, multilane highways, rural one/two/three, urban one/two/three, signals and stops. Although the calibration was done separately for each type of the uninterrupted facility, ultimately the calibration reflected the almost same values for the parameters after all the adjustments. Thus, the uninterrupted facilities were combined, and the suggested parameters for this category is presented in Table 8.

Afterward, the calibration was performed on the signalized arterials and stop sign controlled facilities separately. The FFS, after the calibration process, showed a good fit with the measured reference speed. Then, the signals and stops were combined into interrupted facility type due to having similar parameter values. Finally, the parameters were optimized as mentioned in Table 8.

**Table 8 FFS Calibration Results**

| Facility Types | $\sigma$ | $\beta$ |
|---|---|---|
| Uninterrupted Facility | 0.1427 | 0.2092 |
| Interrupted Facility | 0.3907 | 0.18378 |

Previously, HERS-ST technical report (*58*) derived a range for σ value between 0 and 0.19, and β value between 0.1 and 0.31 for all facility types. Although the calibrated parameter values for uninterrupted facility fall within these ranges, σ value for interrupted facility seems to cross the range. The overall calibration process utilized speed data from 80% of the total segments. To validate the calibrated parameters, the remaining 20% of the segments were used. The calibrated parameters performed well for those 20% segments indicating the credibility of those parameters for estimating FFS. Figure 17 combines these calibration and validation results and shows a better fit between measured reference speed and modeled FFS for both facility types.



**(a) Uninterrupted Facility** **(b) Interrupted Facility**

**Figure 17 Estimated FFS vs Reference Speed**

Although σ value for interrupted facility violates the recommended range, the calibrated values for σ and β can be accepted considering the validation results. Overall, the calibrated values are recommended for estimating FFS using HERS-ST.

This study applied the speed data of Kentucky road segments, where deemed adequate, to improve the performance of traditional FFS model. This application brought confidence over the traditional models. The models can be used for the segments with no speed data. Transportation agenesis can use the same approach of utilizing the speed data to have enhanced performance from the traditional models.

# Chapter 5 Conclusions

## 5.1 Summary

Probe speed data are widely used for estimating state-wide performance measures. The accuracy of these measures depends on adequate speed data. This study proposed a method to evaluate the quality of probe speed data. The method estimated minimum sample rate of speed data for a segment by adopting a bootstrapping approach without requiring an assumption about the underlying distribution of the population. It produced a predefined number of replications using the speed data, which were treated as a population. A tolerance limit of 5% was set as a convergence error for the sample mean of these replicated samples. The whole method was iterated over different sample rates until the error converged to the tolerance limit. The minimum sample rate used for the convergence into the tolerance limit was reported for each road segment. Using this method on the Kentucky based speed data from 2017, the minimum sample rates were obtained for all the segments. The results recommended a minimum sample rate of 10% for both uninterrupted and interrupted facility types in Kentucky.

The minimum sample rates resulted from the bootstrapping approach were compared with data availability to identify the segments with adequate data. A total number of 7,117 segments from uninterrupted and 7,594 segments from interrupted facilities in Kentucky were observed to satisfy the minimum sample rate requirement. In the case of uninterrupted facility, more than 90% of freeways, multilane highways, and urban roads have adequate speed data compared to the minimum sample rate. However, only half of the total rural roads have adequate speed data due to low traffic volume. Further, 92% of the signalized road segments have adequate speed data, whereas only 47% of the total stop sign controlled roads fulfill the requirement.

## 5.2 Applications

Using the minimum sample rates from the bootstrapping, factors affecting this were identified. The factors can provide a general estimate on the data adequacy for a particular application as well as help the agencies during data acquisition process.

RF regression model was used as a tool to identify the factors. After analyzing VI from the model, FC, Section Length, and Speed Limit were found to be the important variables for uninterrupted facility type. Conversely, for interrupted facility, Signal Density, Section Length, Speed Limit, and Intersection Density were observed to be the significant variables. In addition, the RF model outperformed NN and liner regression models for both cases. Therefore, it was recommended to determine the minimum sample rate of a new segment. If one wants to have an idea on the data collection before purchasing from data vendors, they might adopt this model to know the required minimum speed data for their applications.

Speed data of the identified segments were used to improve the performance of the traditional FFS model. Previous research demonstrated the performance of the model using inadequate data (*57*). The existing parameters of the model were also validated using an inadequate dataset. The model may not always produce a good estimate of the FFS using the default parameters. That is why this study decided to calibrate the parameters of the FFS model using actual data with adequacy. During the calibration process, it was also observed that the traditional model is quite sensitive to the lane width and traffic control devices. The adequate speed data used in this study addressed these limitations and helped to calibrate the parameters to improve model performance. It brought more confidence in using traditional models by transportation agencies.

The findings of this study helped to identify the road sections having good coverage of speed data using the required minimum sample rate. Moreover, to obtain reliable congestion measures for the road segments and to improve transportation models, the minimum sample rate is a decision parameter which examines the data quality. After knowing that the availability is greater than the minimum required sample rate, FFS for a specific facility can be determined directly using the measured speed data collected over the year. Furthermore, the minimum sample rate gives an idea of the variation of travel time on a specific corridor. For example, a larger sample rate indicates unstable travel time pattern and vice versa.

### 5.3 Limitations and Future Work

In this study, the bootstrapping approach produced replications from the available measured speed data, considering the dataset as a population. Most of the freeway segments had speed data availability of more than 90%. These speed data, used in bootstrapping replications, are considered as a close approximation of the true population. However, 71% of the rural two-lane and stop sign controlled segments had speed data availability below 30%. This study excluded those segments while performing bootstrapping on the dataset. The reason is that the data associated with those segments only correspond to a subset of the true population and may not represent the true population as well as may produce biased results during the minimum sample rate estimation process. The author will attempt to collect probe speed data for the next 2 or 3 years for these 71% segments. In future, it is expected to acquire more data with a better coverage of probe vehicles traversing on those roads. If speed data with greater than 30% availability can be collected, these will be utilized to apply bootstrapping for the 71% segments and estimate the minimum sample rates.

This study used only one-years' worth of data while applying the bootstrapping on different facilities. However, the probe data collection range for all the facilities can be extended over 3 or 4 years instead of one to observe whether the estimated sample rates are consistent with this study's results or not. Apart from that, the bootstrapping method in this study uses 2,000 replicated samples due the limitation in computational time. Nevertheless, a set of 5,000 or 10,000 replications can be explored to see if the bootstrapped minimum sample rate is sensitive towards the replication numbers or not.

# REFERENCES

1. S. Eshragh, S. E. Young, E. Sharifi, M. Hamedi, K. F. Sadabadi, Indirect Validation of Probe Speed Data on Arterial Corridors. *Transportation Research Record: Journal of the Transportation Research Board*, 105-111 (2017).
2. R. M. Juster, S. E. Young, E. Sharifi, "Probe-Based Arterial Performance Measures Validation," (2015).
3. X. Zhang, M. J. T. R. R. J. o. t. T. R. B. Chen, Genetic algorithm–based routing problem considering the travel reliability under asymmetrical travel time distributions. 114-121 (2016).
4. X. Zhang, Incorporating travel time reliability into transportation network modeling. (2017).
5. A. Haghani, M. Hamedi, K. F. Sadabadi, I-95 Corridor coalition vehicle probe project: Validation of INRIX data. *I-95 Corridor Coalition* **9**, (2009).
6. E. Sharifi *et al.*, "Quality assessment of outsourced probe data on signalized arterials: Nine case studies in Mid-Atlantic region," (2016).
7. A. D. Patire, M. Wright, B. Prodhomme, A. M. Bayen, How much GPS data do we need? *Transportation Research Part C: Emerging Technologies* **58**, 325-342 (2015).
8. FDOT:, Use of Multiple Data Sources for Monitoring Mobility Performance (2015). (2015).
9. Y. Wang, B. N. Araghi, Y. Malinovskiy, J. Corey, T. Cheng, "Error assessment for emerging traffic data collection devices," (2014).
10. M. Chen, X. Zhang, Collection and Analysis of 2013-2014 Travel Time Data. (2017).
11. K. Jha, M. W. Burris, W. L. Eisele, D. L. Schrank, T. J. Lomax, Estimating Reference Speed from Probe-based Travel Speed Data for Performance Measurement. (2018).
12. T. Lomax *et al.*, "Refining the Real-Timed Urban Mobility Report," (2012).
13. Y. Zhang, M. Hamedi, A. Haghani, S. Mahapatra, X. Zhang, "How Data Affect Travel Time Reliability Measures: An Empirical Study," (2015).
14. M. Yun, W. Qin, Minimum Sampling Size of Floating Cars for Urban Link Travel Time Distribution Estimation. *Transportation Research Record*, 0361198119834297 (2019).
15. J. Bates, J. Polak, P. Jones, A. Cook, The valuation of reliability for personal travel. *Transportation Research Part E: Logistics and Transportation Review* **37**, 191-229 (2001).
16. M. Chen, G. Yu, P. Chen, Y. Wang, A copula-based approach for estimating the travel time reliability of urban arterial. *Transportation Research Part C: Emerging Technologies* **82**, 1-23 (2017).
17. P. Chen, K. Yin, J. Sun, Application of finite mixture of regression model with varying mixing probabilities to estimation of urban arterial travel times. *Transportation Research Record: Journal of the Transportation Research Board*, 96-105 (2014).

18. E. Kazagli, H. Koutsopoulos, Estimation of arterial travel time from automatic number plate recognition data. *Transportation Research Record: Journal of the Transportation Research Board*, 22-31 (2013).

19. W. Zhou, S. Zhao, K. Liu, in *ICCTP 2011: Towards Sustainable Transportation Systems*. (2011), pp. 1434-1441.

20. M. Chen, S. Chien, Determining the number of probe vehicles for freeway travel time estimation by microscopic simulation. *Transportation Research Record: Journal of the Transportation Research Board*, 61-68 (2000).

21. M. Cetin, G. F. List, Y. Zhou, Factors affecting minimum number of probes required for reliable estimation of travel time. *Transportation research record* **1917**, 37-44 (2005).

22. T. Miwa, D. Kiuchi, T. Yamamoto, T. Morikawa, Development of map matching algorithm for low frequency probe data. *Transportation Research Part C: Emerging Technologies* **22**, 132-145 (2012).

23. Z. Ma, H. N. Koutsopoulos, L. Ferreira, M. Mesbah, Estimation of trip travel time distribution using a generalized Markov chain approach. *Transportation Research Part C: Emerging Technologies* **74**, 1-21 (2017).

24. S. Yang, A. Malik, Y.-J. Wu, Travel Time Reliability Using the Hasofer–Lind–Rackwitz–Fiessler Algorithm and Kernel Density Estimation. *Transportation Research Record* **2442**, 85-95 (2014).

25. A. Polus, A study of travel time and reliability on arterial routes. *Transportation* **8**, 141-151 (1979).

26. J. Van Lint, H. Van Zuylen, Monitoring and predicting freeway travel time reliability: Using width and skew of day-to-day travel time distribution. *Transportation Research Record: Journal of the Transportation Research Board*, 54-62 (2005).

27. E. Emam, H. Ai-Deek, Using real-life dual-loop detector data to develop new methodology for estimating freeway travel time reliability. *Transportation Research Record: Journal of the Transportation Research Board*, 140-150 (2006).

28. A. Higatani *et al.*, Empirical analysis of travel time reliability measures in Hanshin expressway network. *Journal of Intelligent Transportation Systems* **13**, 28-38 (2009).

29. J. Kwon, T. Barkley, R. Hranac, K. Petty, N. Compin, Decomposition of travel time reliability into various sources: incidents, weather, work zones, special events, and base capacity. *Transportation Research Record: Journal of the Transportation Research Board*, 28-33 (2011).

30. M. Yazici, C. Kamga, K. Mouskos, Analysis of travel time reliability in New York city based on day-of-week and time-of-day periods. *Transportation Research Record: Journal of the Transportation Research Board*, 83-95 (2012).

31. S. Yang, Y.-J. Wu, "Minimum Sample Size for Measuring Travel Time Reliability," (2015).

32. S. Yang, P. Cooke, How accurate is your travel time reliability?—Measuring accuracy using bootstrapping and lognormal mixture models. *Journal of Intelligent Transportation Systems*, 1-15 (2018).

33. V. Varsha, G. H. Pandey, K. R. Rao, B. Bindhu, Determination of Sample Size for Speed Measurement on Urban Arterials. *Transportation Research Procedia* **17**, 384-390 (2016).
34. J. C. Oppenlander, Sample size determination for travel time and delay studies. *Traffic Engineering* **46**, (1976).
35. C. A. Quiroga, D. Bullock, Determination of sample sizes for travel time studies. *ITE Journal* **68**, 92-98 (1998).
36. S. Li, K. Zhu, B. Van Gelder, J. Nagle, C. Tuttle, Reconsideration of sample size requirements for field traffic data collection with global positioning system devices. *Transportation Research Record: Journal of the Transportation Research Board*, 17-22 (2002).
37. A. G. Barnett, J. C. Van Der Pols, A. J. Dobson, Regression to the mean: what it is and how to deal with it. *International journal of epidemiology* **34**, 215-220 (2004).
38. P. Park, D. Lord, Investigating regression to the mean in before-and-after speed data analysis. *Transportation Research Record: Journal of the Transportation Research Board*, 52-58 (2010).
39. E. Green, J. Ripy, M. Chen, X. Zhang, in *Transportation Research Board 92 nd Annual Meeting, Transportation Research Board*. (2013), vol. 92, pp. 1-15.
40. M. Chen, X. Zhang, E. Green, "Analysis of Historical Travel Time Data," (2015).
41. K. Srinivasan, P. Jovanis, Determination of number of probe vehicles required for reliable travel time measurement in urban network. *Transportation Research Record: Journal of the Transportation Research Board*, 15-22 (1996).
42. A. Toppen, K. Wunderlich, *Travel time data collection for measurement of advanced traveler information systems accuracy*. (Mitretek Systems, 2003), vol. 23.
43. F. Guo, H. Rakha, S. Park, Multistate model for travel time reliability. *Transportation Research Record* **2188**, 46-54 (2010).
44. S. Yang, Y.-J. Wu, Z. Yin, Y. Feng, Estimating Freeway Travel Times Using the General Motors Model. *Transportation Research Record: Journal of the Transportation Research Board*, 83-94 (2016).
45. B. Efron, Bootstrap Methods: Another Look at the Jackknife. The Annals of Statistics. **7**, 1-26 (1979).
46. B. Efron, R. J. Tibshirani, *An introduction to the bootstrap*. (CRC press, 1994).
47. D. J. Hand, Principles of data mining. *Drug safety* **30**, 621-622 (2007).
48. H. Han, X. Guo, H. Yu, in *2016 7th IEEE International Conference on Software Engineering and Service Science (ICSESS)*. (IEEE, 2016), pp. 219-224.
49. L. Breiman, Random forests. *Machine learning* **45**, 5-32 (2001).
50. A. Liaw, M. Wiener. (2002).
51. V. Bax, W. J. A. G. Francesconi, Environmental predictors of forest change: An analysis of natural predisposition to deforestation in the tropical Andes region, Peru. **91**, 99-110 (2018).
52. V. Svetnik *et al.*, Random forest: a classification and regression tool for compound classification and QSAR modeling. **43**, 1947-1958 (2003).

53. J. S. Evans, M. A. Murphy, Z. A. Holden, S. A. Cushman, in *Predictive species and habitat modeling in landscape ecology*. (Springer, 2011), pp. 139-159.

54. B. Heung, C. E. Bulmer, M. G. J. G. Schmidt, Predictive soil parent material mapping at a regional-scale: a random forest approach. **214**, 141-154 (2014).

55. R. Genuer, J.-M. Poggi, C. Tuleau-Malot, Variable selection using random forests. *Pattern Recognition Letters* **31**, 2225-2236 (2010).

56. D. Saha, P. Alluri, A. J. J. o. A. T. Gan, A random forests approach to prioritize Highway Safety Manual (HSM) variables for data collection. **50**, 522-540 (2016).

57. M. Chen, H. Gong, Speed Estimation for Air Quality Analysis. *In, Kentucky Transportation Center*, (2005).

58. D. Lee, M. Burris, in *HERS-ST Highway Economic Requirements System-State Version: Technical Report* (2005).

# VITA

- Fahmida Rahman
- Place of Birth: Chittagong, Bangladesh
- Educational Institutions attended:
    - Bangladesh University of Engineering & Technology (BUET)
        - Bachelors of Science in Civil Engineering
- Professionals Position held:
    - Graduate Research Assistant at University of Kentucky
    - Graduate Teaching Assistant at University of Kentucky
    - Graduate Research Assistant at BUET
- Scholastic Honors
    - University of Kentucky
        - Chi Epsilon
    - BUET
        - Merit Scholarships