Speed Profile Variation as a Surrogate Measure of Road Safety Based on GPS-Equipped Vehicle Data

A Dissertation
Presented to
The Academic Faculty

By

Saroch Boonsiripant

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in Civil Engineering

Georgia Institute of Technology

May 2009

Speed Profile Variation as a Surrogate Measure of Road Safety Based on GPS-Equipped Vehicle Data

Approved by:

Dr. Michael P. Hunter, Advisor
School of Civil and Environmental Engineering
Georgia Institute of Technology

Dr. Randall L. Guensler
School of Civil and Environmental Engineering
Georgia Institute of Technology

Dr. Michael O. Rodgers
School of Civil and Environmental Engineering
Georgia Institute of Technology

Dr. Kwok-Leung Tsui
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Karen K. Dixon
School of Civil and Construction Engineering
Oregon State University

Date Approved: March 26, 2009

Dedicated to my parents

Pornphun and Surachart Boonsiripant

ACKNOWLEDGEMENTS

When looking back to all those years I spent in the Ph.D. program at Georgia Tech, I realized that the journey I have been through is much more valuable than the destination I have reached. I met many great people along my journey and would like to thank them for their supports over the past four years.

I would like to express my gratitude to all my thesis committee members for their guidance on this research work. The first person I would like to thank is Dr. Kwok-Leung Tsui. My interest in statistics was inspired by his excellent teaching in a few of his statistics courses. My gratitude also goes to Dr. Randall Guensler for his comprehensive data and invaluable guidance and suggestions. Moreover, I feel very fortunate to have an opportunity to work with Dr. Karen Dixon on a FHWA research project four years ago which helped me formed the scope of this research. I also would like to thank Dr. Michael Rodgers for his closely involvement during the last phase of my dissertation work. Discussing research with him is always a pleasure to me because of his thorough understanding and broad knowledge on the related subjects. Most importantly, I would like to thank my advisor, Dr. Michael Hunter, who has been an outstanding advisor and mentor over the past four years. I could not remember how many time he had to skip lunch or miss dinner with his family when trying to help me go through the research problems. I am extremely grateful to have had the opportunity to work with him.

I have special thanks to Dr. John Leonard who gave me the most valuable book I ever read, Getting Things Done. It does really help me getting many things done and keeping my life in proper balance.

My gratitude also goes to all my friends around for their true friendship and encouragement. To name a few, my thanks go to Dr. Tudor Bodea, Dr. Hsing-Chung Chu, Dr. Angshuman Guin, Hoe Kyoung Kim, Piyanut Itarut, Chanin Ruangthaveekoon, Wonho Suh, Dr. Chanin Tongchitpakdee, and Seungkook Wu.

Finally, I would like to thank my wonderful parents for their unconditional endless love and continued support throughout my journey. My dearest girlfriend, Monthisa Kortrakul (Peach) deserves my special gratitude. I thank her for the enormous support, patience, and faith she has in me over the past four years. Without her, I would never have accomplished what I have today.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AADT	Average Annual Daily Traffic
ANOVA	Analysis of Variance
CV	Coefficient of Variation
FHWA	Federal Highway Administration
GDOT	Georgia Department of Transportation
GIS	Geographic Information System
GLM	Generalized Linear Model
GPS	Global Positioning System
GT	Georgia Institute of Technology
HPMS	Highway Performance Monitoring System
HSIP	Highway Safety Improvement Program
ML	Maximum Likelihood
NCHRP	National Cooperative Highway Research Program
pdf	Probability Density Function
PDOP	Positional Dilution of Precision
RC	Roadway Characteristics
RCLINK	Roadway Characteristics Link
RMS	Root Mean Square
RSS	Residual Sum of Squares
SAFETEA-LU	Safety, Accountable, Flexible, Efficient Transportation Equity Act: A Legacy for Users
SAT	Number of Satellites
SD	Standard Deviation

SHSP	State Strategic Highway Safety Plan
STARS	State Traffic and Report Statistics
TWLTL	Two Way Left Turn Lane
QQ plot	Quantile-Quantile plot
VII	Vehicle Infrastructure Integration

SUMMARY

Road network screening for potentially high incident locations is the first step in a road safety improvement program. During the screening process, road network crash data are required for the identification of high crash locations, a.k.a., black spots. In situations where historical crash data are limited or not available, *surrogate safety measures*, such as traffic and roadway characteristics are often considered. A surrogate safety measure is an indirect measure of safety, which attempts to assess the safety of a road facility through means other than crash data. Among speed characteristics measurements *speed variation* is often used as a surrogate measure of safety. There are a number of studies that attempt to establish a relationship between speed variation and crash risk but the existence form of such a relationship is still hotly debated in the literature. The increasing use of Global Positioning System (GPS) devices for collecting traffic operations data, such as vehicle speed and travel time, was led to interest in using GPS data derived measures as potential indices for roadway safety. As the deployment of GPS-instrumented vehicles becomes more prevalent, we may be able to use this new data streams to better evaluate roadway safety. Our hypothesis is that vehicle speed characteristics may be used to reveal roadways with safety issues such as poorly-designed road geometries, limited sight distance, and high conflict movements from/to side streets.

The primary objective of this research is to explore the use of speed variation over a roadway segment as an indirect means to estimate crash frequency of the facility. This estimated crash frequency can be used as a substitute when historical crash data are unavailable or a proactive means to identify sites that need further engineering studies.

To accomplish this objective, sample operating speed and incident data were collected for corridors in the Metro Atlanta area. To measure operating speeds, second-by-second speed data were obtained from more than 460 GPS-equipped vehicles participating in the Commute Atlanta Study over the 2004 calendar year. Incident data was provided by the Georgia Department of Transportation Office of Traffic Safety and Design. Based on the speed and incident data, several definitions of speed variation are considered as potential surrogate safety measures. The quantified relationships between surrogate measures and crash frequency are developed using Binary Recursive Partitioning methods and a Generalized Linear Modeling (GLM) approach.

This research effort is expected to result in several contributions. First, this study will develop a methodology to determine speed profile under various conditions using vehicle activity data. Second, a speed variation definition suitable for GPS data that can be used as a surrogate safety measure will be recommended. Lastly, the process will provide safety prediction models for identifying high crash locations in the network screening process for urban streets.

Chapter 1. INTRODUCTION

1.1 BACKGROUND

Despite the gradual reduction in fatality and injury rates over the past several years, more than 43,000 people were killed and 2.7 million were injured on the highways in 2005 (*1*). With these high numbers of fatalities and injuries, there is an urgent need for the public agencies to more effectively allocate their limited budget for safety improvement projects. This need was highlighted when the "Safe, Accountable, Flexible, Efficient Transportation Equity Act: A Legacy for Users (SAFETEA-LU)" legislation passed in 2005 and $5.1 billion in funding was allocated to the Highway Safety Improvement Program (HSIP) to achieve significant reductions in fatalities and serious injuries due to motor vehicle crashes (*2*). As part of this program, each state is required to submit State Strategic Highway Safety Plans (SHSP) to identify highway safety problems, develop an evaluation process to assess safety improvements, and use these criteria to prioritize safety improvement projects.

For any road safety improvement program, the process usually involves three steps: screening the road network for high crash locations, conducting detailed engineering studies, and prioritizing/implementing safety improvement projects (*3*). The first step, road network screening, requires historical crash data. However, the data are often limited or not available or engineers are required to evaluate safety on a particular corridor in a short timeframe.

When sufficient crash data are not readily available, another approach to identifying safety problems is to use indirect safety measures, i.e., a surrogate safety measures. A surrogate safety measure is an indirect measure of safety, which attempts to assess the safety of a facility through means other than crash data (*4*). Road accidents are influenced by many factors such as excessive speed, road geometric design, traffic volume, weather, reasons for travelling, driver's physical and mental conditions, and safety campaigns (*5*). Speed characteristics are influenced by three major components, i.e., drivers/vehicles, roadway, and roadside environment (*6*). Figure 1 summarizes the relationship between speed characteristics, safety, and the main influential factors. Since the driver, roadway, and vehicle components influence both speed characteristics and road safety, there might be a possibility of using some speed characteristics measures as a surrogate safety measures. A number of studies (*7-16*) suggest that there is a relationship between speed characteristics and road safety.

Most of the previous studies in the literature drew conclusions based on speed data collected at one or a few spots along the studied corridors. Even though speed data along a corridor may provide a better understanding of design speed vs. operating speed consistency, few of these studies (*17*, *18*) investigated the relationship between speed characteristics along a corridor and road safety due to limitations in their data collection methods.

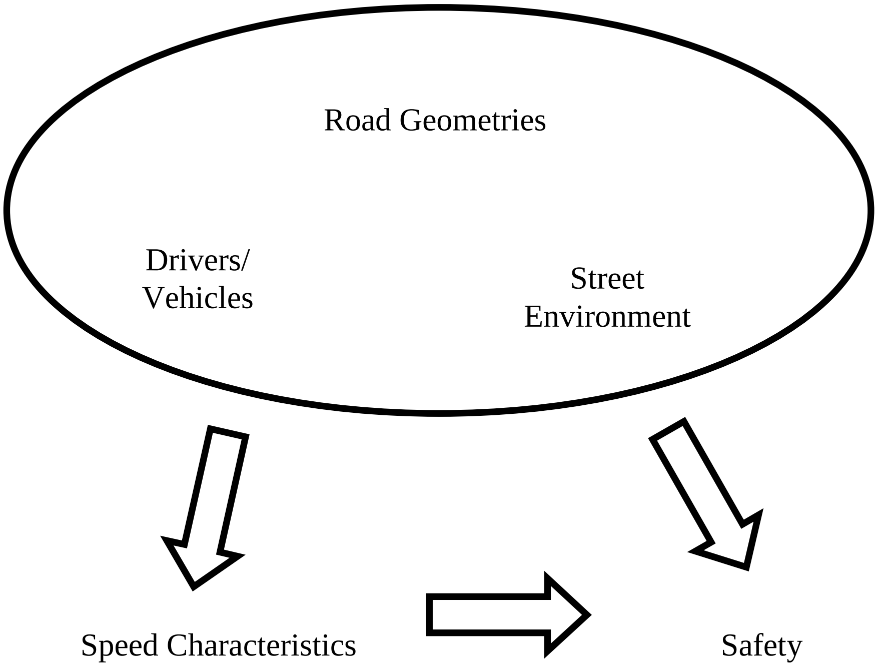


Figure 1: Relationships among speed characteristics, road safety, and influential factors

With the emergence of the Vehicle Infrastructure Integration (VII) initiative (*19*, *20*), significant amounts of traffic data, such as vehicular speed, travel time, and other data, will likely be widely available in the future. This has led to interest in using several measures derived from such extensive traffic data as potential roadway safety indices. The hypothesis is that speed characteristics derived from one-dimension speed data would reveal roadways with design issues such as sharp curves, limited sight distance, mountainous terrain, driveway deficiencies, etc., that could result in higher crash risks.

1.2 PROBLEM DEFINITION

When the historical crash data are not available, the expected crash frequency may potentially be estimated using surrogate measures based on road geometries and speed characteristics. For the road network screening purpose, collecting road geometric and roadside environmental information of all the road segments is often not feasible.

Most of the previous studies that measure speed have employed automated traffic counters or laser/radar speed measurements at specific points along the roadway, assuming that the monitored spot speeds or speed profiles collected along a tangent or a horizontal curve are representative of the speeds of the entire corridor. However, this assumption may not hold when the road has geometric elements such as horizontal or vertical curves, limited sight distance, high driveway density as these factors affect vehicle speed (*6*).

This research uses GPS-equipped vehicle data to quantify the relationship between observed roadway speed characteristics, as revealed by the instrumented vehicle data, and observed crash frequency on these roadways, as revealed in the regional crash database. With the use of GPS technology, vehicles are tracked at second-by-second resolution over an entire trip, any time of the day, under any weather condition. Therefore, this study does not need to assume that a vehicle speed measured at one location is a representative of speeds over the entire road section.

The hypothesis of this research is that the speed variance can be used as a surrogate measure of safety to evaluate urban streets. This research aims to provide a simple screening tool to assist transportation safety engineers and practitioners in identifying roadway segments with safety issues during the evaluation process of safety improvement programs. The developed model would require only the speed profile data to identify potential corridors with a crash risk significantly above expectation.

1.3 RESEARCH OBJECTIVES

The objectives of this research effort are as follows:

- To determine the speed profile, including operating, average, and various percentiles, of a roadway using GPS-instrumented vehicle data.

- To develop *profile-based* traffic attributes using the instrumented vehicle data for potential use as surrogate safety measures.

- To investigate the relationships of road safety to the proposed surrogate safety measures.

1.4 RESEARCH CONTRIBUTIONS

This research effort is expected to provide the following contributions:

- Demonstrate methodology to determine free-flow speed, average speed, and speed percentiles using vehicle activity data.

- Review speed variation definitions found in the literature and propose new measures that can be used as surrogate safety measures.

- Develop a safety prediction model that requires only speed profile data for identifying potential high crash locations in the network screening process for urban streets.

1.5 DISSERTATION OUTLINE

This study proposes to use several traffic attributes derived from instrumented vehicles to predict the crash frequency of the urban streets. Chapter 2 discusses the previous studies that relate speed characteristics to road safety. Most of the previous research efforts were seen to measure the variation of speed at a specific location. However, these point-specific measures do not always show a strong relationship with road safety.

Chapter 3 describes the data used in this study. The corridor selection methodology is also included in this chapter.

In practice, the operating speed is measured during the free-flow condition. However, there is no direct information from the vehicle activity data to determine whether a trip was made under free-flow conditions. Therefore, data filtering processes were developed in Chapter 4 to detect trips that are not likely to be under free-flow condition. Also included this chapter is the crash data processing to obtain crash data in the scope of interest.

Chapter 5 proposes several new traffic attributes based on speed profile data obtained from the previous chapter. The proposed attributes include speed-related and stop-related measures. Most of these measures quantify variation of speed along the corridor, rather than variation of speed at one point.

Since multiple filters were used to obtain likely free-flow speed data, it is vital to understand the effect of each filter on the proposed speed measures. Chapter 6 provides sensitivity analyses of data filters to the speed measure.

Chapter 7 presents the development of crash prediction models for different facility classes. The summary of findings, research contributions, and future recommended work is described in Chapter 8.

The remainder of this dissertation includes Appendix A – the summary of data processing results in Chapter 4 and Appendix B – the sensitivity analysis results of spacing distance in Chapter 5.

Chapter 2. LITERATURE REVIEW

The objective of this research is to develop crash prediction models based on speed characteristics. Since this study utilizes GPS second-by-second speed data along corridors, several speed characteristics can be defined based on this one-dimensional speed data. The first step in the realization of this objective is to assess previous studies that used speed characteristics to quantify road safety.

2.1 CHAPTER ORGANIZATION

This chapter begins with an overview of previous research regarding the speed-safety relationship in Section 2.2. This is followed in Section 2.3 by speed variation definitions from previous research efforts. The chapter concludes with a presentation of statistical models that were used in the past to predict crash occurrences on roadways with the use of speed characteristics as the predictor variable in Section 2.4.

2.2 RELATIONSHIP BETWEEN SPEED CHARACTERISTICS AND SAFETY

Previous research regarding the relationship between speed characteristics and crashes can be divided into three primary groups according to the experimental designs: comparison of pre-crash speed and prevailing speed; comparison of crash-involved driver’s speed and prevailing speed; and lastly, comparison of aggregated speeds among different roadways (*11*).

2.2.1 Pre-crash Speed vs. Prevailing Speed

The first type of research design involves measuring pre-crash and prevailing speeds and relating the difference between these speeds to safety. This type of research takes an *event-based* approach, i.e., comparing the speed from a crash event with the prevailing facility speed during non-crash period (*11*). Pre-crash speed is usually obtained from multiple sources such as police reports and crash construction techniques.

One of the first attempts to examine the relationship between vehicle speeds and crash risk was undertaken by Solomon (*15*). Solomon estimated pre-crash traveling speeds on selected rural highway segments, including 35 sites in 11 States. The author compared pre-crash traveling speeds with speed measurements during normal conditions, and found that many vehicles involved in rural highway crashes were traveling well above or well below the average speed under normal conditions. This variation is characterized by the U-shape form in Figure 2. Even though this early research demonstrated a promising opportunity for using speed variation as a surrogate safety measure, its methodology, namely, pre-speed measurement presents several challenges. The pre-crash speed data were collected from police reports or estimated from a similar event, both sources with potential accuracy issues. The author did not describe how he estimated pre-crash speeds when they were not available from the previous sources. The methodology also assumes a uniform speed throughout the segment length, despite changes in terrain and geometry.

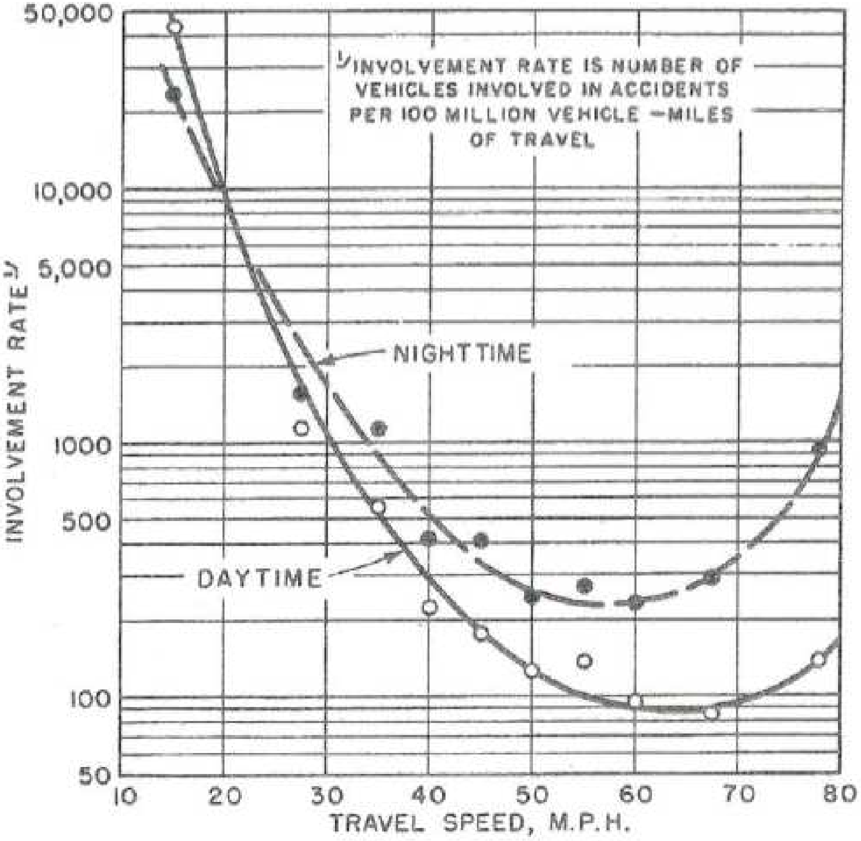


Figure 2: Crash involvement rate by variation from the average speed on study section, day and night (*15*), Reproduced by (*21*)

Kockelman and Murray (*12*) used a more aggregated approach to measure pre-crash speed. The authors collected speed data using loop detectors installed on six Southern California freeways. The traffic counter device provided a data stream at a 30-second aggregation level that includes average speed and traffic density. The authors investigated 744 crashes that occurred during a 1-month period on the study freeway segments and compared their accompanied pre-crash speed variations with the speed variation during normal conditions. The study concluded there is no evidence that speed or speed variation has a relationship with crash occurrence. However, the authors noted several data limitations to their study. For instance, the crash times from the police reports were rarely precise. In addition, speeds are based on using 30-second aggregated data; thus, speed variation had to be inferred from the variation in average speeds over a series of intervals and over a series of lanes.

Kloeden et al. (*11*) determined the relationship between free-flow speed and crash involvement using a case study design, i.e., comparing the speed of a vehicle involved in a crash with speeds of other vehicles travelling at the same time and location but not involved in the crash. The methodology was generally similar to Solomon's study except that pre-crash speeds were determined using computer-aided crash reconstruction techniques developed by the authors, rather than solely based on police reports. The authors concluded that crash-involved vehicles were generally traveling faster than those not involved in crashes. Contrary to Solomon's findings, the study showed that slow-speed vehicles were not associated with high crash risk.

In summary, this type of research design can show the first moment relationship between speed difference and crash involvement. That is, traveling speed of the crash-involved vehicle was directly compared with the prevailing speed of the facility. However, pre-crash travelling speed in the police reports usually was an estimate from the witnesses, drivers, or police (*11*), and therefore, as stated by Ogle (*22*), reduced the soundness of their findings and conclusions. Pre-crash speed can also be expensive to obtain with the current technology available.

2.2.2 Speed of Crash-Involved Drivers vs. Speed of Non-Crash-Involved Drivers

This type of study compares the speed characteristics of drivers with and without crash involvement during the period of interest. This approach is considered a *driver-based* approach to study the relationship between different speed characteristics and crash involvement (*11*). The hypothesis of this type of research is that drivers with crash involvement histories operate their vehicles differently from the drivers without crash involvement history, e.g., individuals that drive faster or accelerate/decelerate more abruptly are more prone to accidents. This type of research is useful for classifying the safety of drivers by their driving behaviors.

One of these driver-based studies was conducted by Fildes et al. (*8*). The authors examined the relationship between driver attributes and speed characteristics in Victoria, Australia. As part of this study, the authors measured vehicle speeds on two urban arterials and two rural undivided highways. After recording their speeds, drivers were stopped at the downstream location and were interviewed to determine if they were involved in any accidents during the past five years and also other related details about the incident(s). The study found that drivers with speeds above the 85th percentile had a higher crash risk than any other drivers. They also found that drivers with a measured speed less than 15th percentile were the least likely to be involved in a crash.

West et al. (*16*) studied relationships between driving behaviors and crash involvement over a three year period. Forty-eight drivers were asked to drive on a predefine route, a mix between urban and motorway routes, and then report their driving behaviors.

Additionally, each driver was accompanied by an observer to validate the reliability of the driver’s self-report. The authors carried out a multiple logistic regression analysis with crash involvement as the dependent variable. They found that the observed speed on motorway has a positive relationship with self-reported crash involvement. In other words, drivers with high driving speeds were associated with at least one crash during the past three years.

Jun (*23*) utilized GPS-measured activity data to compare the driving behavior of two driver groups, those with and without crash-involved experiences over a 14-month period. The author found that driving behaviors such as speeding pattern and hard acceleration/deceleration activity are among the most important factors for determining potential crash involvement rate of an individual.

In summary, driver-based study designs investigate the difference between the driving behaviors of drivers with and without past crash involvement. One assumption of this approach is that driving behaviors do not change after the drivers have accidents. In reality, drivers might be more cautious with their driving after they experience accidents (*11*).

2.2.3 Aggregated Speed Characteristics of Different Roadways

The last category of studies investigates the relationship between aggregated speed characteristics and crash frequency/rate. This study design is considered a *facility-based* approach, i.e., comparing speed characteristics and safety associated with different road segments (*11*). A number of studies have been conducted using this research design because facility speed data tend to be easier to obtain than the data required in the first two described research approaches. The hypothesis is that poorly designed roads that results in high crash risk have different speed characteristics from lower crash risk roads.

Garber and Gadiraju (*9*) investigated whether a discrepancy between design speed and speed limit influences operating speed variability, which in turn influences crash occurrences. Traffic data and crash data were collected over 36 sites in Virginia including urban freeways, rural freeways, urban arterials, rural arterials, and rural major collectors. Individual vehicle speeds and traffic volume were collected for 24-hour periods using automated traffic data recorders. Design speed data was obtained from the highway log sheets. The authors performed an ANOVA test and found that average speed, speed variance, design speed, and highway type have a significant effect on crash rate. A regression model was developed to quantify the relationship between crash rate and speed variance. It was concluded that the crash rates increased with increasing speed variance for all facility types. In addition, the difference between posted and design speeds has a significant effect on speed variance.

Lave (*13*) proposed a concept of coordination between drivers on the road. In this study, the author used speed variance as a measure of the coordination. For example, high speed variation among drivers on a road segment of interest would refer to low driver coordination and vice versa. The author hypothesized that low driver coordination led to higher fatality rates. In the study, fatal crash data and driving speed data were collected for the period 1981-1982 from 50 states for six classifications of roadways (i.e., rural interstates, arterials, and collectors; and urban freeways, highways, and arterials). Speed variance was calculated as the 85th percentile speed minus the average speed. Several regression analyses were performed with fatality rate as a dependent variable. The author found that mean speed was not statistically significant in his crash prediction models; however, variation from the mean was significant.

Anderson et al. (*17*) studied the relationship of safety to several geometric design consistency measures for rural two-lane highways. One of these measures was the speed reduction on a horizontal curve relative to the preceding tangent or curve. The speed reduction values for 5,287 horizontal curves were estimated from speed prediction equations. The authors found a positive relationship between the crash frequency and speed reduction on a horizontal curve. In other words, the greater the speed reduction experienced by drivers on a horizontal curve, the greater the crash involvement of that curve.

In summary, the facility-based study examines the relationships of the aggregated speed characteristics and safety at multiple road facilities. The use of aggregated speeds of the roadway can avoid the problem of estimating the pre-crash speeds of vehicles as seen in Section 2.2.1. Nevertheless, one criticism to the facility-based approach is that the speed-safety relationship established from this approach is rather weak because the aggregated speed measured over the study period might not reflect the actual speed distribution at the time of the crash occurrences (*11*, *24*).

2.3 SURROGATE SAFETY DEFINITIONS

Numerous studies have considered speed-related measures as a potential parameter in the investigation of roadway safety. These measures are often defined differently based on the purpose of the study and the available data collection method. This section summarizes the speed-related measures found in previous research.

2.3.1 Speed

It is clear that higher pre-crashed speed generates a higher impact and therefore increases in likely crash severity. Additionally, at a higher speed, the driver has less time to respond to the incident and is less likely to successfully avoid the crash.

Aljanahi (*5*) investigated the effect of the following speed measures on the expected number of crashes:

- The 85th percentile speed
- The 93th percentile speed
- Root mean square of measured speeds, $RMS = \sqrt{\frac{1}{n}\sum_{i=1}^{n} v_i^2}$, where $v_i$is the ith individual speed point and i = 1,…,n.

Mean speed is also used by several studies to establish the relationship between speed and road safety (*5, 9, 13*).

2.3.2 Speed Variance

Standard deviation and variance of speeds measured at a specific location were used in several studies (*5*, *9*). The formula follows the conventional variance calculation:

$$\sigma_v{}^2 = \frac{\sum_{i=1}^{n}(v_i - \bar{v})^2}{n-1},$$

where $\sigma_v{}^2$ is the speed variance and $\bar{v}$ is the mean of the *n* speed measurements $v_1 \ldots v_n$ at a specific location.

2.3.3 Other Forms of Speed Dispersion

Lave (*13*) used speed variance as a measure of the dispersion of speeds among drivers. Since the actual variance or standard deviation of speeds was not available in the dataset, the author approximated the speed variance using the difference between the 85th percentile speed and the mean speed at a given location. The difference was assumed to be one standard deviation of observed speeds and was written as:

$$SV = \mathrm{V}_{85} - \bar{V}$$

where SV is the speed variance, V_{85} is the 85th percentile speed, and $\bar{V}$ is the average speed at a given point.

Aljanahi et al. (*5*) estimated the expected crash frequency using the following measures, Coefficient of Upper Speed (CUSS) and Skewness Index (SI), as the surrogates for variance of speeds:

$$CUSS = \frac{V_{85} - V_{50}}{V_{50}}$$

where V_{85} and V_{50} are the 85th and the 50th percentile speeds, respectively, and

$$SI = 2x\frac{V_{93} - V_{50}}{V_{93} - V_{7}}$$

where V_{93}, V_{50} and V_7 are the 93th, 50th , and 7th percentile speeds, respectively.

2.3.4 Speed Reduction

Speed reduction from tangent to horizontal curve sections has been proposed as a measure of design consistency (*17, 25*). It is defined as:

$$\Delta V85 = V85_t - V85_c$$

where $V85_t$ is the 85th percentile speed on a tangent section and $V85_c$ is the 85th percentile speed on the following curvature. The final model showed that the higher the speed reduction at a curvature, the greater the expected number of accidents at that location.

2.3.5 Acceleration Noise

Acceleration noise was first proposed by Herman et al (*26*) in 1959 as a means to measure traffic conditions and driving behavior. This measure is different from the previous speed measures in that acceleration noise is derived from speed data of an individual vehicle recorded along the corridor, rather than speed data of multiple vehicles at a single measurement point.

Acceleration noise (σ) is defined to be the root-mean-square of the acceleration, which can be formulated as follows:

$\sigma^2 = \frac{1}{T}\int_0^T (a(t) - a_{av})^2 dt$, and

$$a_{av} = \frac{1}{T}\int_0^T a(t)dt = \frac{1}{T}(v(T) - v(0))$$

where *v(t)* and *a(t)* are the speed and acceleration of a car at time *t* and a_{av} is the average acceleration of the car for a trip taking time *T*.

The value of acceleration noise varies by drivers and traffic conditions. Herman et al. found that a driver driving 5-10 mph faster than average traffic speed resulted in higher acceleration noise.

A few year later, Jones and Potts (*27*) tried to use this parameter to quantify road, driver, and traffic condition in Adelaide, Australia. The authors examined the effect of different roads, drivers, and traffic conditions on acceleration noise. Eight runs were made by two drivers on three road sections (with two road sections containing significant horizontal curvature) during daylight traffic conditions.

The results showed that the acceleration noise was significantly greater on roads with more horizontal curvatures. In addition, on the same section, a down grade tends to result in greater acceleration noise than an up grade. The authors explained that on a down grade it is more difficult to maintain a constant speed when negotiating a sharp curve compared with an up grade. This suggests that the interaction effect between two geometric features, e.g., sharp curve on a downhill section, can be captured using the acceleration noise parameter.

Even though the authors did not directly determine the relationship between the acceleration noise and number of accidents on different road sections, the authors concluded that a road with multiple curves, which is more likely to cause a crash, also tends to yield a greater acceleration noise. Table 1 summarizes the studies in the past with their research focus and definitions of speed variation. Few studies attempted to capture speed along corridor segments primarily due to equipment limitations. Knowledge gained from this section will be used to develop a new definition of speed variation based on GPS data.

Table 1: Summary of speed variation definitions from previous studies

Year	Authors	Research Topic	Facility Type	Location	Speed Data Collection	Speed variation Definition
1962	Jones and Potts (*27*)	Effects of roads, drivers, and traffic on acceleration noise	Urban/ Suburban Roads	Adelaid Hills, Australia	Tachograph	Root-mean-square of acceleration
1964	Solomon(*15*)	Measuring pre-crash speed	Rural highways	11 States in the U.S.	Estimated from crash report, spot speed study	Difference between pre-crash and mean speeds
1985	Lave(*13*)	Aggregate speed and fatality rate	Interstates, arterials, and collectors	50 States in the U.S.	NA	Difference between mean speed and 85th percentile speed

Table 1: Summary of speed variation definitions from previous studies(Continued)

Year	Authors	Research Topic	Facility Type	Location	Speed Data Collection	Speed variation Definition
1989	Garber and Gadiraju (*9*)	Aggregated speed and crash involvement	Interstates and arterials in rural and urban areas; Collectors in rural area	Multiple locations in Virginia	Traffic data recorder	Variance of speed from the mean
1999	Anderson (*17*)	Design consistency measures and crash frequency	Two-lane rural highways	State of Washington	Estimated from speed prediction models	The 85th percentile speed difference between two successive segments
1999	Aljananhi(*5*)	Aggregated speed and crash involvement	Highways	UK and Bahrain	Pneumatic sensors	Std Deviation, CUSS[a], SI[b]
2002	Yuan and Garber (*28*)	Aggregated speed and crash involvement	Rural interstates	10 States in the U.S.	NA	Speed variance
2006	Jun (*23*)	Driver's Speed Characteristics and Crash History	Freeways, arterials, and local roads	Atlanta, GA	GPS-observed travel data	Difference between driving and posted speeds, acceleration noise, cruise duration, etc
2006	Abdel-Aty(*29*)	Aggregated speed and crash involvement	Urban freeways	Orlando, FL	Loop detector	Coefficient of variation of speed
2007	Kockelman and Murray (*12*)	Aggregated speed and crash involvement	Urban freeways	Orange County, CA	Loop detector	Standard deviation of aggregated speeds

Note:

[a] Coefficient of Upper Speed, $CUSS = \frac{V_{85}-V_{50}}{V_{50}}$

[b] Skewness Index, $SI = 2x\frac{V_{93}-V_{50}}{V_{93}-V_{7}}$

2.4 EXISTING CRASH PREDICTION MODELS

This section describes statistical models and model assumptions that have been used by previous researchers. Until the last decade, most road safety research assumed that crash frequency and traffic volume have a linear relationship.

A crash prediction model was developed to evaluate the effect of median treatments on urban arterials in Phoenix, AZ and City of Omaha, NE (*30*). The authors assumed negative binomial distribution of the residuals and used maximum-likelihood techniques to estimate model parameters. The crash frequency was also assumed to have non-linear relationships with traffic volume and segment length. The 189 selected segments were at least 0.75 mile in length and at least 350 ft away from signalized intersections. The three-year crash data associated with the study segments included 7,125 midblock accidents. The researchers determined whether a crash was associated with the signalized intersections from the crash report by using the “intersection-related” field found in the reports. The researchers computed the crash rate for raised-curb, TWLTL, and undivided median treatment groups. Based on this preliminary analysis, the raised-curb median has the lowest crash rate, followed by the TWLTL, and undivided treatment. With regard to land uses, business and office type land uses were found to have a higher crash rate than residential or industrial land uses. The statistical analysis involved two stages: 1) use of analysis of variance (ANOVA) to determine factors that had a significant effect on crash frequency and 2) calibration of the crash prediction model using the Generalized Linear Model (GLM) approach. The ANOVA results suggested that land uses could be grouped as business/ office and residential/ industrial as land uses within the same group have similar crash trends. In addition, driveway density and unsignalized side street were found to be significant factors for crash frequency on segments with a business/ office land use group, but not on segments with a residential/ industrial land use group.

Alijanahi et al. (*5*) investigated the relationship between several measures of speed and crash rate on highways, 9 sites in the U.K. and 10 sites in Bahrain. Speed and traffic flow data were collected at a selected spot on each site using pneumatic sensors. The researchers used 5-year crash data in the U.K. and 4-year crash data in Bahrain. The multiplicative form was used to construct the crash model:

$$\mu = kLF^{a}H^{b}S^{c}$$

where μ is the expected number of accidents per four years (in Bahrain) and five years (in the U.K.), *L* is the length of road segment in kilometers, *F* is the traffic flow (10^5 vehicles per year), *H* is the percent truck; and *S* is a measure of speed characteristics. The model response, μ, was assumed to have a Poisson distribution and the model parameters, *a*, *b*, and *c* were estimated by maximum likelihood. The variable *L* has an exponent of 1, therefore, the authors assumed that number of crashes is proportional to the length of road segment. The results showed positive relationships between speed characteristics (e.g., mean speed and speed variability) and crashes per mile. However, these relationships were not statistically significant and the authors suggested further studies.

Design consistency and safety relationship were modeled in (*18*) using the GLM approach. The following model form was used:

$$E(Y) = a_0 L^{a_1} V^{a_2} e^{\sum_{j=1}^{m} B_j x_j}$$

where $E(Y)$ is the expected crash frequency; L is the section length; V is the AADT; x_j is any of the m variables in addition to L and V; and a and b are model parameters.

2.5 STATISTICAL MODELING APPROACH

The proposed statistical analysis tasks for development of the crash prediction models, i.e., Safety Performance Functions (SPFs), will include selecting the safety measurement for the models, model assumption, and model form. The following sections discuss details of each proposed task.

2.5.1 Safety Measurement for Model Development

One of the objectives of this research is to develop a crash prediction model using speed characteristics as predictor variables, allowing transportation safety engineers and practitioners to identify black spots in the road network. Two safety measurements are widely used for ranking sites for safety investigation, namely, crash count and crash rate. However, it is known that the expected crash frequency does not have a linear relationship with traffic flow and, as a result, the crash rate should not be used to compare the safety of two entities (*31-33*). Figure 3 illustrates the relationship between AADT and crash frequency. The slope of this graph represents crash rate, which changes when AADT reaches a certain threshold of each facility type. In other words, different road functional classifications (e.g., minor arterials, collectors, and local streets) have different crash characteristics. This research therefore proposes the use of crash frequency as a dependent variable in the crash model.

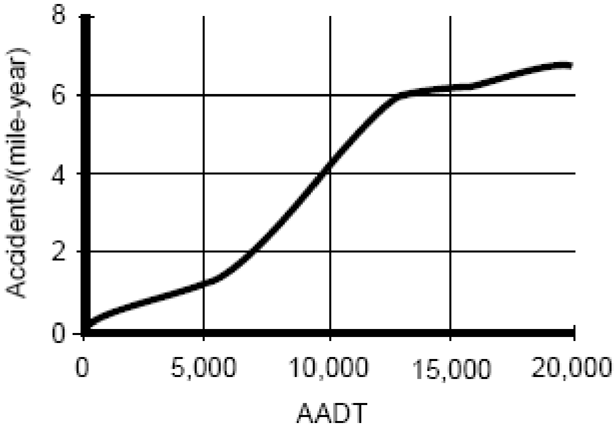


Figure 3: Relationship between Traffic Exposure and safety (*34*)

2.5.2 Model Forms

Hauer (*35*) summarized that there are three forms of statistical models commonly used for road safety research:

Additive model: $Y = L * (\beta_1 X_1 + \beta_2 X_2 + \cdots)$

Multiplicative model: $Y = \mathrm{L} * (\beta_0 {X_1}^{\beta_1} {X_2}^{\beta_2} \ldots)$

Multiplicative model (exponential base): $Y = L * (e^{\beta_0} e^{\beta_1 X_1} e^{\beta_2 X_2} \ldots)$

where Y is the expected crash frequency (crashes per segment length-unit time or crashes per unit time) and L is the segment length.

An Additive model is appropriate for point variables such as driveways and the presence of traffic signs while a multiplicative model is appropriate for segment variables such as lane width or shoulder type, which influence crash risk along the segment. The relationship between the predictor variables and dependent variable is usually determined by exploring the data through graphs and other visualizations. A suggested generic model form by Hauer (*35*) is:

$$Y = \alpha(L * Multiplicative\ Portion + Additive\ Portion)$$

where $Multiplicative\ Portion\ =\ f_0(AADT) * f_1(X_1) * f_2(X_2) * \dots$ and $Additive\ Portion\ = g_1(AADT, X_1') + g_2(AADT, X_2') + g_3(AADT, X_3') + \cdots$

In the equation above, Y indicates the expected number of accidents occurring on a road segment during a time period. In addition, α is a scale parameter which takes into account traits that are not included in this model such as weather and driver demographics and L is the length of segment. The notations $f_0(\cdot), f_1(\cdot)$, and $f_2(\cdot)$ represent functions of the variables $AADT, X_1$, and X_2 which have multiplicative influence to the expected crash frequency while $g_1(\cdot), g_2(\cdot)$, and $g_3(\cdot)$ denote the functions of the variables X_1', X_2', and X_3' which have additive influence to the expected crash frequency.

2.5.3 Modeling Approach

The quantified relationships between surrogate measures and crash frequency are developed using a Generalized Linear Modeling (GLM) approach. The three components of the GLMs are (*36*):

An Error Distribution – the distribution of the dependent variable.

A Link Function – the function that shows how the linear function of the explanatory variables are related to the expected value of the response. The general form is as follows:

$$g(\mu) = \beta_0 x_0 + \beta_1 x_1 + \ldots + \beta_n x_n$$

where $g(\mu)$ is the link function of the expected value of the response.

The Variance Function – the function presents the relationship between variance of the dependent variable and its mean. The general form of the variance function of the response *y* is as follows:

$$Var(y) = \emptyset V(\mu)$$

where ϕ is the dispersion parameter and $V(\mu)$ shows that the variance is a function of the mean. When the error is normally distributed, $V(\mu)$ is 1 and ϕ is σ^2. And when the error has Poisson distribution, $V(\mu)$ is μ and ϕ is 1.

2.5.4 Underlying Distribution Assumption

It is generally accepted that crash count occurrences follow the Poisson process (*17, 33, 34, 37*) for the following reasons:

- Crash frequencies are non-negative integers.

- A high number of crash events at a single location is rare.

The Poisson probability mass function follows the form:

$$P(X = y_i) = \frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!}$$

In the equation above, *P(X = y*$_i$*)* is the probability that the observed number of accidents is y during period i when X is Poisson distributed. Also, μ_i = the expected number of accidents that occur during the period of interest i.

One basic assumption when a random variable x follows Poisson distribution is that its variance is equal to its expected value, or

X~ Poisson, Var(X) = μ.

However, Kononov (*34*) indicated that this property does not hold for crash data in some cases. For instance, some crash data set are better represented by a long tail distribution, which usually indicates high variation. The author proposed to use a negative binomial distribution assumption when the crash data are overdispersed, i.e., the variance of the data are greater than the mean. The *negative binomial probability mass function* follows the form:

$$P(X = y_i) = \frac{\Gamma(\alpha^{-1} + y_i)}{y_i!\,\Gamma(\alpha^{-1})}\left(\frac{\alpha\mu_i}{1 + \mu_i}\right)^{y_i}\left(\frac{1}{1 + \alpha\mu_i}\right)^{\alpha^{-1}}$$

with the variance, $Var(x) = \mu + \alpha\mu^2$

In the equation above, α is the overdispersion parameter estimated by the maximum log-likelihood of the negative binomial function. $\Gamma(r) = (r-1)!$, r is a positive integer.

2.5.5 Tests of Goodness of Fit

The goodness of fit of a Poisson or negative binomial models can be assessed using the following statistical measures (*17, 38, 39*):

- Ordinary multiple correlation coefficient (R^2):

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

where y_i is the observation (crash frequency on corridor i) with $\hat{y}_i$ as the fitted value from the model and $\bar{y}$ as the sample average. The R-square measures how much variation can be explained by the fitted model compared with the model with only an intercept term ($\bar{y}$). Therefore, an R-square closer to one indicates the fitted model is able to explain most of the variation in the data.

- Deviance test: the Poisson deviance, i.e., G-statistics, is of the form:

$$D = 2\sum_{i=1}^{n}\left(y_i \log(y_i/\hat{\mu}_i) - (y_i - \hat{\mu}_i)\right)$$

The deviance follows the X^2distribution with n-p degrees of freedom where there are n observations and p parameters in the model.

- Dispersion parameter (σ_d): A measure of degree of dispersion of the data is of the form:

$$\sigma_d = \frac{Pearson\ X^2}{n - p}$$

and,

$$Pearson\ X^2 = \sum_i \frac{\left(y_i - E(\hat{\mu}_i)\right)^2}{Var(\hat{\mu}_i)}.$$

Values greater than 1 indicate under-dispersion while the values less than 1 indicate over-dispersion. The ideal condition, where the true variance equals mean, represents the dispersion value of one.

2.6 SUMMARY

This chapter reviewed the previous research regarding speed and safety relationships. Previous research works can be generally classified into three groups: event-based, driver-based, and facility-based studies. This dissertation uses the facility-based approach, i.e., comparing speed characteristics and safety associated with different road segments.

Several safety surrogate measures including speed, speed variation, speed reduction, and acceleration noise, were also discussed. Among these measures, only speed reduction and acceleration noise capture speed variability along a corridor. As a result, there is a need to develop profile-based traffic attributes using the instrumented vehicle data for potential use as safety surrogate measures.

The last section reviewed the statistical modeling methods used in the road safety analysis. Crash frequency was proposed as the dependent variable in this study as the crash frequency is not proportional to traffic volume. The GLM approach with the assumption of Poisson or negative binomial distribution of the error term is commonly applied in the crash prediction model development.

Chapter 3. DATA

This chapter describes the data collection methodology for speed, crash data, traffic data, and the road characteristics data. Additionally, since the speed data were obtained from the instrumented vehicles of a random selection of drivers in the Atlanta region, speed data might not available on all links. Therefore, a corridor selection methodology is also developed.

3.1 DATA COLLECTION

The following subsections describe data collection methodologies and data related issues. Further detail may be found in the FHWA project report (*6*). Data are classified into four primary groups including speed data, crash data, traffic data, and road characteristics data. The original speed data collection and corridor selection plan and procedure were developed as part of the FHWA project. The project was primarily conceived with determining the effect of roadway features on operating speed. The data collected for the FHWA project was then subsequently used for this effort.

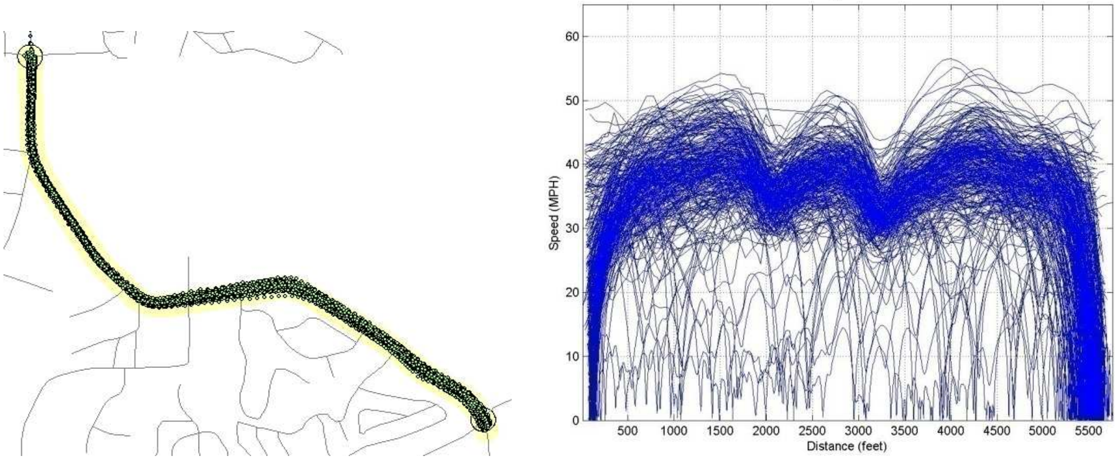
3.1.1 Speed Data

Unlike previous studies that have measured speeds and speed distributions at spot locations the FHWA study and this research focuses on the variability of speeds along stretches of pre-selected corridors using GPS-equipped vehicle speed data. The Commute Atlanta data employed in the studies includes second-by-second vehicle trajectory data collected from January 2004 to December 2004. Each second of Commute Atlanta GPS data contain the trip ID, date, time, vehicle speed, position (latitude-longitude), travel direction, road ID with mile post, and satellite data quality information (used in automated data processing and quality assurance routines). Table 2 summarizes the instrumented vehicle record attributes and their descriptions.

Table 2: Instrumented Vehicle Record Attribute List

Attribute	Description
TRIPID	A combination of Driver ID, Date, Time information
DATE	Date (yyyymmdd)
TIME	Time (hhmmss)
LAT	Latitude
LON	Longitude
SPEED	Speed (mph)
HEAD	Azimuth, angle between north and heading directions (degree)
SAT	Number of satellites
PDOP	Position dilution of precision
RCLINK	Road classification link number (i.e. unique identifier assigned by the state department of transportation to all roadways.)
BEG_MP	Beginning mile point

Using Geographic Information System (GIS) routines, the second-by-second vehicle position data are overlaid on a GIS map and linked to the roadway design and operating parameters (such as speed limit, lane width, curvature, etc.). Figure 4 (a) illustrates a plot of GPS vehicle location, study segment 35, on a second by second basis. Each line in Figure 4 (b) represents the speed trace of an individual trip, from the starting point to the ending point of a corridor.



(a)

(b)

Figure 4: GPS Speed Data of Study Segment 35 Northbound (a) Overlaid on the GIS Map and (b) Speed Profile Plot

3.1.2 Crash Data

By analyzing crash history data provided by the Georgia Department of Transportation (GDOT) within a GIS analytical framework, crashes that occurred in the proximity of the selected corridors can be identified. The four-year (2002-2005) average of crash counts for each roadway link is used to minimize potential year-to-year anomalies.

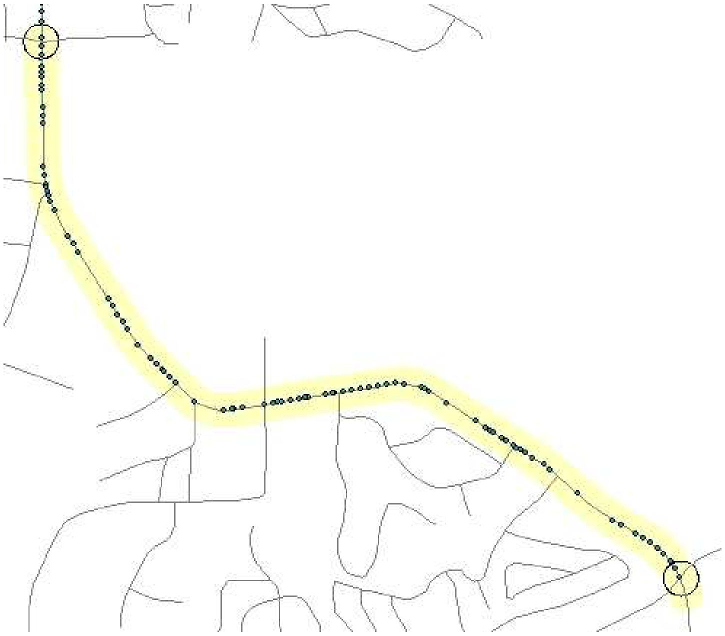


Figure 5: Projection of GDOT Crash Data onto the Road Network

Figure 5 shows the crash data locations along one corridor for the 2002-2005 time period. Hence, crash histories can be linked with spatial speed data, to the extent that crash position data in the database are accurate. Figure 6 shows aggregated crash data by plotting crash counts against the milelog (located at a .01 mile increment, numbered along the corridor) of South Atlanta Street from Azalea Drive to Marietta Highway. Figure 6 illustrates that the number of crashes significantly increases at the signalized intersections at the corridor end points. Since only segmental crashes are of interest in this research effort, crashes that occurred in the intersection vicinity, i.e., within 250-ft radius of the intersection, will be removed from the analysis.

3.1.3 Traffic Data

Traffic count data will be included in the safety prediction model to incorporate the effect of traffic volume on crash involvement. Average Annual Daily Traffic (AADT) data were obtained from the GDOT's State Traffic and Report Statistics (STARS). GDOT conducts annual traffic counts throughout the state roadway system as part of the Highway Performance Monitoring System (HPMS) program. For this study the average of AADT values from 2002 to 2004 were used.

3.1.4 Road Environment Characteristics

Pertinent road characteristics data such as roadway geometry, functional classification, speed limit, land uses, and driveway density were collected from multiple sources such as GDOT's Road Characteristic (RC) file, site visit, aerial maps, etc.

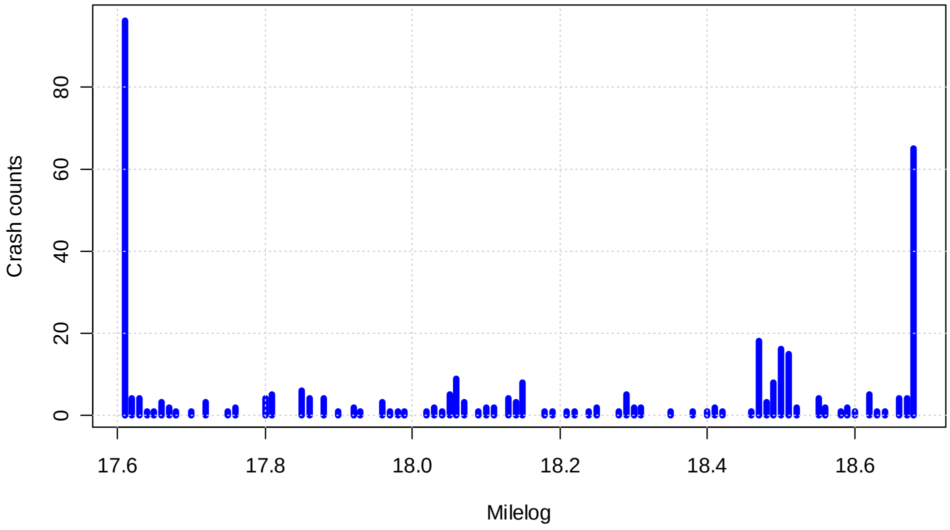


Figure 6: GDOT Crash Data Counts vs. Milelog

3.2 CORRIDOR SELECTION

Since the activity data used in this study were obtained from 460 instrumented-vehicles operating freely, some road link might not have speed data available. This section develops a methodology to collect speed data from the corridors with relative high number of drivers and trips.

3.2.1 Methodology

In designing the FHWA project data analysis plan, two primary aspects were considered when selecting corridors for analysis: 1) maximizing sample size, i.e. the number of drivers and the number of trips, and 2) ensuring that sufficient explanatory variable control would be available in the analysis. That is, sites were selected to ensure a balance in terms of number of road feature types as well as number of trips across different drivers.

To achieve these goals, all corridors within the study area, for which instrumented vehicle data were available, were ranked based on the number of trips being made on each link. Then the top 100 segments in the minor arterial, collector, and local street functional classes were selected. As the focus of the FHWA study was on low- to mid-speed urban facilities, higher functional classes were not included in the final corridor selection. Among top 100 links selected for each functional class, candidate segments were re-ranked by the coefficient of variation (c_v) of the number of trips per driver, i.e., ratio of standard deviation (σ) to the mean (μ):

$$c_v = \frac{\sigma}{\mu}$$

A roadway link with a low coefficient of variation implies that the majority of drivers on this link have similar trip totals, while a high coefficient of variation implies a few drivers accounted for the majority of trips. The object of this ranking was to select those corridors with trip distributed among a higher percentage of drivers.

The RC links were prioritized such that the lower the coefficient of variation, the higher the priority. The corridor prioritization may be inspected visually using GIS software, color-coding the top one-hundred RC links in each road classification based on their coefficient values, i.e., corridors with lower coefficient have a darker color than the ones with higher coefficient. Figure 7 shows a selection of RC links that are included in the top one-hundred lists for the minor arterial (blue), collector street (green), and local street (orange) classifications. In addition, dark color links have higher (i.e., lower c_v) priority than light color links.

To help to ensure that the developed crash prediction model is representative of roadways throughout the Metro Atlanta region the selected corridors were distributed throughout 11 sub-regions of the Metro Atlanta area defined for the FHWA effort. The 11 sub-regions (N1, NE1, SE1, SW1, NW1, N2, NE2, SE2, S, SW2, and NW2) utilize the freeway structure as boundaries (see Figure 8).

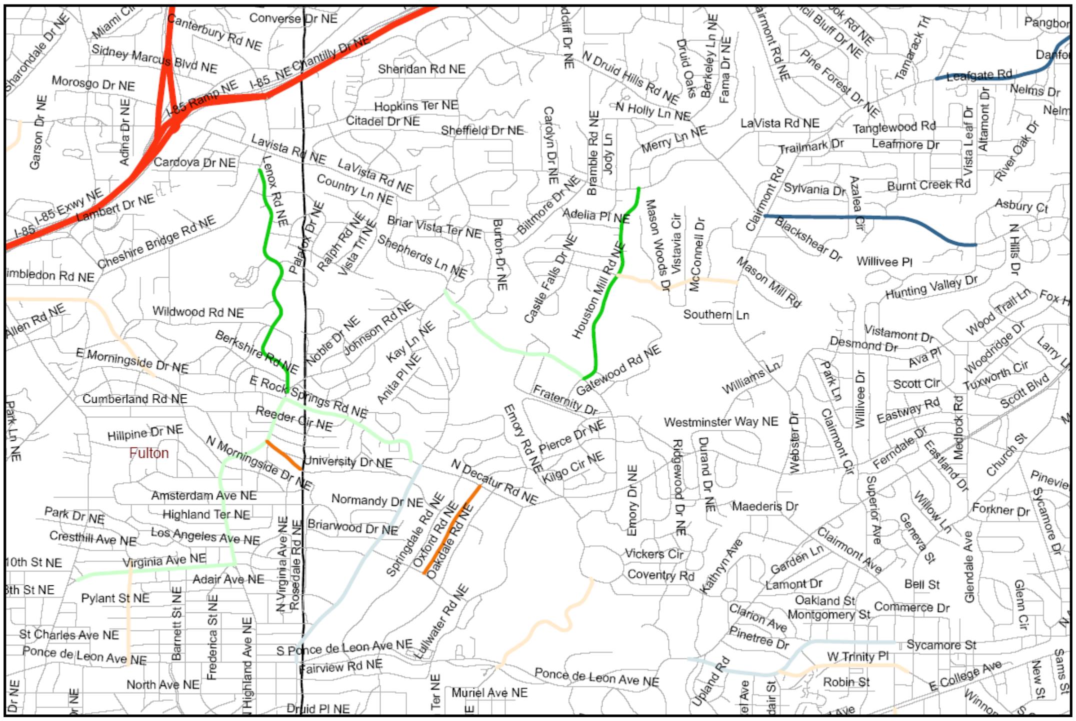


Figure 7: Color-coded RC links

Other corridor selection criteria included: 1) uniform cross-section along the stretch, 2) no mainline traffic control between corridor end points, 3) corridor length greater than 2,000 feet, and 4) speed limit not exceeding 45 mph.

Through visual inspection of the corridors, candidate corridors were selected according to their priority and distribution among the 11 sub regions outlined in corridor section criterion 2. Corridors were also eliminated that did not meet the initial minimum length requirement between traffic control devices. For each selected corridor, design and operations characteristics such as number of driveways, number of side streets, type of end point traffic controls, speed limit, number of lanes, and road functional classification were obtained from field observation.

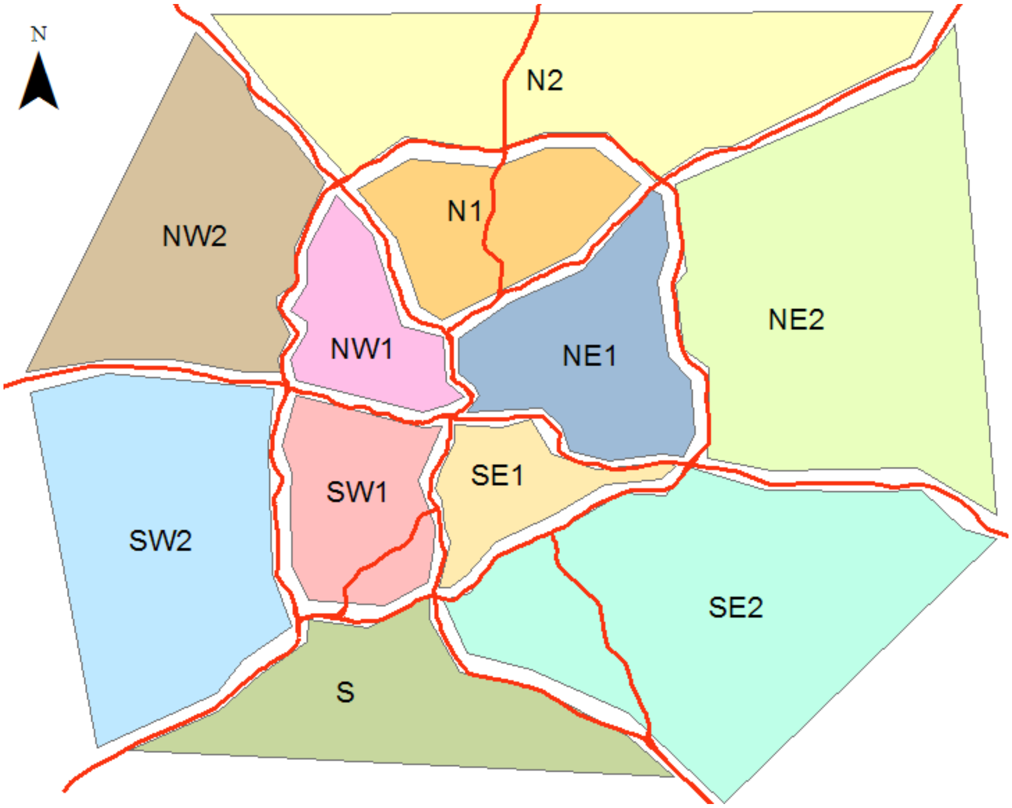


Figure 8: Sub-Area System Map

A total of ninety-three corridors were selected for analysis based upon these criteria. Over the entire 12-month period, a total of 6,661,991 second-by-second data points (roughly equivalent to 1,838 hours of travel time) were collected from 408 drivers. Across the 93 corridors, the average number of drivers per corridor is 56, ranging from 10 to 216 drivers. A total of 77,455 trips were observed across all of the corridors, with each corridor traversed by between 33 and 7,900 trips.

Table 3 shows the demographic information of the participants. It should be noted that there is higher distribution of female drivers in this subset of data than in the Commute

Atlanta database. Distribution of younger drivers also appears to be less than those in the full dataset.

Table 3: Demographic distribution of participants

Age Group	Female		Male		All	
15-24	13	(62%)	8	(38%)	21	(5%)
25-34	33	(62%)	20	(38%)	53	(13%)
35-44	51	(54%)	43	(46%)	94	(23%)
45-54	48	(48%)	51	(52%)	99	(24%)
55-64	52	(58%)	38	(42%)	90	(22%)
65+	22	(43%)	29	(57%)	51	(13%)
Total	219	(54%)	189	(46%)	408	(100%)

3.2.2 Corridor Selection Result

Ninety-two corridors were initially selected for data analysis and modeling. Out of these initial 92 corridors, 33 are Minor Arterials (36%), 32 are Collector Streets (35%), and 27 are Local Streets (29%). Figure 9 and Figure 10 illustrate the distribution of selected corridors. The quantity in the box found in each sub region in Figure 9 indicates the number of selected corridors in that sub region. It is noted that sub-regions SW1 and SE1 are under-represented due to low availability of GPS data in these two sub-regions. This lower availability of GPS data is primarily explained by the sparser density of households in these regions participating in the Commute Atlanta Project (*22*, *40*). The distribution of households (Figure 11) depicts a higher density of the participants in the northern regions that those in the southern region.

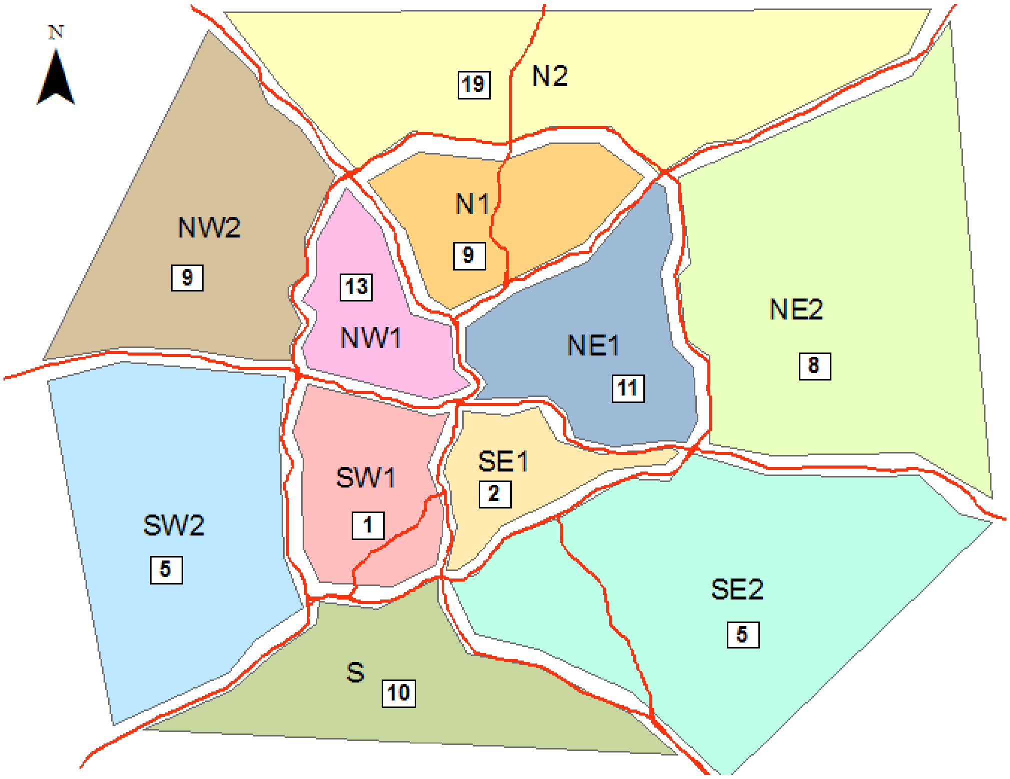


Figure 9: Sub-area System Map with Number of Selected Corridors in each region

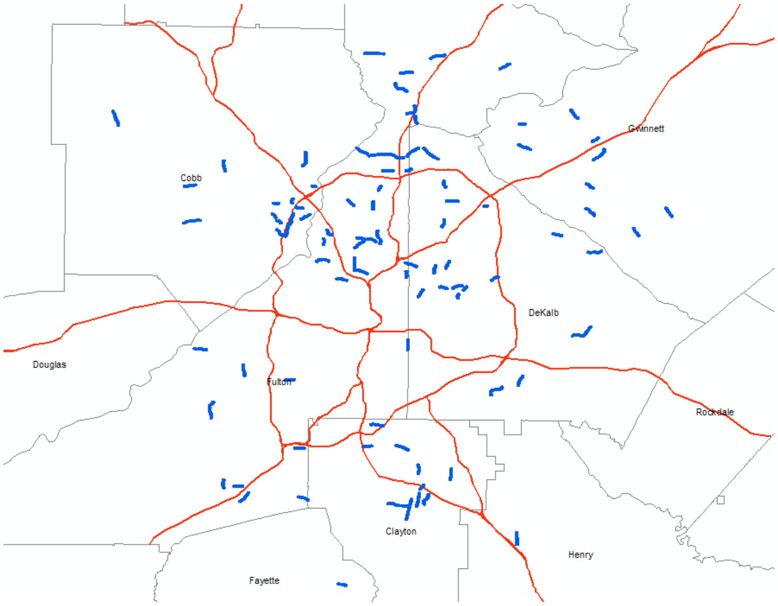


Figure 10: Locations of the 92 Selected Corridors

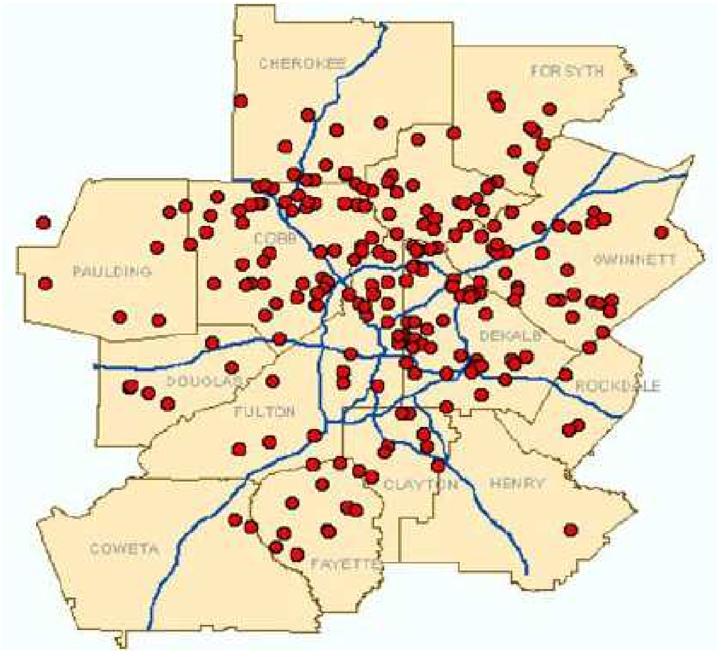


Figure 11: Locations of the Commute Atlanta Project Participating Households (*22*)

3.3 SUMMARY

The data collection and corridor selection methodologies described in this chapter were originally developed for the FHWA project, which concerned the effect of road environment features on operating speed. This study utilized the same GPS dataset, with additional crash data and traffic volume data, to construct a road safety screening tool. Therefore, the data collection and corridor selection methodologies from the FHWA could be adopted for use in this study.

To develop a road safety screening tool, the speed, crash, traffic volume, and road characteristics data are required. The speed data were obtained during the one-year period (2004) from the GPS-instrumented vehicles. The four year period (2002-2005) of crash data were obtained to account for the regression to the mean (RTM) phenomenon characterized by crash data (*31*). The traffic volume of the selected corridors during the same period as crash data were also obtained. Road characteristics data such as speed limit, signalized intersection locations, and road geometric features were obtained during the site visit.

Chapter 4. DATA PROCESSING

With the total of 6,661,991 records of second-by-second GPS data and 1,285,424 crash data records, much of the efforts in this dissertation were spent on processing the data. This chapter described the two key tasks: speed data processing (Section 4.1) and crash data processing (Section 4.2).

4.1 SPEED DATA PROCESSING

The original data processing algorithms were developed as part of the FHWA project “Effects of Urban Street Environment on Operating Speeds” (*6*) and further refined as part of this research effort. Utilizing the FHWA algorithms and refinements in this effort, it is possible to sort the speed data by numerous attributes (i.e., likely free-flow vs. non-free-flow, day vs. night, continuity across the corridor, weather conditions, etc.) to explore potential relationships with safety as will be seen in the following chapters.

A complete list of estimated attributes may be found in Table 2, followed by a discussion of the estimation of each attribute in the data processing section.

Table 4 Estimated Attributes per Vehicle Record

Attribute	Description
GAP	Gap time between the current and previous points (sec)
DIST1	Distance from the current data point to the starting point of the corridor
DIST2	Distance from the current data point to the ending point of the corridor
COMP	Indicator variable, 1 for complete trip, and 0 otherwise
DIR	Direction of travel
LTIME	Local time
NIGHT	Indicator variable, 1 for a trip made during night time, and 0 otherwise
RAIN	Indicator variable, 1 for a trip with raining condition and 0 otherwise
QUEUE	Indicator variable, 1 for a trip with downstream queue longer than 400 ft, and 0 otherwise.
SIGN	Sign of the difference between the current speed and speed filter threshold.
FF1	Indicator variable, a,b,c, or d for free-flow speed type 1 and 0 otherwise.
FF2	Indicator variable, a,b,c, or d for free-flow speed type 2 and 0 otherwise.
ACC	Acceleration value (mph/sec)
CTL	Indicator variable, 1 for a speed point under influence of downstream traffic signal control and 0 otherwise.
DEV	Indicator variable, 1 for a trip with high deviation in speeds and 0 otherwise.
SIGNAL	Indicator variable, 1 for a data point with poor signal quality
PCT80	Indicator variable, 1 when a trip contains at least 80% of good quality data in a trip and 0 otherwise.

Each of the 10 algorithms in this section estimates some trip characteristic for each data record (i.e. every second of instrumented vehicle data is a data record) and appends an associated attribute value to the record. Prior to executing the algorithms, the raw data processed in the Drive Atlanta Lab is sorted into a separate file for each corridor, with the records in each corridor file grouped by driver and sorted by timestamp. The data processing algorithms which are described in the following, are then applied to each corridor data file.

Table 5 illustrates example raw vehicle activity data of driver number one on corridor 00. It is seen that the same driver may traverse the same corridor more than one time during the one-year study period. The different traversals may be distinguished using the time gaps in the data sequence, i.e., one trip was made on January 16th, 2004 and another trip was made on the 19th of the same month.

Table 5: Layout of the Raw Vehicle Activity Data of Driver Number 1 on Corridor 00

DRIVER ID	DATE	TIME	LAT	LON	SPEED	HEAD	SAT	PDOP
DVR_01	20040116	214924	34.030493	-84.248155	41.16	61.65	8	2
DVR_01	20040116	214925	34.030565	-84.24801	41.56	64	8	1.7
DVR_01	20040116	214926	34.030636	-84.247858	42.14	65.83	9	1.5
DVR_01	20040116	214927	34.030704	-84.2477	42.73	67.72	8	1.6
DVR_01	20040116	214928	34.030775	-84.247539	42.95	70.13	8	1.6
DVR_01	20040116	214929	34.030841	-84.247376	43.04	71.79	9	1.5
DVR_01	20040116	214930	34.030911	-84.247211	43.15	72.65	7	2
DVR_01	20040116	214931	34.03098	-84.247048	43.22	73.34	7	2.9
DVR_01	20040119	214958	34.031048	-84.246883	32.88	56.55	6	3.09
DVR_01	20040119	214959	34.031116	-84.24672	34.07	58.16	6	3
DVR_01	20040119	215000	34.031186	-84.246558	35.01	59.53	6	5.19
DVR_01	20040119	215001	34.031258	-84.246395	36.1	61.31	7	2.29

4.1.1 Trip Identification

We define DR_i^d as the i^{th} record for driver d, TR_k^j as the k^{th} record in trip j, and a trip as a period of continuous travel. The objective of the first algorithm is to assign each DR_i^d to a trip, allowing for the identification of each record by trip TR_k^j, in addition to driver

DR_i^d. Each trip over a corridor utilizes a unique trip number, i.e. multiple drivers do not reuse trip IDs.

In the initial portion of the algorithm, a new record attribute GAP is created and appended to each DR_i^d. $GAP(DR_i^d)$ is defined as the time interval between the $i-1^{th}$and i^{th} record of driver d. If $GAP(DR_i^d)$ is greater than ten seconds then record *i* is assigned as the beginning of a new trip. The *Trip Identification* algorithm is executed as follows:

```
Determine Set $J$
initialize $j = 1. k = 1.$
for $d = 1$ to $D$ {
        for $i = 1$ to $I_d$
                $GAP(DR_i^d) = TIME\left(DR_i^d\right) - TIME(DR_{i-1}^d)$
                if $GAP(DR_i^d)$ > 10 seconds {
                        update $k = j$ and $j = j + 1$.
                }
        Set $TR_k^j = DR_i^d$.
        }
}
```

where,
D = the set of all drivers on a corridor
J = the set of all trips on a corridor
K = the set of all records in a trip
I_d = the set of all records for driver d
DR_i^d = the i^{th} record for driver d,
TR_k^j = the k^{th} record in trip j
$TIME(DR_i^d)$ = timestamp of DR_i^d
$GAP(DR_i^d)$ = time interval between $(i-1)^{th}$ and i^{th} record of driver d

4.1.2 Smoothing data using Kalman Filter

While most of GPS receivers, including the SiRF Star II receiver deployed in this study, have an integrated data filtering and smoothing processes to partially mitigate errors in the data stream, some random errors still remain in this dataset (*41, 42*). Quality of speed and location data is critical to the determination of likely free-flow trips. Therefore, to reduce the impact of random errors a modified Kalman filter is utilized. A detailed description of the utilized Kalman filter may be found (*41*).

4.1.3 Trip Continuity

In measuring many of the speed characteristics, it is desirable that a vehicle traverse the entire corridor with no intermediary activity stops. As the data set is a collection of daily trips representative of the many activities individuals undertake uninterrupted traversals are not guaranteed. Vehicles may enter or depart the corridor at internal points, such as a driveway, gas station, etc. For this effort, trips that pass through both corridor boundary intersections are considered complete trips, otherwise the trip is considered incomplete. Recall in the *Trip Identification* algorithm a trip is defined as a continuous second-by-second stream of data (allowing at most a 10 sec gap between records). The impact of this trip definition is that a trip chain, i.e. a driver stopping or diverting along the corridor, would be identified as separate trips in the *Trip Identification* algorithm. Thus, for the *Trip Continuity* algorithm it is only necessary to determine if an individual trip, TR_k^j, passes through both corridor endpoints. For this analysis, a trip is deemed to have passed through an intersection at a corridor end point if the vehicle passes within 100ft of the intersection center.

The algorithm consists of two primary steps: determine the distance from the GPS location of each record to the corridor boundary intersections and check that at least one record location within each trip is within 100ft of the corridor boundary intersections. We manually determined the default orientation of each corridor as south-to-north or west-to-east.

The algorithm is implemented as follows:

Determine distance to corridor boundary intersections.

```
for $j = 1$ to $J$ {
    for $k = 1$ to $K_j$ {
        Calculate $DIST1(TR_k^j)$
        Calculate $DIST2(TR_k^j)$
    }
}
```

Check for passing boundary intersections.

```
for $j = 1$ to $J$
    $COMP(TR_k^j) = \begin{cases} 0 \ \ if \ \left(\left(\min_k DIST1(TR_k^j) > 100\ ft\right) \ or \ \left(\min_k DIST2(TR_k^j) > 100\ ft\right)\right) \\ 1 \ otherwise \end{cases}$
    for all $k$'s in trip $j$.
}
```

where,
$DIST1(TR_k^j)$ = Euclidean distance from point k of trip j to corridor south (east) boundary intersection

$DIST2(TR_k^j)$ = Euclidean distance from point k of trip j to corridor north (west) boundary intersection
$COMP(TR_k^j)$ = Complete trip attribute of trip j: 1 if complete trip, 0 otherwise

Figure 12 shows an example of a complete trip on a study corridor, with a continuous trip that passes through both boundary intersections. Figure 13 illustrates an incomplete trip. This trip (west to east) passed through one boundary corridor but departs the corridor prior to reaching the second boundary intersection.

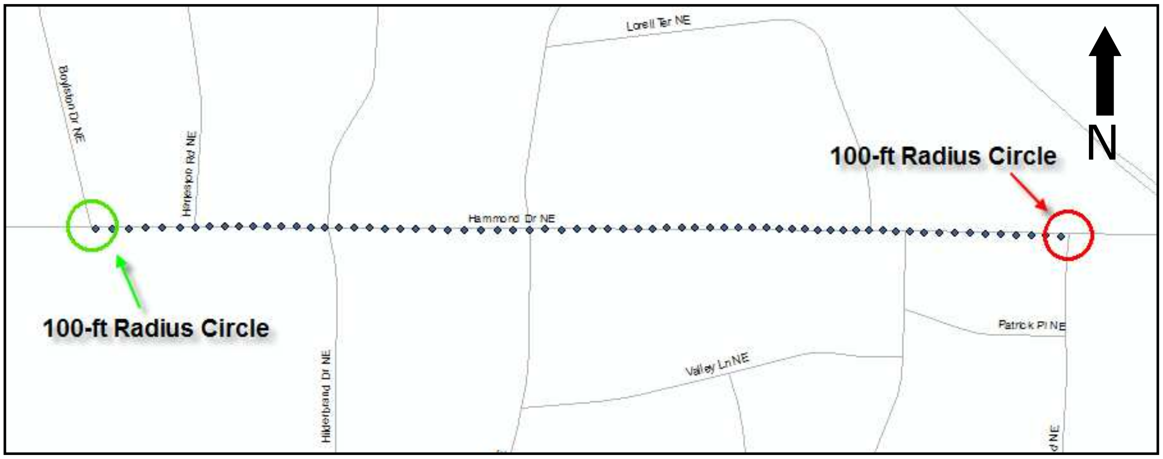


Figure 12: Example of a Complete Trip on Hammond Drive

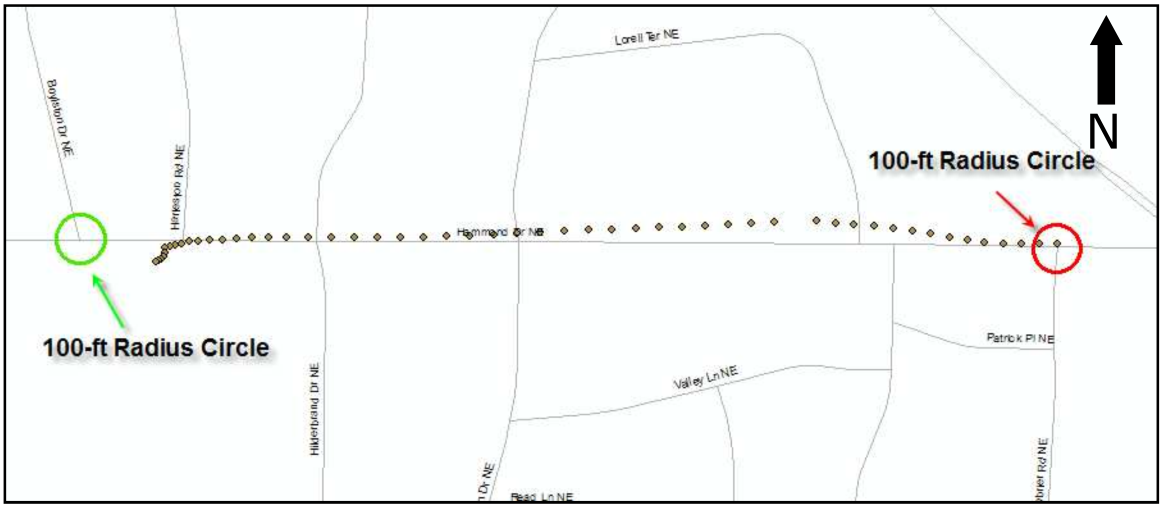


Figure 13: Example of an Incomplete Trip on Hammond Drive

Figure 14 shows the combined effective of the *trip identification* and *trip continuity* algorithms. Shown in Figure 14 is the travel activity of a single vehicle as it crosses the corridor in the westbound direction, leaves the corridor bounds for 385 seconds and then traverses the corridor eastbound. The trip identification initially divides this activity into three trips, two in the westbound direction and one in the east bound direction. The westbound travel is divided into two trips as a result of the 29 second gap between two records. The *trip continuity* algorithm evaluates the three trips, with only the eastbound trip being identified as complete. It is unknown if the westbound vehicle left the corridor, there was an equipment malfunction, or some other reason for the 29 second data gap. Regardless, the data are incomplete and not suitable for the determination of continuous speed characteristics, thus it is desirable that the westbound activity be identified as incomplete. As seen in Figure 13 and Figure 14, the combination of the two algorithms successfully removes vehicle activity that does not represent complete trips or contains incomplete data.

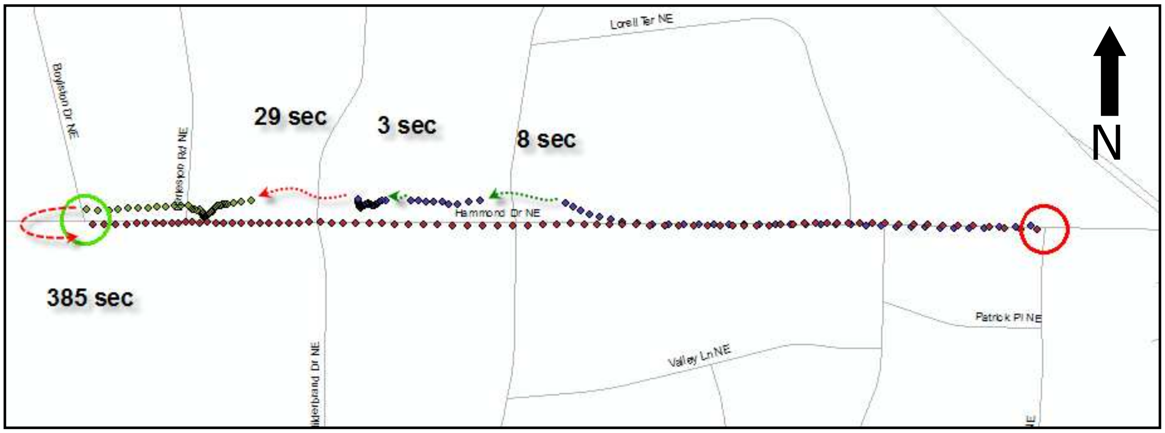


Figure 14: One Entire Trip Separated into Three Sub-Trips

4.1.4 Direction of Travel

In subsequent algorithms, such as the determination of acceleration and decelerations zones (Section 4.1.8), the direction of travel for a trip is required. The *Direction of Travel* algorithm determines the trip heading, and appends an attribute to each trip record with this data. The *Direction of Travel* algorithm compares the distance from the location of the first record of a trip to the south (west) and north (east) boundary intersections. If the starting location is closer to the corridor's south (west) intersection, the direction of travel is northbound (eastbound), otherwise the direction of travel is southbound (westbound).

The algorithm is implemented as follows:

Compare distance to boundary intersections to set heading

```
for $j = 1$ to $J$ {
        if $DIST1(TR_1^j) < DIST2(TR_1^j)$ {
                assign $DIR(TR_k^j)$ = NB (or EB) for all $k$'s in trip $j$
        }
        else{
                assign $DIR(TR_k^j)$ = SB (or WB) for all $k$'s in trip $j$
        }
}
```

where:
$DIR(TR_k^j)$ = the direction of travel of trip j

4.1.5 Local Time and Nighttime

Lighting conditions may influence a driver's speed. Therefore, each trip is assigned an attribute identifying whether it occurs during the day or night, allowing for a sorting of trips by lighting conditions prior to the development of speed models, if desired. A trip is considered a nighttime trip if it starts before sunrise or after sunset. Since the sunrise and sunset time varies significantly throughout the year, sunrise and sunset times specific to each trip were calculated. A Sun altitude of -0.833 degrees is chosen in the determination of sunrise/sunset as it is the position where the upper edge of the disk of the Sun touches the earth's horizon, accounting for atmospheric refraction. We also adjusted the calculated sunrise and sunset times by adding a 30-minute buffer to the sunrise time and subtracting a 30-minute buffer from the sunset time. For this dataset, approximately 23 percent of the complete trips were identified as occurring during nighttime.

The algorithm for identifying whether a trip occurred during nighttime or daytime contains two primary steps. First, it is necessary to create a trip attribute with the local time (Eastern Standard Time for this study), as the GPS timestamps are recorded in

Greenwich Mean Time (GMT). The calculated sunrise/set time are then compared to the trip start time.

The algorithm is a follows:

Calculate local time

```
for $j = 1$ to $J$ {
        for $k = 1$ to $K$
                $LTIME(TR_k^j)$ = local time (EST) of $TIME(TR_k^j)$
        }
}
```

Daytime or nighttime determination

```
for $j = 1$ to $J$ {
        $RISE$ = sunrise time of $DATE(TR_k^j)$
$SET$ = sunset time of $DATE(TR_k^j)$

for all $k$'s in trip $j$:

                $NIGHT(TR_k^j) = \begin{cases} 0 & if\ \ RISE + 30\min < LTIME(TR_1^j) < SET - 30\min \\ 1 & otherwise \end{cases}$

        }
}
```

where:
$LTIME(TR_k^j)$ = Local time (EST) for record k of trip j
$DATE(TR_k^j)$ = Date for record k of trip j
$RISE$ = Local sunrise time for $DATE(TR_k^j)$
SET = Local sunset time for $DATE(TR_k^j)$
$NIGHT(TR_k^j)$= attribute of record k of trip j indicating daytime or nighttime at time of record

4.1.6 Inclement Weather Conditions

Inclement weather may influence a driver’s speed. This step detects trips that likely occurred during rain conditions (snow/ice conditions were not observed during the study period). The determination of potential inclement weather during a trip is based on the hourly precipitation data from several weather stations in Metro Atlanta. These weather stations are located at the Fulton County Airport, DeKalb-Peachtree Airport, and Hartsfield-Jackson Atlanta Airport (see Figure 15). A trip is considered to have likely occurred under inclement weather conditions if measurable rainfall is recorded at the two closest stations during the 2-hour time window before the trip. This rule identified approximately 20 percent of the daytime complete trips as occurring during potential inclement weather conditions. While a portion of the trips identified as inclement weather trips likely did not experience inclement weather, it was decided to implement a conservative rule for trips utilized in the speed model development. However, given this rule it should be noted that it would be inappropriate to create a set of “inclement weather” speed models using trips identified as inclement weather trips, as many of these trips may have occurred under clear weather conditions.

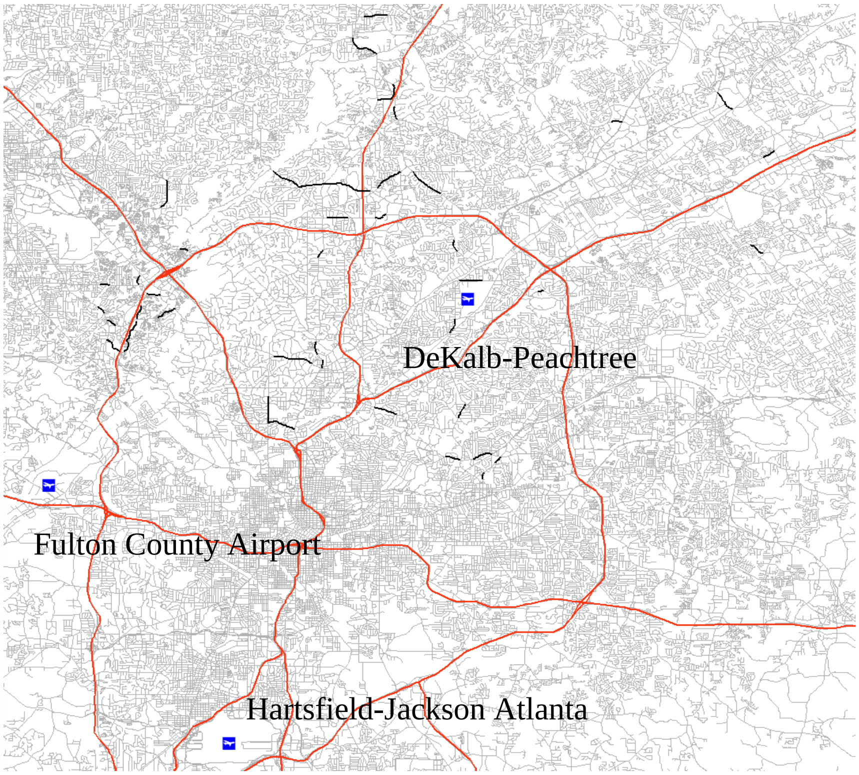


Figure 15: Locations of the Three Weather Stations (Identified by Blue Indication)

The algorithm is implemented as follows:

Determine the two closest weather stations to the first point of trip *j*. Let *w* be weather station ID's from 1 to 3.

```
for $j$ = 1 to $J$ {
    for $w$ = 1 to 3{
            $DIST(w)$ = distance from the station w to the first point in trip $j$, $TR_1^j$
    }

    initialize sum = 0.
    for $w'$ = 1' to 2'{
        for t = ($LTIME(TR_k^j)$ – 2 hours) to $LTIME(TR_k^j)${
                $PRECIP(w', t)$ = hourly precipitation amount at station $w'$, time $t$.
            $sum = sum + PRECIP(w', t)$
        }
    }
```

```
        $RAIN(TR_k^j) = \begin{cases} 1 & if\ \ sum > 0 \\ 0 & otherwise \end{cases}$
        for all $k$'s in trip $j$
}
```

Note that 1’ and 2’ are the ID’s of the two closest stations to trip *j*. It follows that $DIST(1') < DIST(2') < DIST(3')$.

where:
w = Local time (EST) for record *k* of trip *j*
$DATE(TR_1^j)$ = Date for record *k* of trip *j*
$RISE$ = Local sunrise time for $DATE(TR_1^j)$
SET = Local sunset time for $DATE(TR_1^j)$
$NIGHT(TR_k^j)$ = attribute of record *k* of trip *j* indicating daytime or nighttime at time of record

4.1.7 Identify Potentially Non-Free-Flow Trips

In the next algorithm, we apply a series of developed filters is applied that utilize the characteristics of a trips GPS trajectory data to help identify complete trips that were likely non-free-flow trips. As a first step in developing these filters, a Graphic User Interface (GUI) application, called the GPS Speed Profile Viewer, was constructed. This application plots the speed profiles – distance (feet) from the corridor starting point (X-axis) versus vehicle speed in mph (Y-axis) – for all trips, or trips during a user selectable time period, that occurred on a corridor. For example, Figure 16 depicts the speed profile of westbound trips on Corridor No. 20, Hammond Drive, between Perimeter Center Parkway and Peachtree Dunwoody Road.

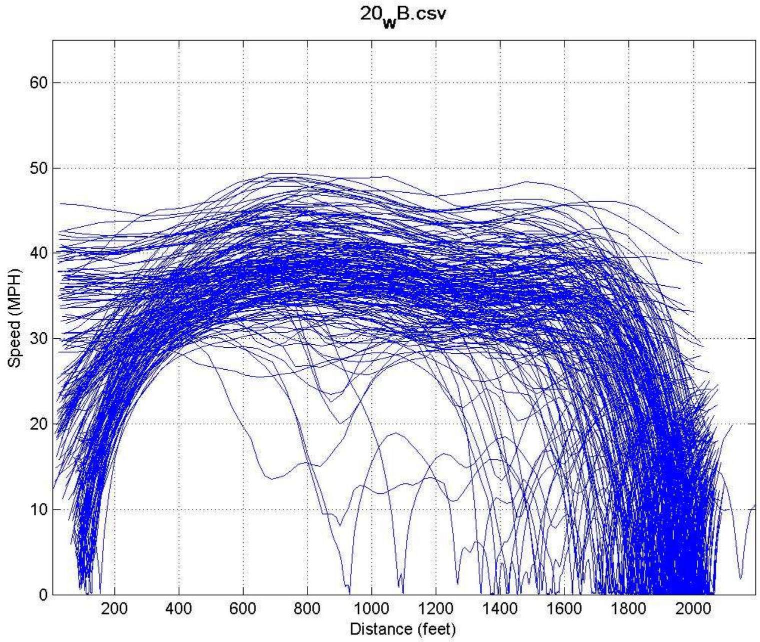


Figure 16: Example Speed Plot using the Speed Profile Viewer

Figure 16 illustrates that a number of vehicles may have stopped or significantly slowed in the corridor mid-section during their trips. From the graph, it is clearly seen that these stopped and slowed vehicle trips are not in free-flow operation. Each of the filters developed to identify these trips is summarized in the following sections.

Initially, a time-of-day filter was considered, defining peak period traffic as non-free-flow and non-peak period traffic as free-flow. However, upon inspection of the candidate corridors several irregular peak hour periods in commercial and warehouse districts were identified as well as and commonly accepted non-peak hours clearly exhibiting non-free-flow trip characteristics. Thus, time-of-travel based filters did not adequately identify likely non-free-flow or likely free-flow trips. To overcome the peak time based filter drawbacks, a combination of filters based on trip characteristics was developed. These are 1) downstream queue filter; 2) fixed speed filter with free-flow pattern recognition, and; 3) variable speed filter with free-flow pattern recognition. The combination of these filters successfully identified the peak and non-peak hour trips that did not exhibit free-flow behavior, in essence enabling the use of variable peak hours with respect to the individual corridors.

A) Downstream Queue Filter.

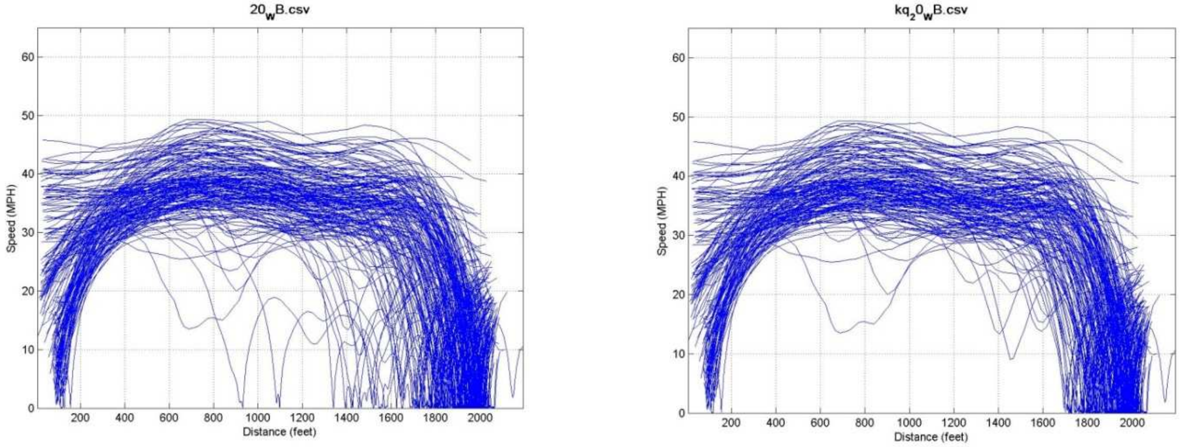
As shown in Figure 16, vehicles are often captured in a queue at the downstream end of a trip. When the stopping location of the vehicle indicates a significant queue length the vehicle should not be assumed as free-flowing on the upstream portion of the corridor, as a lengthy queue characterizes likely congested conditions. A vehicle stopping more than 400-feet upstream of the center of the trip end intersection was selected as the queue length at which free-flow travel was no longer likely. The downstream queue filter identifies trips that have a speed lower than 5 mph in the range from the mid-point of the corridor to 400 feet from the downstream location. The 400-feet value was selected following a pattern recognition and sensitivity analysis on the corridors. Initially, a separate set back value for each functional classification was investigated, however, the 400-ft value conservatively identifies queued vehicles for all locations and subsequent filters capture other irregular trips.

Figure 17 illustrates the effect of removing trips identified by the Downstream Queue filter. For our data set approximately 6 percent of the daytime, non-inclement weather, complete trips were identified using this filter.

The algorithm is implemented as follows:

```
Check the distance from record $k$ in trip $j$ to the ending point
for $j = 1$ to $J$ {
        for $k = 1$ to $K_j${
            $QUEUE(TR_k^j) = \begin{cases} 1, if\ (DIST2(TR_k^j) \geq 400\ \ ft)\ AND\ SPEED(TR_k^j) < 5\ mph \\ 0, otherwise \end{cases}$

        }
}
```



(a)

(b)

Figure 17: Trip Speeds (a) Before and (b) After Applying Downstream Queue Filter

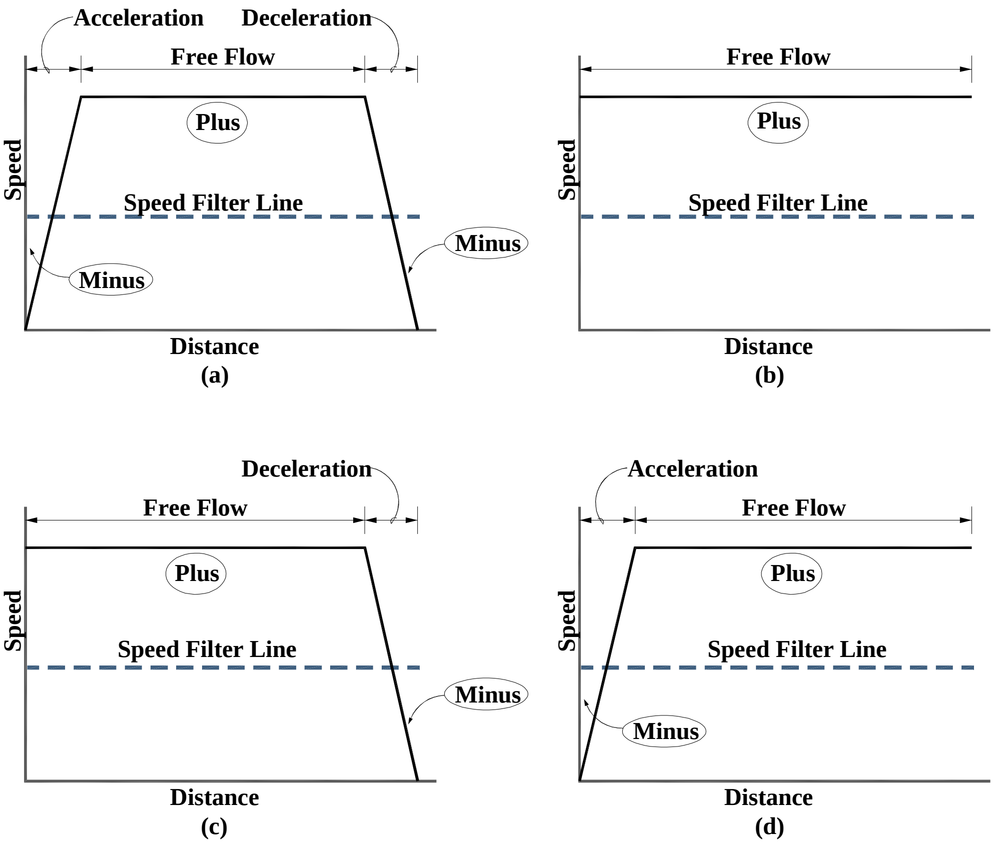


Figure 18: Four Speed Patterns Defining Potential Free-Flow Speed Trips (Figure Credit: Karen Dixon, Oregon State University)

B) Fixed Speed Filter with Free-Flow Pattern Recognition:

Figure 18 represents potential simplified free-flow speed profile trip patterns across a corridor. Occasionally a vehicle may encounter a “non-free-flow” pattern as depicted in Figure 19. Patterns similar to those seem in Figure 19 may occur, for example, when a study vehicle is trailing a vehicle that reduces speed to execute a turn mid-block. To identify trips that are clearly not free-flow due to this phenomenon, it is necessary to identify trips that experienced speeds less than a fixed cutoff value, 10 mph in this step, not located in the corridor boundary acceleration or deceleration zones.

Through visual inspection identifying a trip that violates this rule is a relatively simple matter. However, due to the large number of trips and the desire to test the sensitivity of overall trip loss to the cutoff value an efficient automated implementation of the rule was desired. This was accomplished through the use of a pattern recognition approach where a negative sign represents speeds less than the designated cutoff speed and a positive sign speeds greater than the designated cutoff speed.

The first step in the pattern recognition checks the pattern sign at the mid-point. For a trip to be considered likely free-flow through the corridor, it is assumed that the vehicle must be traveling at free-flow speed by the corridor mid-point. Any trip with a negative pattern sign (i.e., speed less than a pre-determined cutoff value) at the mid-point may safely be assumed to not be traveling at free-flow speed.

Next, the pattern recognition algorithm considers speed data in the area starting from the upstream intersection to 400 feet before the end of the corridor. The 400-foot area defined as a queuing zone is excluded from the pattern recognition since stop locations of vehicles in the downstream queue can be varied. All possible free-flow patterns are depicted in Figure 18 . The pattern recognition algorithm compares an individual trip with each of the four pre-defined free-flow patterns. If the trip does not match with any of the free-flow pattern, it will then be determined as a likely non-free-flow trip. More particularly, the algorithm starts with splitting a trip into two halves at the midpoint and considering speed pattern of each half at a time. The algorithm then finds the matching pattern from the pre-defined patterns below:

- If the first half consists of exactly one sign change from minus to plus and the second half consists of exactly one sign change from plus to minus, the trip matches the free-flow pattern in Figure 18a. The notation for pattern (a) is (–+, +–).

- If a trip does not have any sign change in both halves and has a speed above the cutoff, it matches the free-flow pattern in Figure 18b. The notation for pattern (b) is (++, ++).

- If the first half does not have any sign change and has a speed above the cutoff, and the second half consists of one sign change from plus to minus, the trip matches the free-flow pattern in Figure 18c. The notation for pattern (c) is (++, +-).

- If the first half consists of one sign change from minus to plus and the second half does not have any sign change, the trip matches the free-flow pattern in Figure 18d. The notation for pattern (d) is (-+, ++).

- If a trip does not match any of the above, it is identified as a likely non-free-flow trip.

Examples of non-free-flow trips are depicted in Figure 19. For instance, if a vehicle enters the corridor after being stopped (i.e., the vehicle was stopped at a signal light or stop sign) and accelerates to a speed greater than 10 mph by the time it reaches the midpoint of the corridor, the algorithm would recognize the sign change in the first half of this trip as one sign change from minus to plus. Furthermore, if the same vehicle decelerates and stops due to a signal light or stop sign at the downstream intersection, the algorithm would recognize the speed pattern in the second half as a one sign change from plus to minus. Finally, the algorithm combines the information from the first and second halves and determines this trip as a free-flow trip pattern (a). Now, if the same trip has an additional change from positive to negative to positive - representing vehicle deceleration to a speed below 10 mph and then acceleration to a speed above 10 mph - the trip is identified as a non-free-flow trip and is removed from the free-flow data set.

At the conclusion of the fixed speed filter, approximately eight percent of the trips not identified as likely non-free-flow by the downstream queue filter were identified as likely non-free-flow trips by this filter.

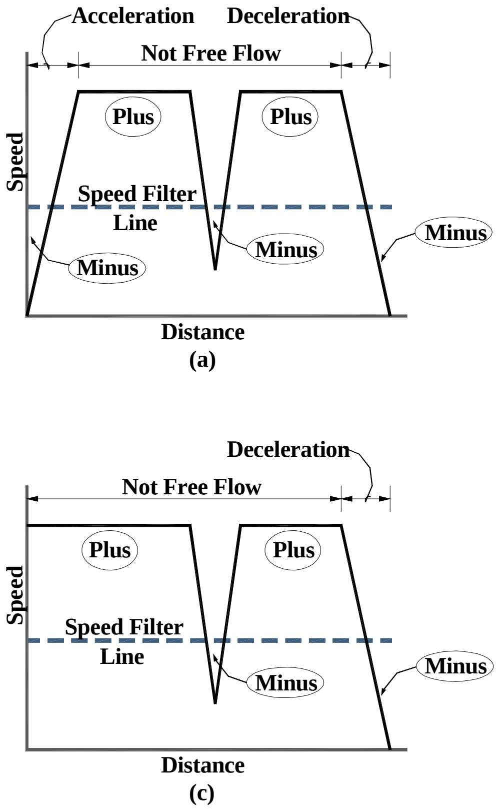
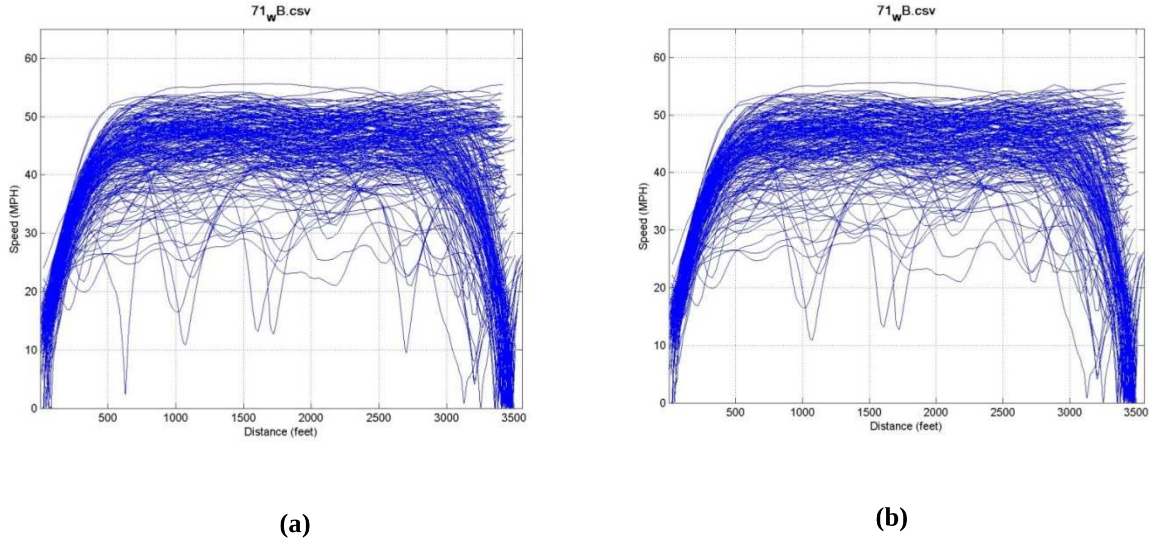


Figure 19: Four Speed Patterns Defining Potential Non-Free-Flow Speed Trips (Figure Credit: Karen Dixon, Oregon State University)



(a)

(b)

Figure 20: Speeds (a) Before and (b) After Applying 10-mph Filter

The algorithm is executed as follows:

```
Compare GPS speeds with 10 mph threshold
for $j = 1$ to $J$ {
    for $k = 1$ to $K_j${
            $SIGN(TR_k^j) = \begin{cases} +1 & if\ SPEED(TR_k^j) \geq 10 \\ -1 & otherwise \end{cases}$
    }
}

Determine pattern parameters
for $j = 1$ to $J$ {
    initialize changes_front = 0 and changes_back = 0
    for $k = 1$ to $K_j - 1${
        if $DIST1(TR_k^j) \leq$ L/2{
            if $SIGN(TR_k^j) \neq SIGN(TR_{k+1}^j)${
                from_front = $SIGN(TR_k^j)$
                to_front = $SIGN(TR_{k+1}^j)$
                changes_front = changes_front + 1
            }
        }
        else{
            if $SIGN(TR_k^j) \neq SIGN(TR_{k+1}^j)${
                from_back = $SIGN(TR_k^j)$
                to_back = $SIGN(TR_{k+1}^j)$
                changes_back = changes_back + 1
            }
        }
        if $DIST1(TR_k^j) = L/2${
            midSpeed = $SPEED(TR_k^j)$
        }
    }
}
```

Assign trip pattern (FF1) as defined in Table 6. Trip patterns that are not defined by this table will be marked as non-free-flow trips.

Table 6: Potential Free-Flow (FF1) Speed patterns

midSpeed	changes_front	from_front	to_front	changes_back	from_back	to_back	FF1
>10	0	1	1	0	1	1	c
				1	-1	1	d
	1	-1	1	0	1	1	b
				1	1	-1	a
Else							0

Note that L = corridor length

changes_front = number of sign changes within the upstream half of the corridor

changes_back = number of sign changes within the downstream half of the corridor

from_front = sign notation before sign change in the upstream half of the corridor

to_front = sign notation after sign change in the upstream half of the corridor

from_back = sign notation before sign change in the downstream half of the corridor

to_back = sign notation after sign change in the downstream half of the corridor

C) Variable speed filter with Free-Flow Pattern Recognition

The previous step removed trips that contain speeds less than 10 mph not located in the acceleration and deceleration zones. However, from Figure 20b, it is noticeable that some remaining trips are still under likely non-free-flow conditions. The ten-mph speed filter line works well on low speed roadways, e.g., 25 mph posted speed limit, but frequently fails to catch non-free-flow trips on a higher speed facilities, e.g., 40 or 45 mph posted speed limit. As a result, a variable speed filter is developed to identify the non-free-flow trips remaining not identified by the ten-mph filter. The variable free-flow speed cutoff on each corridor is dependent on each corridor's speed limit and driver's mean speed, as will be seen in the following section. A sensitivity analysis was performed to determine an approximate minimum value for free-flow speed conditions. It is again assumed that by the corridor mid-point a vehicle should be able to achieve free-flow speed.

Mid-Point Free-Flow Speed Determination

This analysis investigated several potential variable cutoff speeds to identify trips not at free-flow speed at the corridor mid point:

- Speed limit-10 mph
- Seventy percent of the drivers' mean speed at the corridor midpoint
- Seventy percent of the speed limit
- Seventy percent of minimum(driver's mean speed, speed limit)

Though all trips are depicted in the speed profile plots, at several sites many of the trips were unique to one driver. As a result, the analysis used the average speed per driver to estimate the mean speed for Options 2 and 4. The following algorithm utilizes the forth condition – seventy percent of minimum between driver's mean speed and speed limit. However, each of the given cutoffs was investigated by changing the threshold value in the following algorithm. The forth condition was ultimately chosen as it seemed to provide the best results for the given dataset.

The algorithm is implemented as follows:

Calculate mean speed among drivers at midpoint of the corridor

$$meanSp = \frac{1}{\mathrm{D}}\sum_{\mathrm{d}=1}^{\mathrm{D}}\frac{1}{\mathrm{J}}\sum_{\mathrm{j}=1}^{\mathrm{J_d}}\mathrm{SPEED}\left(\mathrm{TR}_{\mathrm{mid}}^{\mathrm{j}}\right)$$

where: $SPEED(TR_{mid}^{j})$ = speed at midpoint of trip j.
J_d = number of trips made by driver d
D = total number of drivers in the corridor

Determine the variable speed threshold, *trshld.*
trshld = 0.7*min(meanSp, speed limit)

Compare GPS speeds with the variable speed threshold

```
for j = 1 to J {
    for k = 1 to K_j{
        $SIGN(TR_k^j) = \begin{cases} +1 & if\ SPEED(TR_k^j) \geq trshld \\ -1 & otherwise \end{cases}$
    }
}
```

Determine pattern parameters (similar to fixed speed filter)

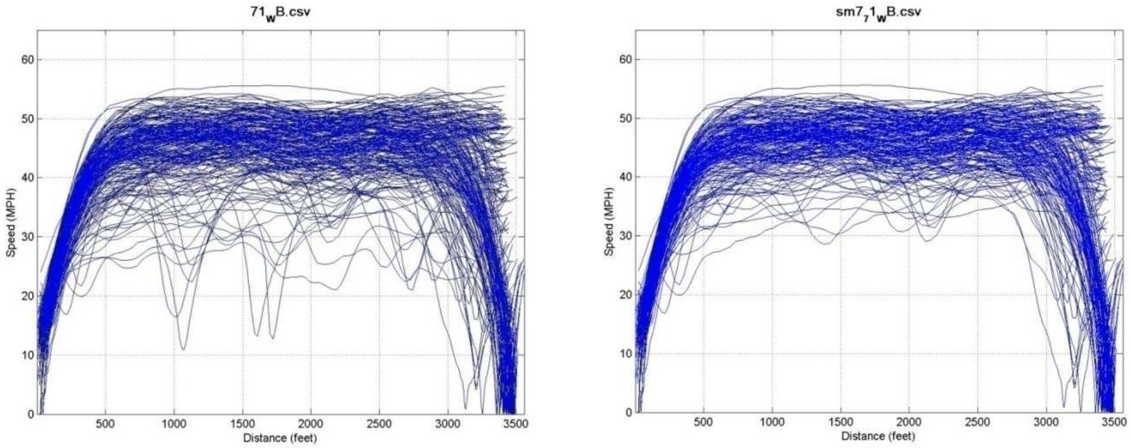
```
for j = 1 to J {
    initialize changes_front = 0 and changes_back = 0
    for k = 1 to K_j-1{
        if $DIST1(TR_k^j) \leq$ L/2{
            if SIGN $(TR_k^j) \neq$ SIGN$(TR_{k+1}^j)${
                from_front = $SIGN(TR_k^j)$
                to_front = SIGN$(TR_{k+1}^j)$
                changes_front = changes_front + 1
            }
        }
        else{
            if SIGN$(TR_k^j) \neq$ SIGN$(TR_{k+1}^j)${
                from_back = SIGN$(TR_k^j)$
                to_back = SIGN$(TR_{k+1}^j)$
                changes_back = changes_back + 1
            }
```

```
                }
                if $DIST1(TR_k^j)$ = L/2{ # check speed at mid point
                        midSpeed = SPEED($TR_k^j$)
                }
        }
}
```

Assign trip pattern (FF2) as defined in Table 7. Trips patterns that are not defined in this table will be marked as likely non-free-flow trips.

Table 7: Potential Free-Flow (FF2) Speed patterns

midSpeed	changes_front	from_front	to_front	changes_back	from_back	to_back	FF2
>trshld	0	NA	NA	0	NA	NA	c
				1	-1	1	d
	1	-1	1	0	NA	NA	b
				1	1	-1	a
Else							0



(a)

(b)

Figure 21: Speeds (a) Before and (b) After Varied Free-Flow Speed Filter

4.1.8 Removing data points in the acceleration and deceleration zones

To obtain free-flow speed conditions, the zones containing acceleration or deceleration effects from the traffic control at the two ends of corridors is determined. This step trims the instrumented vehicle trajectories, removing the data points within the acceleration and deceleration zones. To implement this filter it is necessary to determine the deceleration and acceleration zone distances. A combination of speed and acceleration values was used to detect acceleration and deceleration motions of vehicles due to traffic control. The acceleration value of data point *k*, ACC(TR_k^j) was calculated using the central difference formula,

$$ACC(TR_k^j) = \frac{SPEED(TR_{k+1}^j) - SPEED(TR_{k-1}^j)}{TIME(TR_{k+1}^j) - TIME(TR_{k-1}^j)}$$

The central difference formula is widely used to calculate acceleration and it provides higher accuracy than those derived from forward or backward difference formulas (*43*). Deceleration and acceleration lengths were determined separately as follows:

<u>A) Deceleration zone determination:</u>

In this study, the deceleration zone is defined as the zone in which at least 90 percent of the individual trips begin to decelerate due to traffic control (e.g., stop sign, traffic signal) at the corridor end point. In addition, we defined the start of a vehicle's deceleration process as the location at which its deceleration is greater than 1 mph/sec. The 1 mph/sec deceleration rate accounts for the normal fluctuation in driving behavior during free-flow conditions and GPS receiver error. Finally, the deceleration zone is taken to end at the downstream intersection of the corridor.

The algorithm for the deceleration filter is below:

- Determine midpoint of the corridor, midpoint = L/2.
- For each trip, start from the midpoint moving toward the end of corridor to find the first instance that the trip's speed is less than 10 mph (See Figure 22).
- Starting from the location in the previous step, move upstream and identify the first instance that that the vehicle's deceleration becomes less than 1 mph/sec (Figure 23). This point is the starting location of deceleration due to traffic control for the trip.
- Repeat these steps for every trip in the corridor to obtain a list of deceleration starting locations. After obtaining starting locations for every trip, the locations (distance from downstream intersection) are sorted from high to low.
- Find the 90th percentile location from the list. This is defined as the start of the deceleration zone for the corridor (See Figure 24).

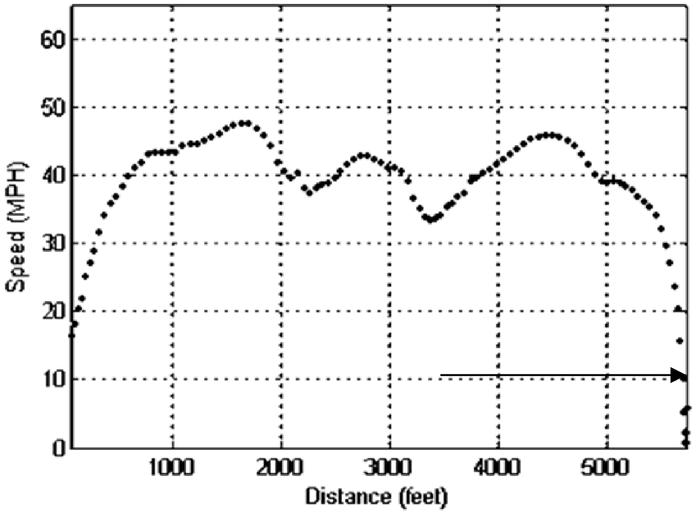


Figure 22: Forward Motion to Identify the First Speed Point Dropping below 10 mph

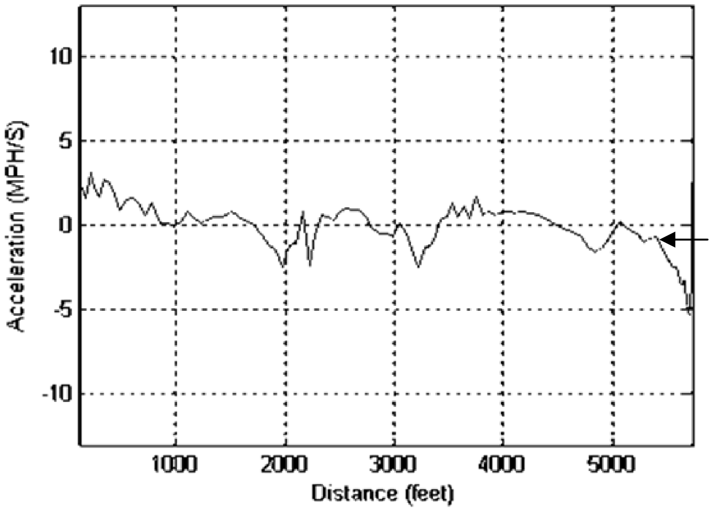


Figure 23: Reverse Motion to Find the Initial Deceleration Location < 1 mph/sec

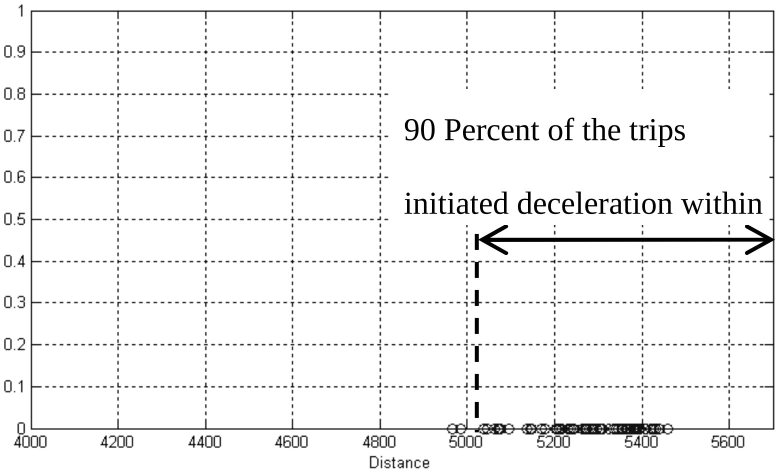


Figure 24: Example Plots of Deceleration Points for a Corridor's Trips

The algorithm is implemented as follows:

```
Calculate acceleration rate for each data point
for $j = 1$ to $J$ {
        for $k = 2$ to $K - 1$
                $ACC(TR_k^j) = \frac{SPEED(TR_{k+1}^j) - SPEED(TR_{k-1}^j)}{TIME(TR_{k+1}^j) - TIME(TR_{k-1}^j)}$
        }
}

find the starting point of deceleration of each trip
lower bound speed, $lowSp$ = 10 mph
for $j = 1$ to $J$ {
        for $k = 1$ to $K$ {
        find min $k$' where $(SPEED(TR_{k'}^j) < lowSp)\ AND\ (DIST2(TR_{k'}^j) < L/2)$
                for $k = k$' to 1 {
                        find $k$" where ACC($TR_{k''}^j$) $\leq$ -1
                        store $DIST2(TR_{k''}^j)$ into $decArray$
                }
        }
}
```

$decDist$ = the 90th percentile of distance in the $decArray$

Determine whether a data point is located within the deceleration zone.

```
for $j = 1$ to $J$ {
    for $k = 1$ to $K_j${
        CTL($TR_k^j$) = $\begin{cases} 1, if\ \ DIST2(TR_k^j) \le decDist \\ 0, otherwise \end{cases}$
    }
}
```

<u>B) Acceleration zone determination:</u>

The acceleration zone is defined as the zone starting at the upstream intersection to the location where 90 percent of the trips begin to operate at cruising speed. As with the deceleration, a trip is assumed to begin operating at cruising speed when acceleration rate drops to less than 1 mph/sec.

Algorithms:

- Determine midpoint of the corridor, midpoint = L/2.

- Based on speed and distance data for a vehicle trip, start from the midpoint moving toward the starting point of corridor to find the first location that a vehicle's speed is less than the cutoff speed line, which is defined as the minimum between speed limit minus 10 mph and 25 mph. For example, if speed limit is 30 mph, the lower bound speed line will be 20 mph and if speed limit is 45 mph, the cutoff speed line will be 25 mph (Figure 25).

- Based on acceleration and distance data, move toward the end of the corridor and search for the first location at which the vehicle's acceleration rate drops below 1 mph (Figure 26).

- Repeat these steps for every trip traversing the same corridor to obtain a list of the trip acceleration ending locations. After obtaining acceleration ending locations for every trip, the locations (distance from upstream intersection) are sorted from low to high.

- Find the 90th percentile location from the list and assign this as the acceleration zone end point location.

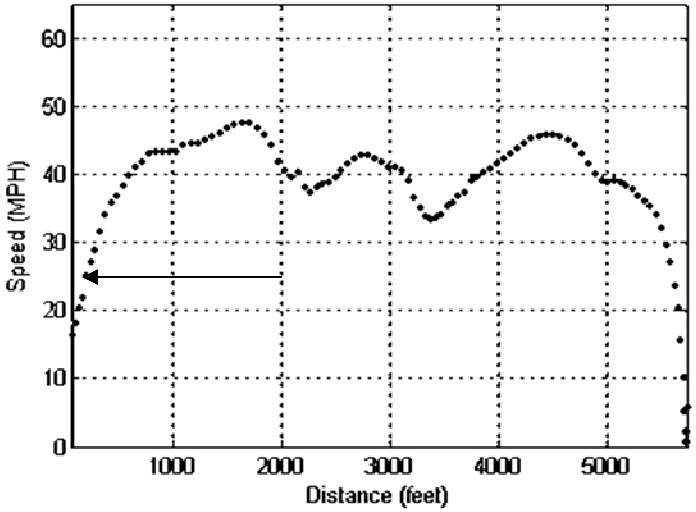


Figure 25: Reverse Motion to Find First Point below Lower Bound Speed Threshold (Corridor ID 35 NB, lower bound line = 25 mph)

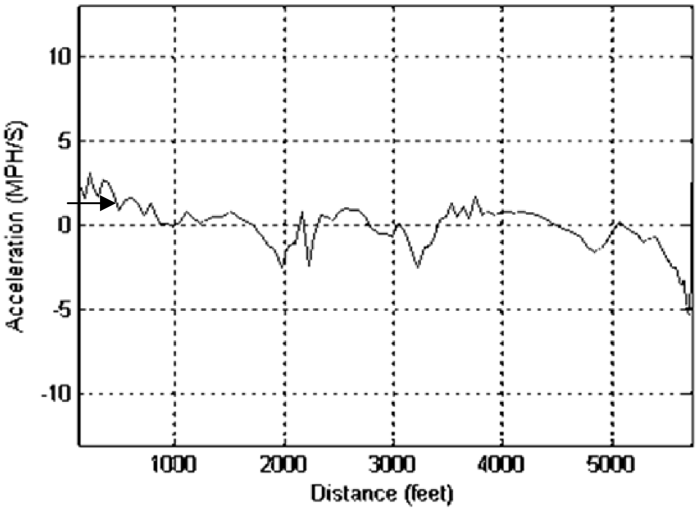


Figure 26: Forward Motion to Find First Location with Acceleration Rate < 1 mph

Note that the lower bound speed lines between acceleration zone and deceleration zone are based on different criteria because these two cases have different traffic characteristics. For deceleration zone case, it is likely to see vehicles slow down to speeds below 10 mph within the corridor boundary. On the other hand , in the acceleration zone, it is likely that vehicles starting their acceleration process from a location upstream of the corridor boundary, e.g., starting from the middle of the upstream traffic queue or starting from the stop bar where the cross street is four-lane road with a median. Therefore, these vehicles may well have a speed greater than 10 mph prior to entering to corridor however still be undergoing the acceleration process.

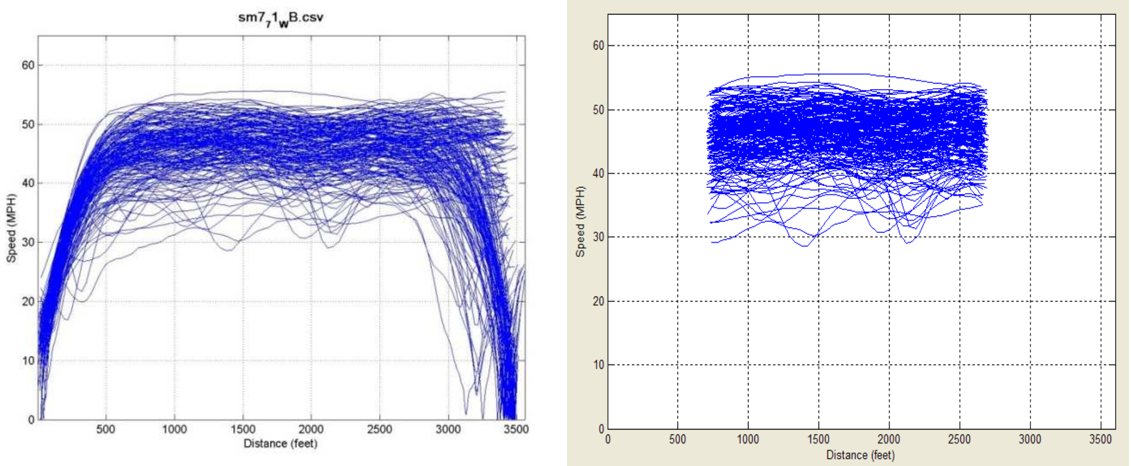
The algorithm is implemented as follows:

```
find the starting point of stable speed of each trip
lower bound speed, $lowSp$ = min(speed limit-10,25)
for $j = 1$ to $J$ {
    for $k = K$ to 1 {
    find max $k'$ where $(SPEED(TR_{k'}^{j}) < lowSp)\ AND\ (DIST1(TR_{k'}^{j}) < L/2)$
        for $k = k'$ to $K$ {
            find min $k''$ where $ACC(TR_{k''}^{j}) > 1$
            store $DIST1(TR_{k}^{j})$ into accArray
        }
    }
}
```

accDist = the 90th percentile of distance in the *accArray*

```
Determine whether a data point is located within the acceleration zone.
for $j = 1$ to $J$ {
    for $k = 1$ to $K_j$ {
            $CTL(TR_{k}^{j}) = \begin{cases} 1, if\ \ DIST2(TR_{k}^{j}) \leq accDist \\ 0, otherwise \end{cases}$
    }
}
```

By implementing the acceleration and deceleration search algorithms on the corridors in this study, it is found that the acceleration zones mostly range from 400-1000 ft (with an exception of 1240 ft on one corridor) with an average of 500 ft. The deceleration distances range from 500-1000 ft with an average of 600 ft. Figure 27 compares speed profiles before and after removing the acceleration and deceleration data points.



(a) **(b)**

Figure 27: Speeds (a) Before and (b) After Acceleration/Deceleration Filter

4.1.9 Remove highly deviated trips

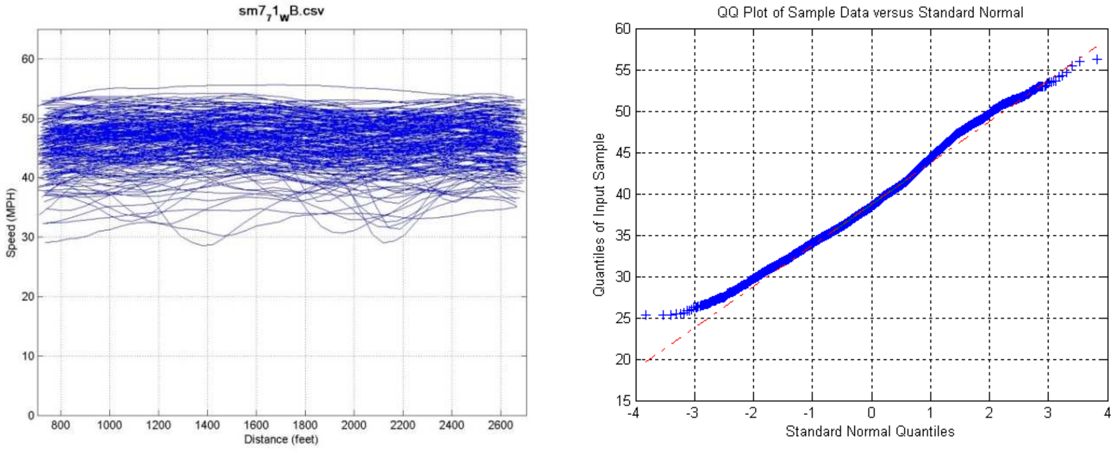
This filter is a final filter to add in the identification of likely free-flow trips. After excluding speed data in the acceleration and deceleration zones from the dataset, there are still some trips with high speed variation. This type of trips implies non-free-flow conditions and should not be included in the analysis. Therefore, a lower bound speed criteria to remove trips with high speed deviation was developed. Quantile-Quantile plot (QQ-plot) was used to visualize the characteristic of speed data. Based on the QQ-plots, majority of the corridors demonstrated a similar pattern of the speed data in which speeds began to deviate approximately minus two standard deviations from the mean. The QQ plot in Figure 28 (b) shows that speed data deviate from normality when data are beyond two standard deviations from the mean. This filter removes trips with speed data exceeding two standard deviations from the mean. Figure 29 depicts the result from the low speed trip filter. Note that approximately 11 percent of the trips from the previous step were detected as highly deviated trips by this filter.

The algorithm is implemented as follows:

find the mean and standard deviation of all data points on the same corridor

$$\overline{V} = \frac{1}{\sum_{j=1}^{J} K_j} \sum_{j=1}^{J} \sum_{k}^{K_j} SPEED(TR_k^j), \qquad \sigma^2 = \frac{1}{\sum_{j=1}^{J} K_j - 1} \sum_{j=1}^{J} \sum_{k=1}^{K_j} \left(SPEED(TR_k^j) - \overline{V} \right)^2$$

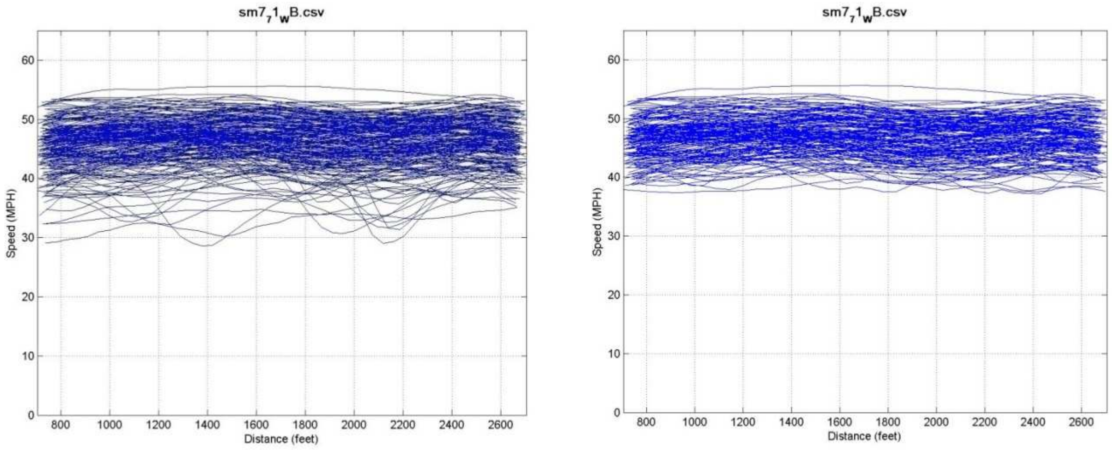
```
for $j = 1$ to $J$ {
        $DEV(TR_k^j) = \begin{cases} 1\ for\ \forall_k\ if\ \min SPEED(TR_k^j) \le \bar{V} - 2\sigma \\ 0\ for\ \forall_k\ otherwise \end{cases}$
}
```



(b) QQ Plot

(a) Speed profile

Figure 28: Speeds and Quantile-Quantile Plot of Corridor No. 21 Westbound



(a) **(b)**

Figure 29: Speeds (a) Before and (b) After Applying Low Speed Filter

4.1.10 Check Quality of GPS signal

Criteria on Number of Satellite (SAT) and Position Dilution of Precision (PDOP) should be established based on the acceptable data accuracy, data availability, and other characteristics of GPS data collected for this study. PDOP is an indicator of the reliability of the GPS data and is geometrically equivalent to the inverse of the volume of the pyramid formed by the satellites in view and the GPS receiver (*44*). In this study, acceptable quality GPS data are defined as data with minimum number of satellites (SAT) of 4 and range of PDOP value between 1 and 8 (*22, 42*). Additionally, the minimum percentage of acceptable quality data for each trip has been set to 80%, meaning that if more than 80 percent of data points from one trip pass the GPS signal criteria, this trip will be included. As a result, 22 percent of the trips from the previous step were detected as trips with more than 20 percent low quality GPS data by this filter.

It is important to note that the GPS signal quality filter is applied after the acceleration/ deceleration filter because the 20-percent criteria is based on number of data points (generally equivalent to travel time in seconds). If the low quality signal condition occurs when the vehicle is stopped in the queue waiting for the green light, it is likely that the trip will a significant percentage of low quality data points, however these exist outside the midblock areas of interest. Therefore, data points in the acceleration and deceleration zones should be detected and excluded before checking the GPS signal quality to maximize the number of usable trips.

The algorithm is implemented as follows:

```
Determine quality of GPS data points
for $j = 1$ to $J$ {
        for $k = 1$ to $K_j$ {
                $SIGNAL(TR_k^j) = \begin{cases} 1, \ if \ (SAT \geq 4) AND \ (PDOP \leq 8) \\ 0, \ otherwise \end{cases}$
        }
}
Determine percentage of good quality data for each trip
for $j = 1$ to $J$ {
        $PCT80(TR_k^j) = \begin{cases} 1, \quad if \ \frac{1}{K_j} \sum_{k=1}^{K_j} SIGNAL(TR_k^j) \geq 0.80 \\ 0, \quad otherwise \end{cases}$
}
```

Once GPS data had gone through the data processing, we use the following attribute values to determine free-flow condition trips:

$$COMP(TR_{.}^{j}) = 1, NIGHT(TR_{.}^{j}) = 0, RAIN(TR_{.}^{j}) = 0,\ QUEUE(TR_{.}^{j}) = 0,$$

$$FF1(TR_{.}^{j}) \neq 0, FF2(TR_{.}^{j}) \neq 0, CTL(TR_{.}^{j}) = 0, DEV(TR_{.}^{j}) = 1, PCT(TR_{.}^{j}) = 1.$$

After the data processing, the number of corridors was reduced to 61 due to data availability criteria chosen to ensure sufficient data for later modeling efforts. These criteria were 1) the effective corridor length is greater than 1000 ft and 2) the number of drivers during the free-flow condition is at least ten. Approximately 66 percent of total trips on the remaining 61 corridors were identified as potentially non-free-flow patterns or low quality of data. There are total of 15,158 trips made by 408 drivers within the selected corridors during year 2004 and the 406,398 second-by-second GPS data points, equivalent to 113 hours of travel.

4.2 CRASH DATA PROCESSING

The four-year crash data from January 2002 to December 2005 are obtained from GDOT's crash database system. These data are stored in Microsoft Office Access (mdb) format. More than 1,200,000 accidents occurred within the State of Georgia during this four-year period. Multiple tools were utilized to query and manipulate crash data including Microsoft Office Access 2003, Perl programming language, and ArcMap GIS software.

The GDOT Crash database contains more than 100 data attributes for each crash record. Table 8 excerpts data attributes that are relevant to the purpose of this study. The 14 fields that are considered in this study include crash ID, date and time, road characteristic (RC) link number, milepoint, latitude and longitude, annual average daily traffic (AADT), first harmful event, weather condition, light condition, pavement surface condition, manner of collision among vehicles involved in the crash, contributing factor, and traffic control type. Note that first harmful event is defined as the first event in a traffic collision to result in injury or property damage.

Several steps of data manipulations were developed to prepare and filter data before application to the model development, including determining crash location, identifying nighttime and inclement weather incidents, and incidents related to pedestrian, bicycle, and animal.

Table 8: Relevant Crash Attributes from GDOT Crash Database

Field Name	Description	Coded Values	
ACC_ID	Accident ID		
ACC_JULDT	Accident Date	mm/dd/yyyy	
ACC_ATIME	Accident Time		
ACC_HE1_TYPE	First Harmful Event	01-Overturn 02-Fire/Explosion 03-Immersion 04-Jackknife 05-Other Non-Collision 06-Pedestrian 07-Pedalcycle 08-Railway Train 09-Animal 10-Parked Motor Vehicle 11-Motor Vehicle in Motion 12-Motor Vehicle in Motion - In Other Roadway 13-Other Object (Not Fixed) 14-Deer 15-Impact Attenuator 16-Bridge Pier/Abutment 17-Bridge Parapet End	18-Bridge Rail 19-Guardrail Face 20-Guardrail End 21-Median Barrier 22-Highway Traffic Sign Post 23-Overhead Sign Support 24-Luminaire/Light Support 25-Utility Pole 26-Other Post 27-Culvert 28-Curb 29-Ditch 30-Embankment 31-Fence 32-Mailbox 33-Tree 34-Other Fixed Object
ACC_WEAT_TYPE	Weather	1-Clear 2-Cloudy 3-Rain 4-Snow	5-Sleet 6-Fog 7-Other
ACC_LITE_TYPE	Light Condition	1-Daylight 2-Dusk 3-Dawn	4-Dark-Lighted 5-Dark-Not Lighted
ACC_SURF_TYPE	Surface Condition	1-Dry 2-Wet 3-Snowy	4-Icy 5-Other
ACC_MNRC_TYPE	Manner of Collision	1-Angle 2-Head On 3-Rear End 4-Sideswipe - Same Direction	5-Sideswipe - Opposite Direction 6-Not A Collision With A Motor Vehicle

Table 8: Relevant Crash Attributes from GDOT Crash Database (Continued)

Field Name	Description	Coded Values
VEH_CONF1_TYPE	Contributing Factor 1	01-No Contributing Factors 02-D.U.I 03-Following too Close 04-Failed to Yield 05-Exceeding Speed Limit 06-Disregard Stop Sign/Signal 07-Wrong Side of Road 08-Weather Conditions 09-Improper Passing 10-Driver Lost Control 11-Changed Lanes Improperly 12-Object or Animal 13-Improper Turn 14-Parked Improperly 15-Mechanical or Vehicle Failure 16-Surface Defects 17-Misjudged Clearance 18-Improper Backing 19-No Signal/Improper Signal 20-Driver Condition 21-Driverless Vehicle 22-Too Fast for Conditions 23-Improper Passing of School Bus 24-Disregard Police Officer 25-Distracted 26-Other
VEH_TRCNTL_TYPE	Traffic Control	1-No Stop Present 2-Traffic Signal 3-RR Signal/Sign 4-Warning Sign 5-Stop or Yield Sign 6-No Passing Zone 7-Lanes 8-Other
LOC_RCLINK_IDENTIFIER	RC Link Number	
LOC_ACC_MILELOG	Milelog	
LOC_X	Longitude	
LOC_Y	Latitude	
LOC_SIGNAL_TYPE	Road Signal Type	S-Traffic Control Device (Red,Amber,Green) P-Traffic Control w/Pedestrian Signalization A-Stop Sign F-Flasher-Other than Overhead Beacon L-Traffic Control Device with Turn Arrow B-Beacon-Overhead Flashing Amber R-Beacon-Overhead Flashing Red C-Stop All Direction Y-Yield Sign W-Yield Sign Opposite Direction of Inventory O-Stop Sign Opposite Direction of Inventory

4.2.1 Crash Location

Crash records can be overlaid on the GIS street layer using latitude (LOC_Y) and longitude (LOC_X) information. However, the latitude and longitude information is not available for every record in the GDOT's crash database – almost 40 percent of the records (505,410 out of 1,285,424 records) do not have latitude and longitude information.

Another way to locate accidents is the linear referencing system. In this system, the location of an event is determined by a linear measure along a reference element. GDOT initially stores crash data using the linear referencing system and then converting the locations to the Cartesian coordinate because the milelog are constantly changing but the Cartesian coordinate is more absolute. The GDOT's crash database uses the fields "RC link Number" and "milelog" to locate an event.

The "Linear Referencing Tools" in the ArcToolbox package was used to locate crash records with missing latitude and longitude coordinate. First, the tool "Create Routes" was used to create a referencing route. The GIS road network obtained from the Georgia Institute of Technology DRIVE lab was input as a base map to create a referencing route. Next, the crash records were located on the referencing road network using the "Make Route Event Layer" tool. This tool located crash locations based on the associated RC Link ID and milepoint information.

Note that 47,659 of crash records (approximately four percent of the total crashes) do not contain either Cartesian or linear referencing coordinate informations; and therefore, are disregarded from the study.

4.2.2 Intersection-Related Crashes

Since this study focuses on road segments rather than intersections, accidents located within the 250-ft distance from the traffic-controlled intersections were identified as intersection-related crashes and removed from the further analysis. The information about the signal type at the intersection was obtained from field visits.

After removing the accidents located in the 250-ft radius from the intersection, the records were further verified using the field “LOC_SIGNAL_TYPE” available in the GDOT crash database. Even though, crashes located at least 250 feet away from the intersection are unlikely to be intersection-related crashes, there is approximately 10 percent of the records are coded as one of the followings:

LOC_SIGNAL_TYPE = “S” (Traffic Control Device)

= “P” (Traffic Control w/Pedestrian Signalization)

= “C” (Stop All Direction)

= “L” (Traffic Control Device with Turn Arrow)

The crashes with the codes above were removed from the analysis as the crash was likely intersection related.

4.2.3 Weather and Light Condition

In this step, crashes that occurred during night and/or inclement weather conditions were removed as the study was aimed at investigating normal operating conditions. The crash records with the following attributes were removed:

ACC_WEAT_TYPE = 3(Rain) and 4(Snow)

ACC_SURF_TYPE = 2(Wet), 3(Snowy), 4(Icy), and 5(Other)

ACC_LITE_TYPE = 2(Dusk), 3(Dawn), 4 and 5 (Dark)

4.2.4 Crash Type

Crashes associated with non-motor vehicle factors such as animals, pedestrians and bicycles were removed. The crash records with the following attributes were removed: ACC_HE1_TYPE = 6 (Pedestrian), 7 (Pedalcycle), 9 (Animal), and 14 (Deer). Table 9 depicts the crash counts and percentage by the first harmful event type. Eighty-seven percent of the crashes were associated with the motor vehicle in motion and less than two percent of the total crashes were associated with pedestrians, pedalcycles, and animals. Approximately two percent of the crashes during the four-year period were associated with the roadside utility poles.

Table 9: Crash Counts and Percentage by First Harmful Event Type

First Harmful Event	Counts	Percent
Motor vehicle in motion	3024	87.20%
Non collision	22	0.63%
Pedestrian	16	0.46%
Pedalcycle	13	0.37%
Animal	2	0.06%
Deer	34	0.98%
Utility pole	72	2.08%
Others	357	10.29%
Total	3468	100.00%

4.2.5 Results

Of the 1,285,424 accidents in the State of Georgia, there are 3,120 accidents located on the 86 study corridors (excluding intersection-related accidents) during the four-year period.

Of the 3,120 segment crashes, 25 percent occurred during inclement weather. Approximately two percent of the segment crashes involved a pedestrian or bicycle. In addition, 22 percent of the segment crashes occurred during nighttime. After applying the filters in sections 4.2.2, 4.2.3, and 4.2.4, 60 percent of the accidents records remained and were used in the model development.

4.3 SUMMARY

This chapter described the data processing algorithms developed for speed and crash data. The speed data processing algorithms from the FHWA project were used in this study to identify speed profiles during various conditions. Trip attributes that can be determined using the developed data processing algorithms include trip continuity, direction of travel, daylight condition, weather condition, likely free-flow trip pattern, traffic-controlled influence zone, and GPS signal quality.

Several steps of data manipulations were developed to prepare and filter data so that they can be associated with the processed speed data. First, crash locations were determined using the linear referencing method. Next, crash events located within the 250-ft radius of an intersection were removed. Light and weather conditions were determined from the crash record attributes. Finally, crashes related to pedestrians, pedalcycles, and animals were removed from the dataset.

Chapter 5. TRAFFIC ATTRIBUTES

5.1 INTRODUCTION

As seen in the literature review, many speed-related traffic parameters have been defined in previous studies. Most of these parameters are point-based, i.e., measurements at a specific location. In this chapter, we propose several traffic parameters derived from one-dimensional spatial data to capture different traffic characteristics along the corridor. The proposed traffic attributes can be grouped into three main categories: speed-related measures, stop pattern measures, and other measures. The speed-related measures attempt to indicate the consistency of vehicle speeds on the roadway while the stop pattern measures attempt to capture the movement conflicts along the corridor.

5.2 SPEED-RELATED MEASURES

In this section, speed variations and other speed-related measures based on previous studies (*22*, *23*) are examined for use as potential surrogate measures of road safety. All speed measures in this section exclude data from the acceleration and deceleration zones as this study is focused on midblock performance.

5.2.1 Speed Variation (SD85)

The 85th percentile speed is selected as it is widely used by roadway designers and practitioners to represent the normal operating condition of the roadway (*45*). The speed variation parameter (SD85) captures the variation of the 85th percentile speed at pre-specified intervals along the corridor and is calculated using the following formula:

$$SD85 = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(v_{85,i} - \bar{v}_{85}\right)^2}.$$

In the equation above, $v_{85,\mathrm{i}}$ indicates the 85th percentile speed at the ith location, where there are N equally-spaced locations along the corridor, and $\bar{v}_{85}$ is the mean of the 85th percentile speeds along the corridor. As SD85 is a variability measure of operating speed, only likely free-flow trips are included in the measure calculation.

The South Atlanta Road corridor (Southbound) was selected to demonstrate the calculation of speed measures. This corridor is a two-lane road with a reversible lane is classified by GDOT as a minor arterial. In addition, the primary land use of this corridor is dense residential (apartment complexes) and commercial. The corridor is bounded by traffic signals. Approximately 680 trips were recorded` during the one-year study period with 45% of the total trips made under the likely free-flow condition, based on the algorithm presented in the previous chapter. Seven percent of the trips entered or exited the corridor at a midblock location. The speed profile of all trips made on this corridor is illustrated in Figure 30.

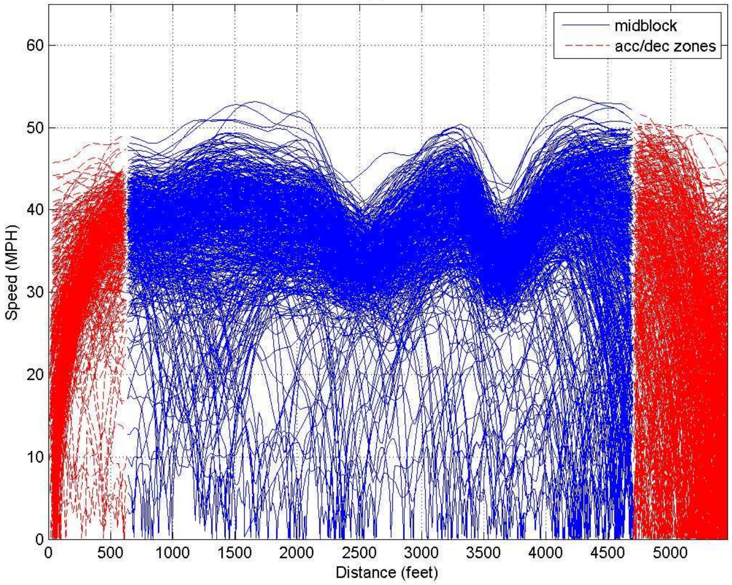


Figure 30: Speed Profiles on the South Atlanta Road Southbound with the Red Dashed Line Indicating the Acceleration and Deceleration Zones

Sensitivity Analysis of Sampling Distance

Since spacing distance between two measured points may influence the calculated speed variation, this subsection investigates the sensitivity of the speed measures to the spacing distance.

The purpose of this effort is to determine the most appropriate sampling distance for the speed variation calculation, for this dataset. An unduly short sampling distance will increase the computational time without gaining additional information while an overly long sampling distance will not yield the true variability of speed along the corridor. For example, Figure 31 depicts the 85th percentile speed profiles of the South Atlanta Road Southbound based on two sampling distances, 100 ft and 1000 ft. It is seen that the speed variation of the 100-ft sampling distance follows the natural frequency of the 85th percentile speed along the corridor. The 1000-ft sampling distance, on the other hand, does not adequately capture the 85th percentile speed variation. For example, the speed variation between 2000 ft and 3000 ft is entirely missed.

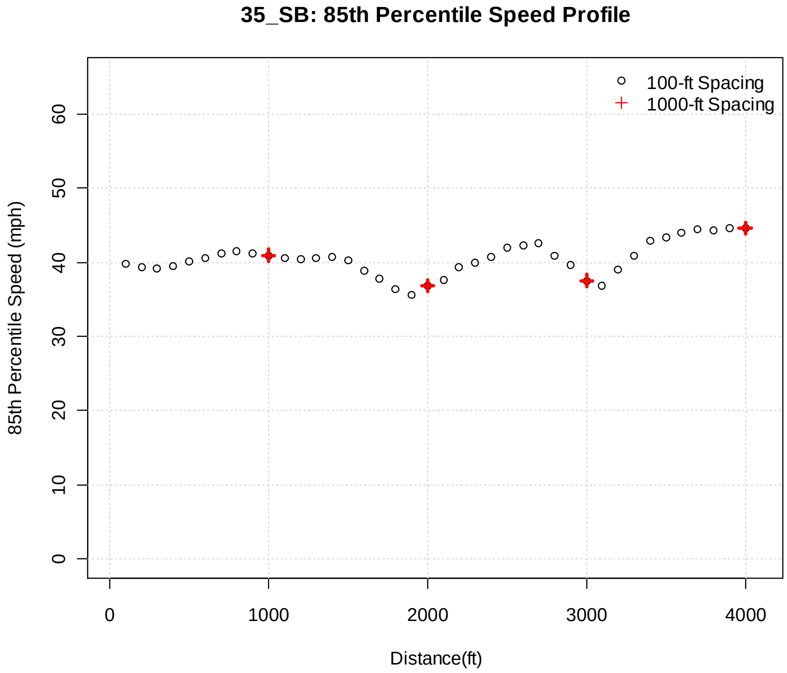


Figure 31: Profile of the 85th Percentile Speed on the Corridor ID 35 Southbound using 100-ft and 1000-ft Sampling Distances

To investigate the impact of spacing, the $SD85$ values were calculated by varying the spacing distance from 50 to 1000 feet, using a 25-ft increment. Figure 32 illustrates the sensitivity of the speed variation due to sampling distance for 12 corridors in the study. Each line in the plot represents $SD85$ of the specific corridor at the respective sampling distance. In general, it can be seen that the speed variation value is relatively constant in the narrow spacing region, i.e., between 50 to 200 ft while the calculated values tend to vary at higher sampling distances. Therefore, the spacing distance of 200 ft is recommended for the speed variation parameter. Results for other corridors are summarized in Appendix B. Note that the $SD85$ has an increasing trend with greater spacing distance in general. This is because standard deviation is inversely proportional with number of sampling points (N). Therefore, $SD85$ tends to increase as number of sampling points decreases.

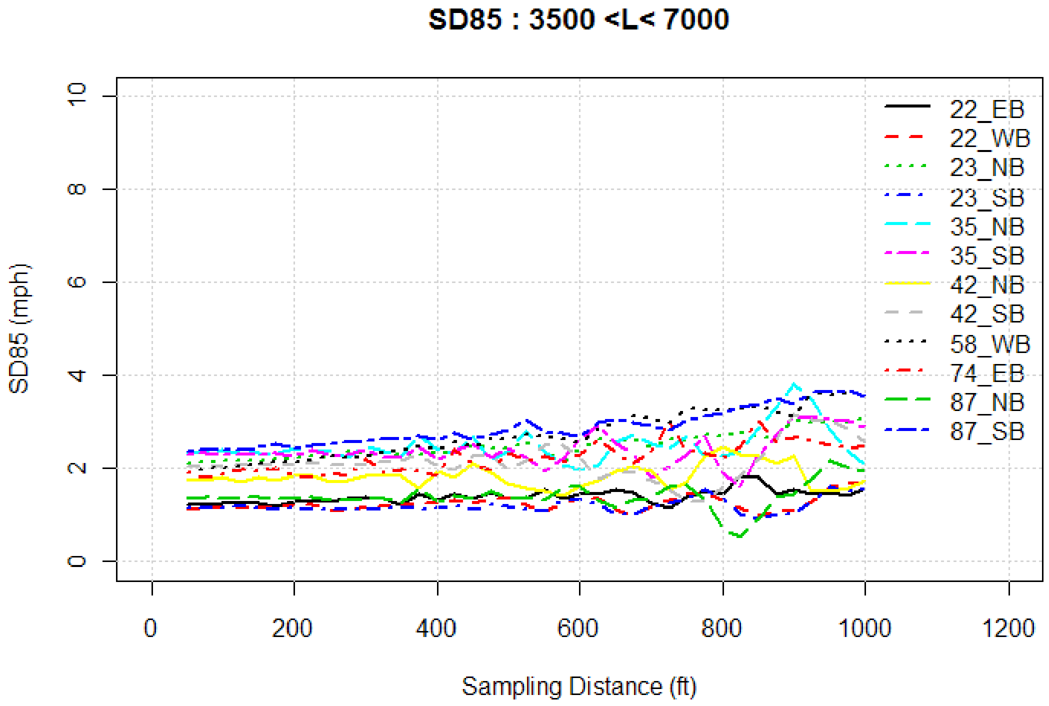


Figure 32: Sensitivity of Speed Variation to the Spacing Distance

5.2.2 Mean of 85th Percentile Speed (M85)

Several studies concluded that higher speed roadways results in a higher crash risk. The mean of the 85th percentile speed ($M85$) along the corridor is included in this study in an attempt to investigate this hypothesis. As with $SD85$, $M85$ measure is based on likely free-flow trip data. The $M85$ is formulated as:

$$M85 = \frac{\sum_{i=1}^{N} v_{85,i}}{N}$$

Figure 33 illustrates the relationship between $SD85$ and $M85$. The measure M85 represents the average 85th percentile speed of the corridor while variation of the 85th percentile speed profile around the mean is represented by $SD85$. For South Atlanta Road Southbound, the $SD85$ is 2.324 mph and the $M85$ is 40.5 mph.

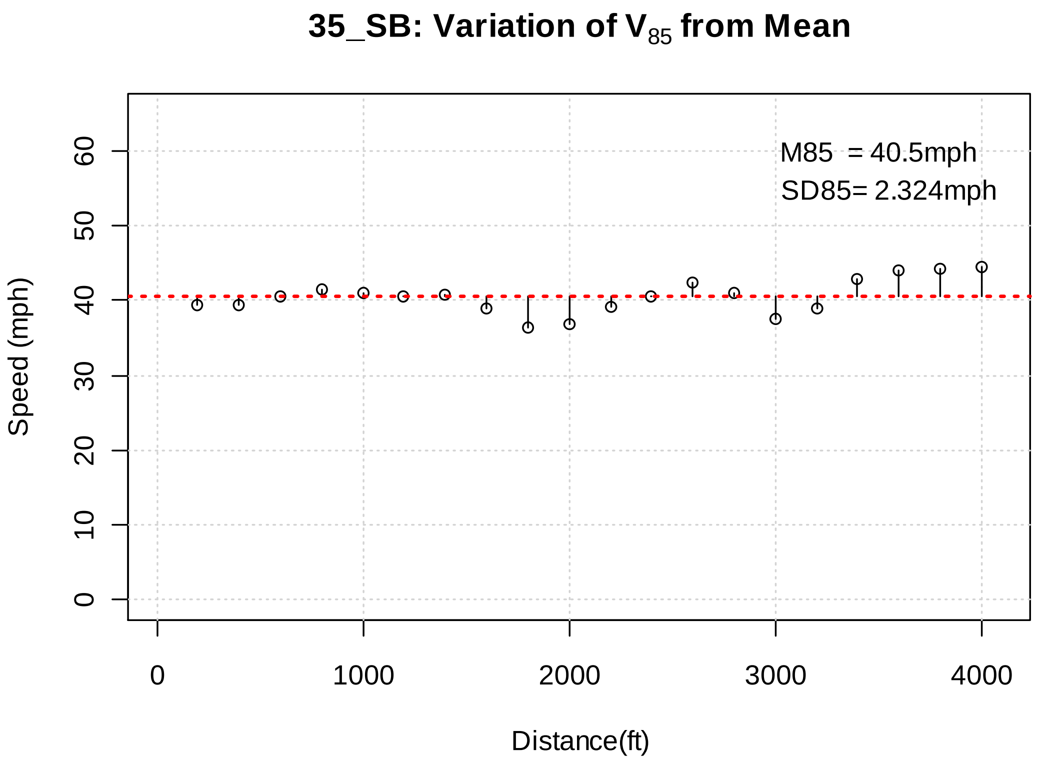


Figure 33: Variation of the 85th Percentile Speed from Mean on South Atlanta Road Southbound

Sensitivity Analysis of Sampling Distance

Figure 34 shows that, unlike $SD85$, the speed parameter $M85$ is relatively insensitive to the spacing distance. In addition, the M85 value does not increase as sampling distance increase. Therefore, the 200-ft sampling distance recommended for $SD85$ is also reasonable to calculate the $M85$.

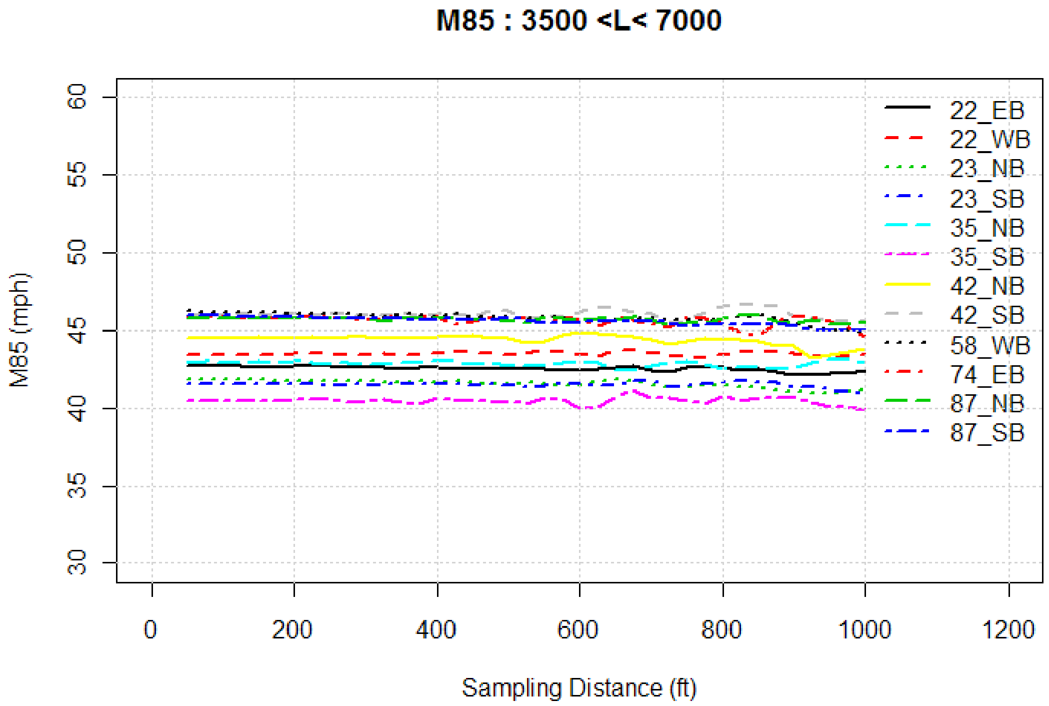


Figure 34: Sensitivity of the Average 85th Percentile speed to the Spacing Distance

5.2.3 Coefficient of variation (CV85)

Similar to $SD85$, this measure, $CV85$, also attempt to represents 85th percentile speed variation. In $CV85$, the $SD85$ is divided by the mean speed to reduce the effect of speed magnitude. The coefficient of variation ($CV85$) is calculated as follows:

$$CV85 = \frac{SD85}{M85}$$

Sensitivity Analysis of Sampling Distance

Figure 35 shows that the fluctuation of $CV85$ has a similar pattern as that of $SD85$. This is expected as $CV85$ is $SD85$ scaled by $M85$, which is relatively constant at any sampling distance. From Figure 35, it is seen that CV85 is stable from 50 to 200 feet and is increasingly variable at larger spacing distance. As a result, the spacing of 200-ft was selected for CV85 calculation.

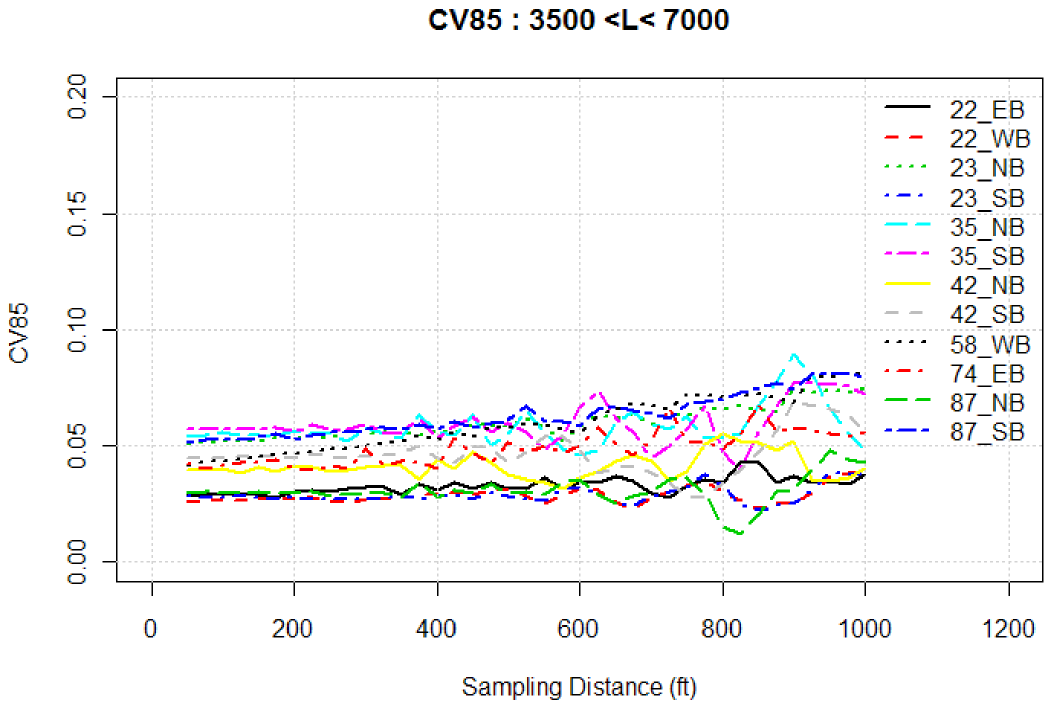


Figure 35: Sensitivity of the Coefficient of Variation to the Spacing Distance

5.2.4 Interquartile Range of 85th Percentile speed (IQ85)

The Interquartile Range of 85th Percentile speed ($IQ85$) is defined as the difference between the third and the first quartile of the 85th percentile speed based on likely free-flow trip data along the corridor. It is formulated as:

$$IQ85 = Q3(v_{85}) - Q1(v_{85})$$

where $Q3(v_{85})$ and $Q1(v_{85})$ are the third and the first quartiles of the 85th percentile speed sample ($v_{85,1}, v_{85,2}, \dots, v_{85,N}$) of the N locations along the corridor.

As with $SD85$ and $CV85$, this parameter measures the fluctuation of the 85th percentile speed along the corridor. However, $IQ85$ does not depend on the number of samples (N) as does $SD85$ and $CV85$. That is, $IQ85$ does not increase as N decreases.

The interquartile range of the 85th percentile speed profile on South Atlanta Road Southbound is illustrated in Figure 36 where the upper line is the third quartile and the lower line is the first quartile. The interquartile for this profile is 2.5 mph.

<u>Sensitivity Analysis of Sampling Distance</u>

Figure 37 shows that the fluctuation of $IQ85$ due to sampling distance. Unlike $SD85$, $IQ85$ does not have an increasing trend when the sampling distance is larger. From the figure, it is seen that $IQ85$ is stable from 50 to 200 feet and increasingly variable at larger spacing distance. As a result, the spacing of 200-ft was selected for $IQ85$ calculation.

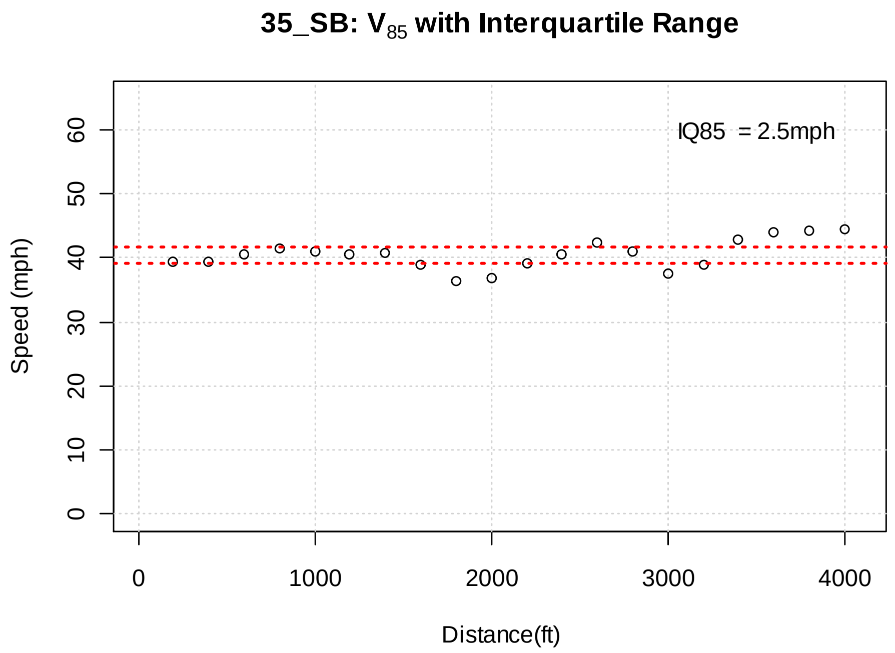


Figure 36: The 85th Percentile Speed Profile with the Interquartile Range Marked by Dashed Line

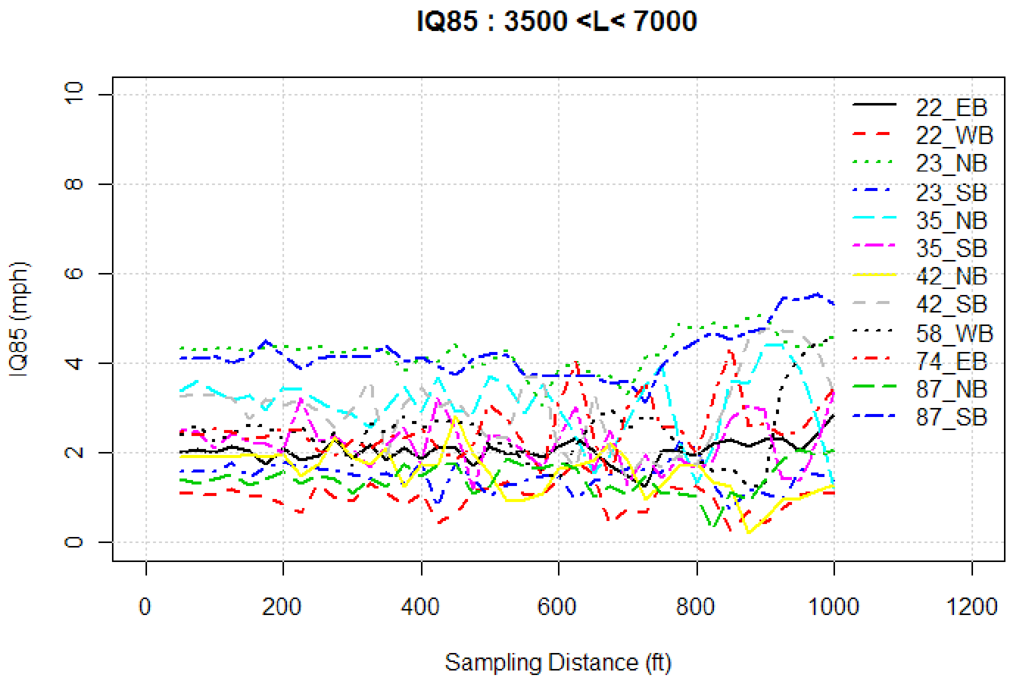


Figure 37: Sensitivity of the Interquartile Range of 85th Percentile Speed to the Spacing Distance

5.2.5 Variation of the 85th speeds Percentile from the speed limit (SVLIM)

The difference between the operating speed and design speed at a particular location can be used to measure the design consistency of a single road element (*46*). However, the design speed is not readily known for most of the corridors. Thus, speed limit was utilized, which is typically related to the design speed (*45*). In this study, since we have speed measurements along the corridor, a new measure ($SVLIM$) is proposed to quantify the design consistency along the roadway:

$$SVLIM = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(v_{85,i} - V_{LIM}\right)^2}$$

where $v_{85,\mathrm{i}}$ indicates the 85th percentile speed at the ith location where there are N equally-spaced locations along the corridor, and V_{LIM} is the speed limit of the corridor. Note that the 85th percentile speed is assumed to represent of the operating speed of the roadway. The calculation is similar to the speed measure in Section 5.2.1. The only difference is that the variation is calculated by comparing the observations with the corresponding speed limit.

The variation of the 85th percentile speed from the speed limit is illustrated in Figure 38. The variation from speed limit is almost 6 mph on this corridor.

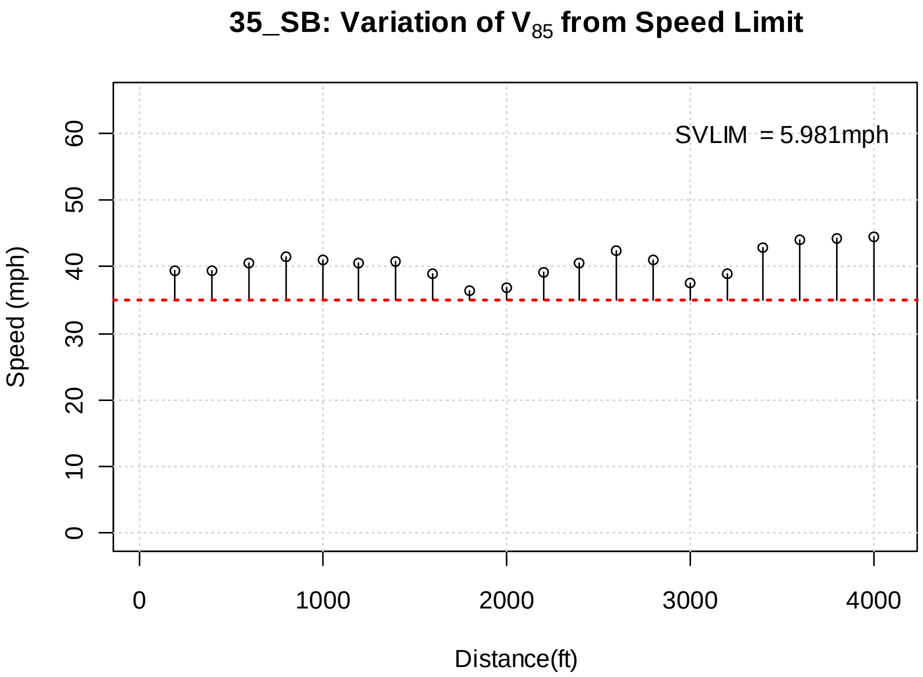


Figure 38: Variation of the 85th Percentile Speed from Speed Limit on South Atlanta Road Southbound

Sensitivity Analysis of Sampling Distance

Figure 39 shows the sensitivity of $SVLIM$ to the sampling distance. It is seen that $SVLIM$ is stable from 50 to 200 feet and increasingly variable at larger spacing distance. As a result, the spacing of 200-ft was selected for $SVLIM$ calculation.

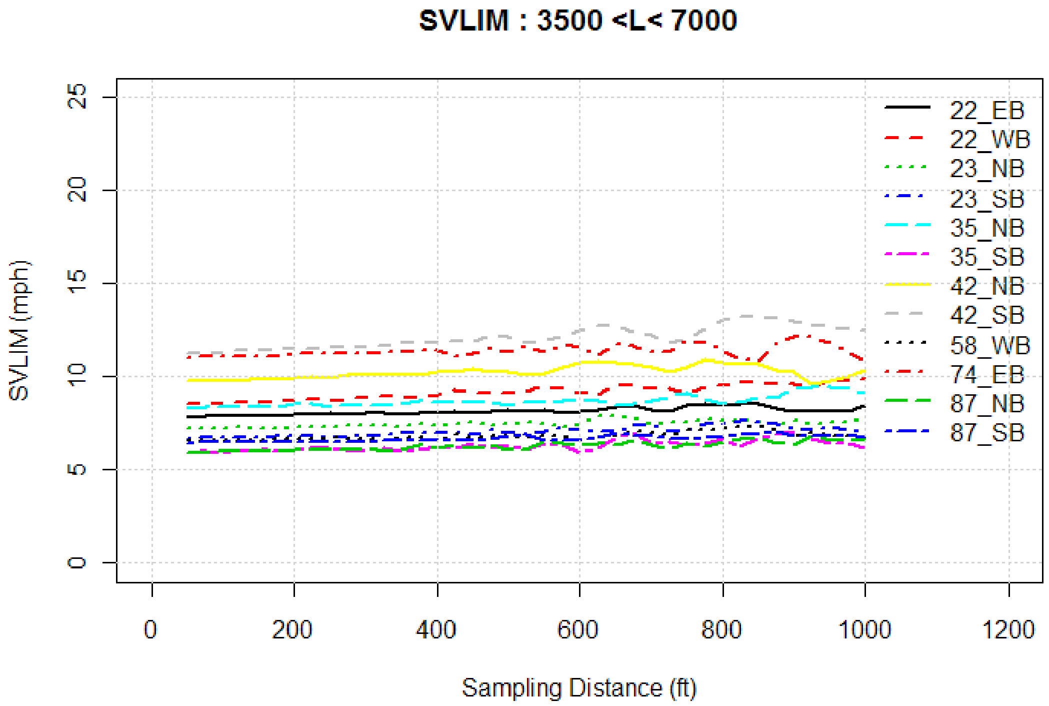


Figure 39: Sensitivity of the Coefficient of Variation (SVLIM) to the Spacing Distance

5.2.6 Mean of Speed Band (M_BND)

Speed band (Δ_B) is defined as the difference between the 95th percentile and the 5th percentile ($v_{95} - v_5$) speed during likely free-flow condition at a specific point. This point-specific measure is intended to capture the variation of multiple trips' speeds at a single location. The speed band profile of South Atlanta Road Southbound is shown in Figure 40. It is seen that this variation measure is not constant throughout the corridor.

Therefore, the proposed measure is the average of the variability of the speed throughout the corridor and is of the form:

$$M\_BND = \frac{1}{N}\sum_{i=1}^{N}\Delta_{B,i}$$

where $\Delta_{B,i}$ is the speed band at location i of the N sampling locations on the corridor. The mean of speed band of the South Atlanta Road Southbound corridor is approximately 9 mph.

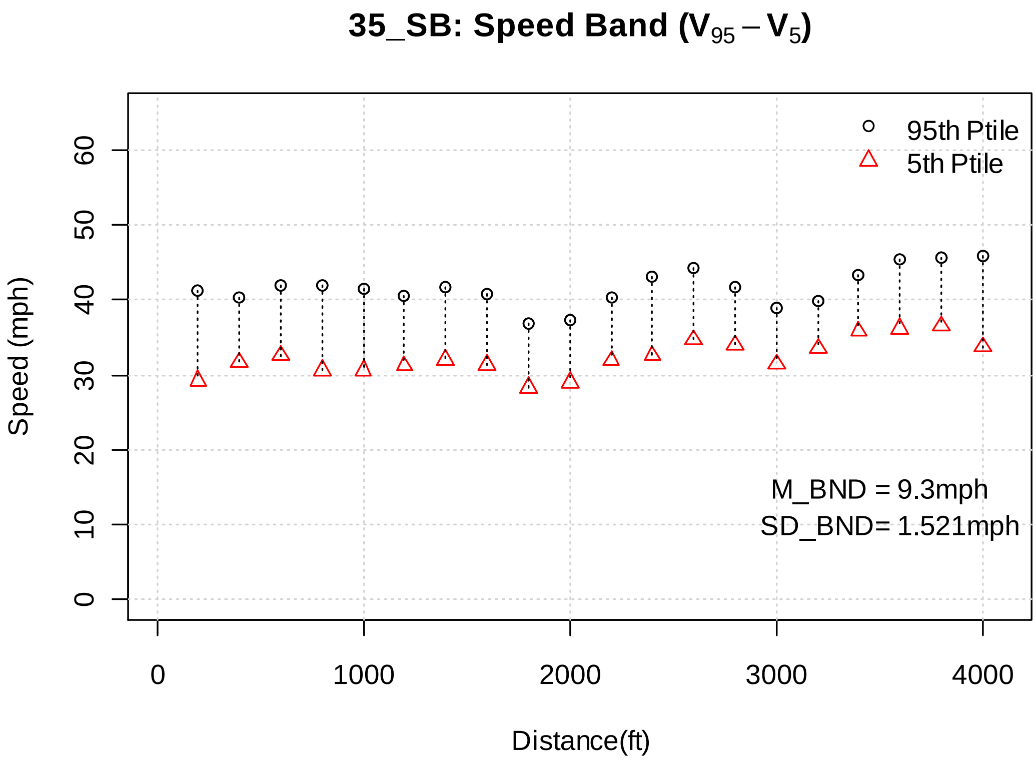


Figure 40: The 95th and 5th Percentile Speed Profile of South Atlanta Road Southbound

Sensitivity Analysis of Sampling Distance

Figure 41 shows the sensitivity of M_BND to the sampling distance. It is seen that M_BND is stable from 50 to 400 feet and with increasing variability at larger intervals. To be consistent with other measures, the spacing of 200-ft was selected for M_BND calculation.

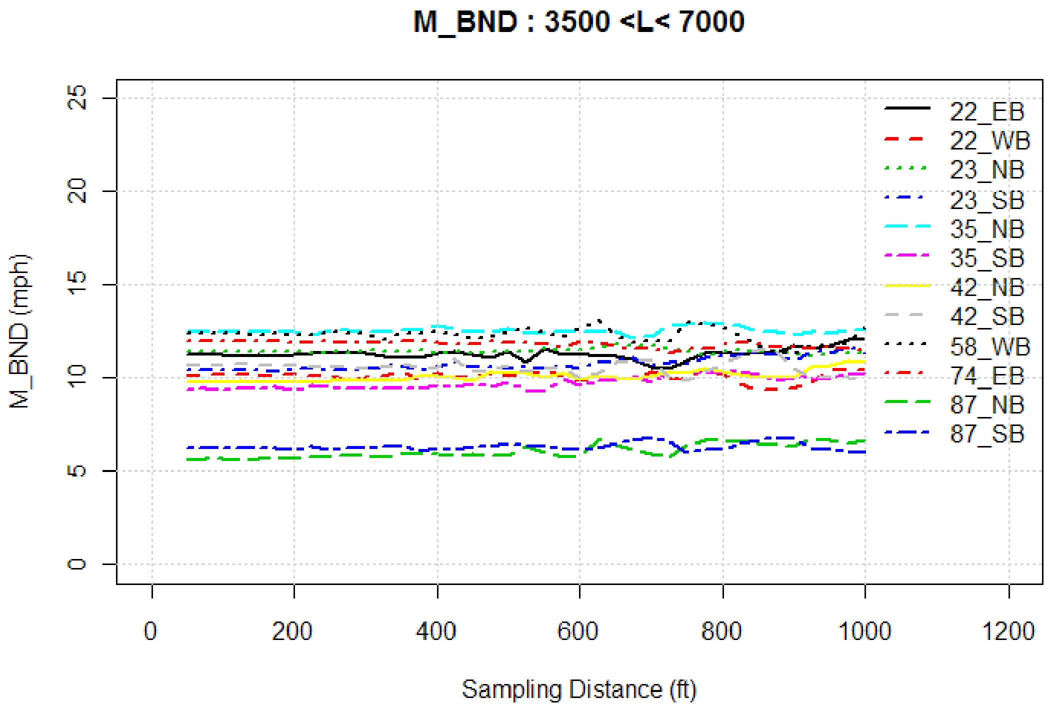


Figure 41: Sensitivity of the Speed Band (M_BND) to the Sampling Distance Interval

5.2.7 Variation of Speed Band (SD_BND)

The variation of the speed band is closely related to the mean of the speed band. This measure is designed to capture the variation in the variability of the speed (i.e., Δ_B) along the corridor and is formulated as:

$$SD\_BND = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(\Delta_{B,i} - M\_BND\right)^2}$$

Note that $\Delta_{B,i}$ indicates the speed band ($v_{95,i} - v_{5,i}$) at the ith location, where there are N equally-spaced locations along the corridor, and M_BND is as previously defined. The variation of speed band South Atlanta Road Southbound is seen in Figure 40. The SD_BND for this corridor is around 1.5 mph.

Sensitivity Analysis of Sampling Distance

Figure 42 shows the sensitivity of SD_BND to the sampling distance. It is seen that SD_BND is stable from 50 to 200 feet with increasing variability at larger sampling intervals. As a result, the spacing of 200-ft was selected for SD_BND calculation.

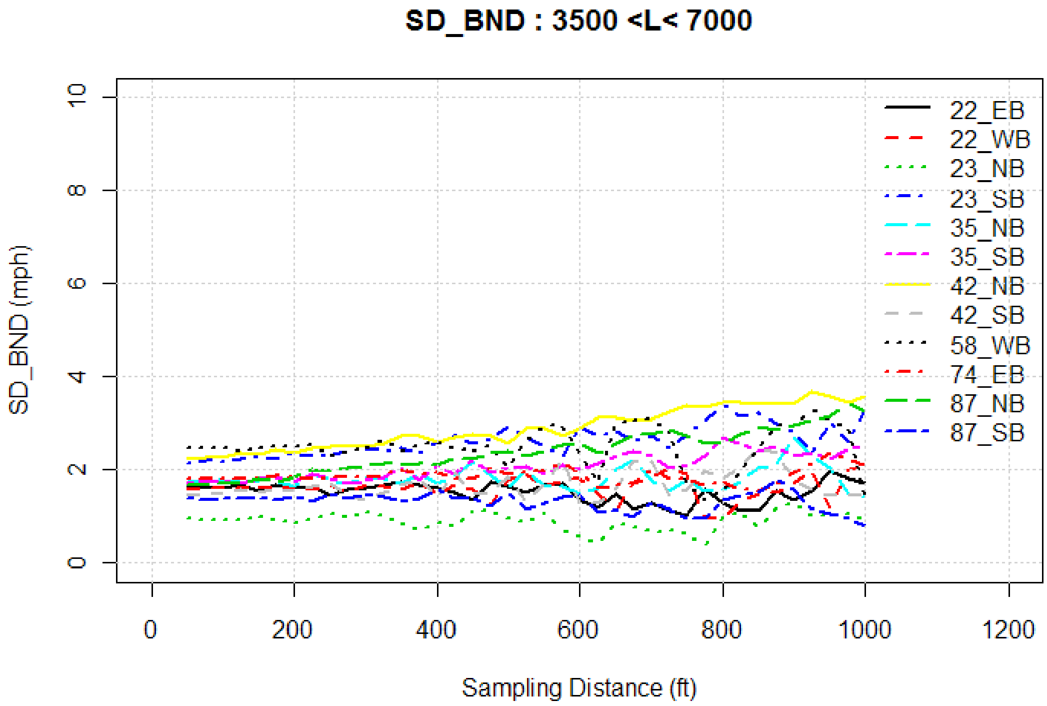


Figure 42: Sensitivity of the Variation of Speed Band (SD_BND) to the Spacing Distance

5.2.8 Acceleration noise (AN)

Acceleration noise is defined as the root mean squared of the acceleration (*27*):

$$\sigma^2 = \frac{1}{T}\int_0^T (a(t) - a_{av})^2 dt$$

or

$$\sigma^2 = \frac{1}{T}\int_0^T a(t)^2 dt - (a_{av})^2$$

where *v(t)* and *a(t)* are the speed and acceleration of a car at time *t* and a_{av} is the average acceleration of the car for a trip taken during time *T*. This time-averaged calculation gives more weight to the low speed data points; therefore, when the vehicle is stationary during the trip, the authors suggested omitting the stopped time from the calculation to avoid bias from the low speed data points. Alternatively, the acceleration noise could be defined in terms of space averages, i.e., averaging acceleration values at every certain distance (*27*).

In this study, the acceleration noises were calculated using both likely free-flow trips only and all vehicle trips during daylight and dry road surface condition. The datasets are generated using the algorithm developed in the previous chapter. It is hypothesized that the acceleration noise during free-flow condition is a function of the noise caused by road geometries and drivers. The dataset of all trips during daylight and dry conditions includes free-flow trips as well as non-free-flow trips. Therefore, the acceleration noise should include that resulting from traffic conditions, road geometries, and drivers.

Under the free-flow condition, acceleration noise calculated using the time-averaged and space-averaged methods are expected to be similar. However, when including non-free-flow trips, the presence of stopped data points may reduce the calculated acceleration noise when using the time-averaged method.

Space-Averaged Acceleration Noise

The value of the space-averaged acceleration noise (${\sigma_t}^2$) of a trip is approximated by:

$${\sigma_t}^2 = \frac{1}{\mathrm{N}} \sum_{i=1}^{N} [a_i - \bar{a}]^2$$

Note that a_i is the acceleration rate at location *i* where there are N equally-spaced locations along the corridor, and $\bar{a}$ is the mean of the N acceleration samples. To evaluate the sensitivity of the spacing interval, intervals were tested from 50 ft to 200 ft. It is further noted that the acceleration noise, ${\sigma_t}^2$ is the acceleration noise of a single trip. However, the subject of interest in this study is the corridor rather than the trip. That is, acceleration noise in this study is considered as a property or characteristic of the corridor. Therefore, acceleration noise from multiple trips should be aggregated to one single value to represent the corridor.

To accomplish this, first, acceleration noise of multiple trips made by the same driver are averaged to represent the driver's acceleration noise. We obtain the driver's acceleration noise (${\sigma_d}^2$) by aggregating the acceleration noises (${\sigma_t}^2$) of T trips made by the same driver:

$${\sigma_d}^2 = \frac{1}{T} \sum_{t=1}^{T} {\sigma_t}^2$$

where ${\sigma_d}^2$ is the average acceleration noise of driver d.

Next, acceleration noise from multiple drivers are averaged to represent the corridor's acceleration noise. We obtain the corridor's acceleration noise (${\sigma_c}^2$) by aggregating acceleration noises (${\sigma_d}^2$) of D drivers on the same corridor:

$${\sigma_c}^2 = \frac{1}{D}\sum_{d=1}^{D} {\sigma_d}^2$$

where ${\sigma_c}^2$ is the acceleration noise of corridor c

As aforementioned, ${\sigma_c}^2$ is calculated for both free-flow condition and all trips during daylight and dry conditions. These are denoted as AN_FF for the acceleration noise of the free-flow trips and AN_AF for the acceleration noise of the all trips during daylight and dry conditions.

Sensitivity Analysis of Sampling Distance

Figure 43 shows the sensitivity of AN_FF to the sampling distance and Figure 44 shows the sensitivity of AN_AF to the sampling distance. It is seen that, in general, AN_AF is higher and more variable than AN_FF. This is because AN_AF includes trips under likely non-free-flow condition. In addition, both AN_AF and AN_FF are relatively stable from 50 to 200 feet and becomes more variable at larger sampling distances. As a result, the spacing of 200-ft was selected for the acceleration noise calculation.

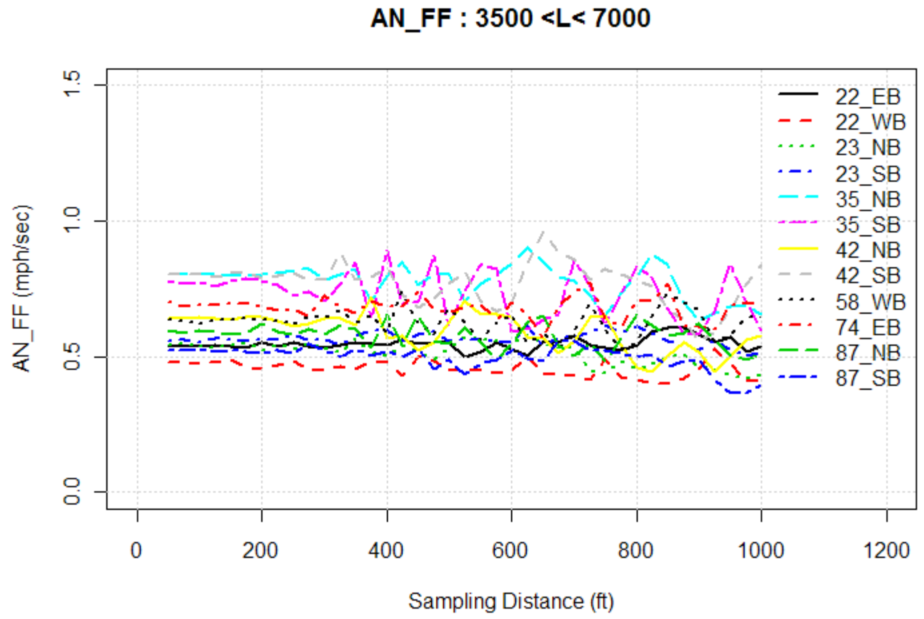


Figure 43: Sensitivity of Acceleration Noise under Free-Flow Condition (AN_FF) to the Sampling Interval

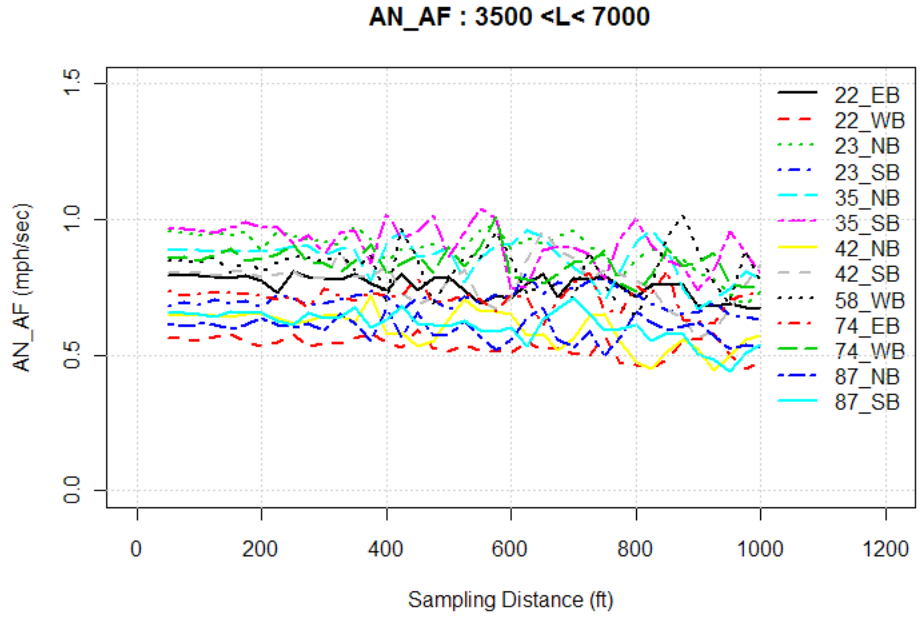


Figure 44: Sensitivity of Acceleration Noise for All Trips under Daylight and Dry Conditions (AN_AF) to the Sampling Interval

Time-Averaged Acceleration Noise

Figure 45 represents a non-free-flow trip traversing South Atlanta Road Southbound corridor during the weekday PM peak. The total travel time through the corridor was 198 seconds. The stopped time – defined as speed lower than 5 mph – of this trip was 99 seconds, equivalent to 50% of the total travel time. Stationary data points associated with the zero acceleration rates as shown in Figure 46. As a result, including the stationary data points in the calculation reduces the magnitude of the acceleration noise.

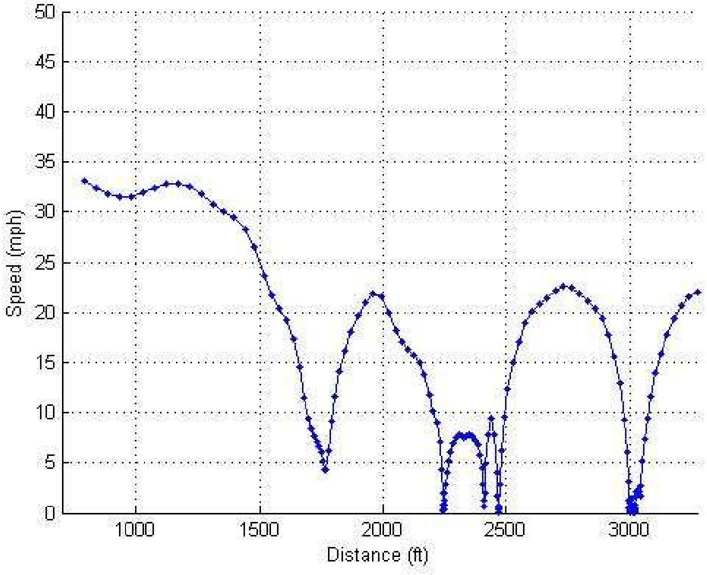


Figure 45: Speed Profile of a Non-Free-Flow Trip Traversing the Corridor South Atlanta Road Southbound

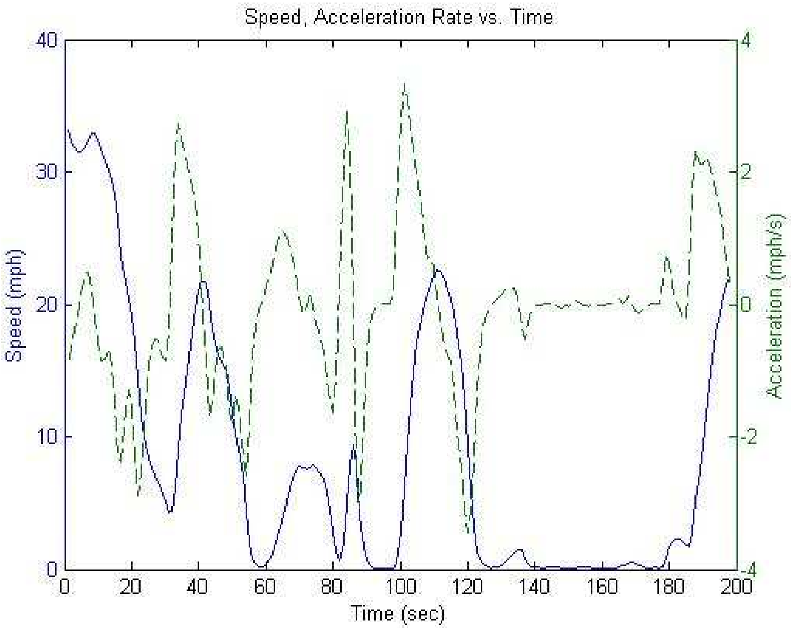


Figure 46: Speed (Left Axis) and Acceleration Rate (Right Axis) vs. Time of a Non-Free-Flow Trip

In this study, the time-averaged acceleration noise (${\sigma_t}^2$) of a single trip is approximated by:

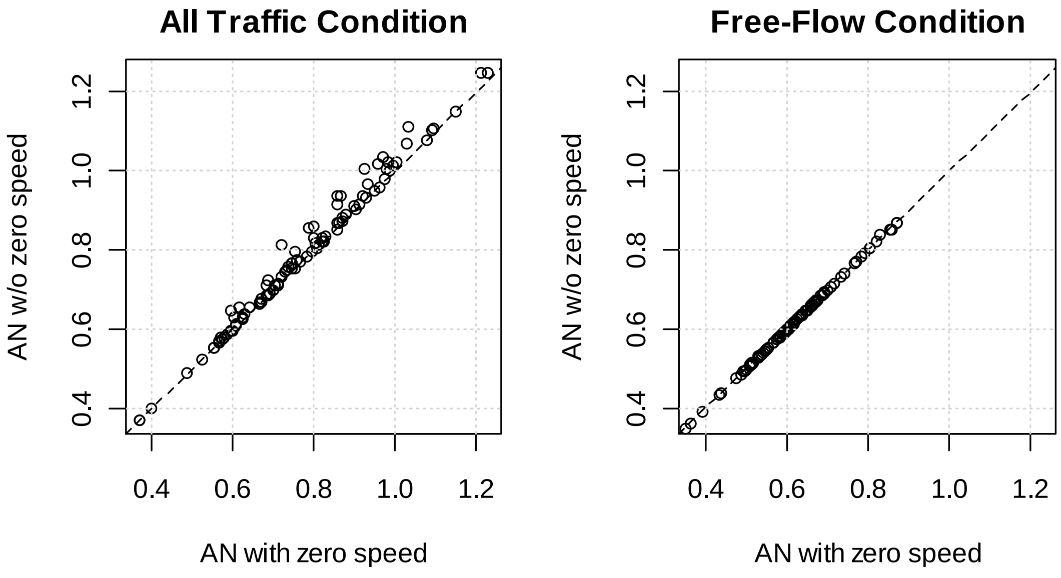
$${\sigma_t}^2 = \frac{1}{\mathrm{N}}\sum_{i=1}^{N}[a_i - \bar{a}]^2$$

where a_i is the acceleration at time i and is measured at every one second throughout the N seconds of the trip's travel time, and $\bar{a}$ is the average acceleration of the trip. The time-averaged acceleration noise for driver (${\sigma_d}^2$) and corridor (${\sigma_c}^2$) were obtained in the same manner as in the space-averaged case.

Comparison of With and Without Stopped Time Acceleration Noises

The corridor's acceleration noise values for every study segments were calculated and plotted in Figure 47. The left hand side figure (a) depicts AN values under all traffic condition while the right hand side figure (b) depicts AN values under free-flow condition (i.e., all trips during daylight and dry condtions). The x-axis represents the acceleration noise calculated from all data points and the y-axis represents the acceleration noise calculated from the same dataset but without stopped time – defined as data points having measured speed less than 5 mph. Figure 47(a) shows that all the points are located at or above the x=y diagonal line. This means that, after removing stationary data points, AN tends to increase for all traffic condition. Note, the stopped time does not appear to have a significant impact on the acceleration noise in this study because the corridor's acceleration noise was averaged from multiple trips and multiple drivers. Figure 47 (b) shows that acceleration noise under likely free-flow condition are not significantly affected by the calculation method, i.e., with or without stopped time in the dataset. This is as expected as the free-flow condition does not include any trips with a low or zero speed.

In summary, the stopped time causes the underestimation of the acceleration noise under all traffic condition; therefore, it is suggested to remove the stopped time before calculating the time-averaged acceleration noise.



(a) **(b)**

Figure 47: Comparison of Acceleration Noises (mph/sec) With and Without Stationary Data Points under (a) All Traffic Condition and (b) Free-Flow Condition

Comparison of Time-Averaged and Spaced Average Acceleration Noises

Since there are two possible methods to calculate the acceleration noise, this sub-section determines which calculation method is appropriate to the dataset in this study. It is noted that the 200-ft sampling distance was used to calculate the space-averaged acceleration noise under free-flow and all traffic conditions.

Figure 48 illustrates the acceleration noise computed by the time-averaged and the space-averaged methods. Figure 48 (a) shows the acceleration noise given all trips under daylight and dry conditions. When the acceleration noise is high, the time-averaged method tends to give a higher value. This is because the time-averaged method tends to oversample at lower speed which also tend to have higher acceleration/deceleration rates. Figure 48 (b) shows the acceleration noise under potentially free-flow condition. It is seen that under this condition, the acceleration noise from both methods yield similar results.

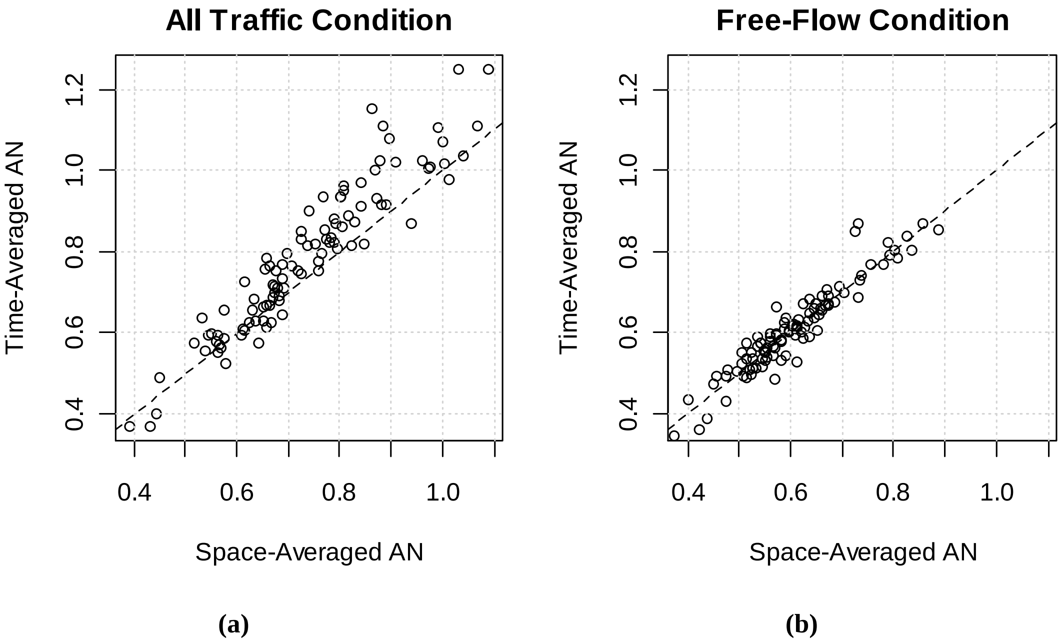


Figure 48: Relationships between Time-Averaged and Space-Averaged Acceleration Noise (mph/sec) under (a) All Traffic Condition and (b) Free-Flow Condition

As a result, this study selects the time-averaged acceleration noise calculation method as it is reflective of the variation due to traffic interference when considering all trips and provides a similar result to the space-averaged during the free-flow condition.

5.3 STOP-RELATED MEASURES

Crash frequency is highly correlated with roadside features such as driveway density, side street density, median type, and adjacent land use (*30*, *32*, *47*, *48*). Several measures related to stopping maneuvers are proposed to represent traffic characteristics influenced by roadside features mentioned previously. Unless otherwise stated, all stop-related measures utilize all trip data during daytime and dry conditions.

5.3.1 Stop Frequency per Trip per Mile (STOPS)

The stop frequency per trip per mile is essentially the ratio of number of stops while the instrumented vehicles traverse the midblock section of the corridor to number of trips that traversed the corridor per unit length. As stated, only trips made during the daylight and dry road surface condition are considered. A single stop is defined as the moment when the vehicle reduces its speed under 5 mph until the speed again exceeds 5 mph. There can be more than one stop along the corridor in a single trip. For example, the speed profile plotted in Figure 49 experienced five stops which the vehicle traversed South Atlanta Road Southbound. The parameter $STOPS$ can be formulated as:

$$STOPS = \frac{No.of\ Stops}{No.of\ Trips * Length}$$

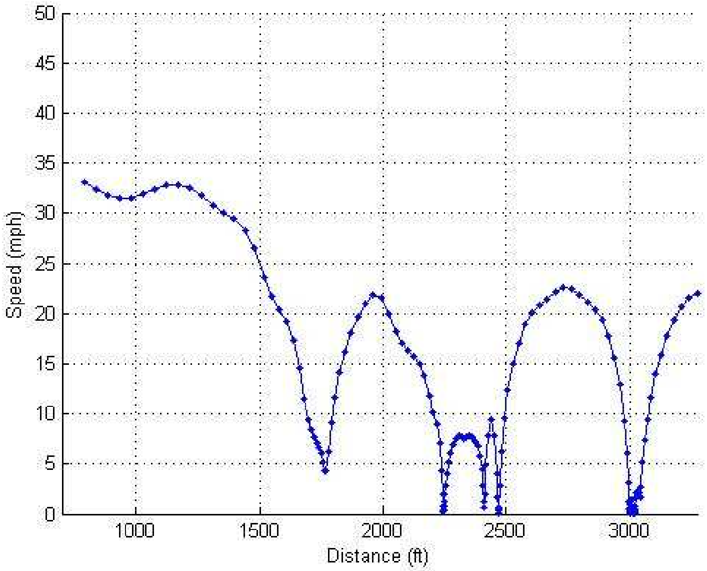
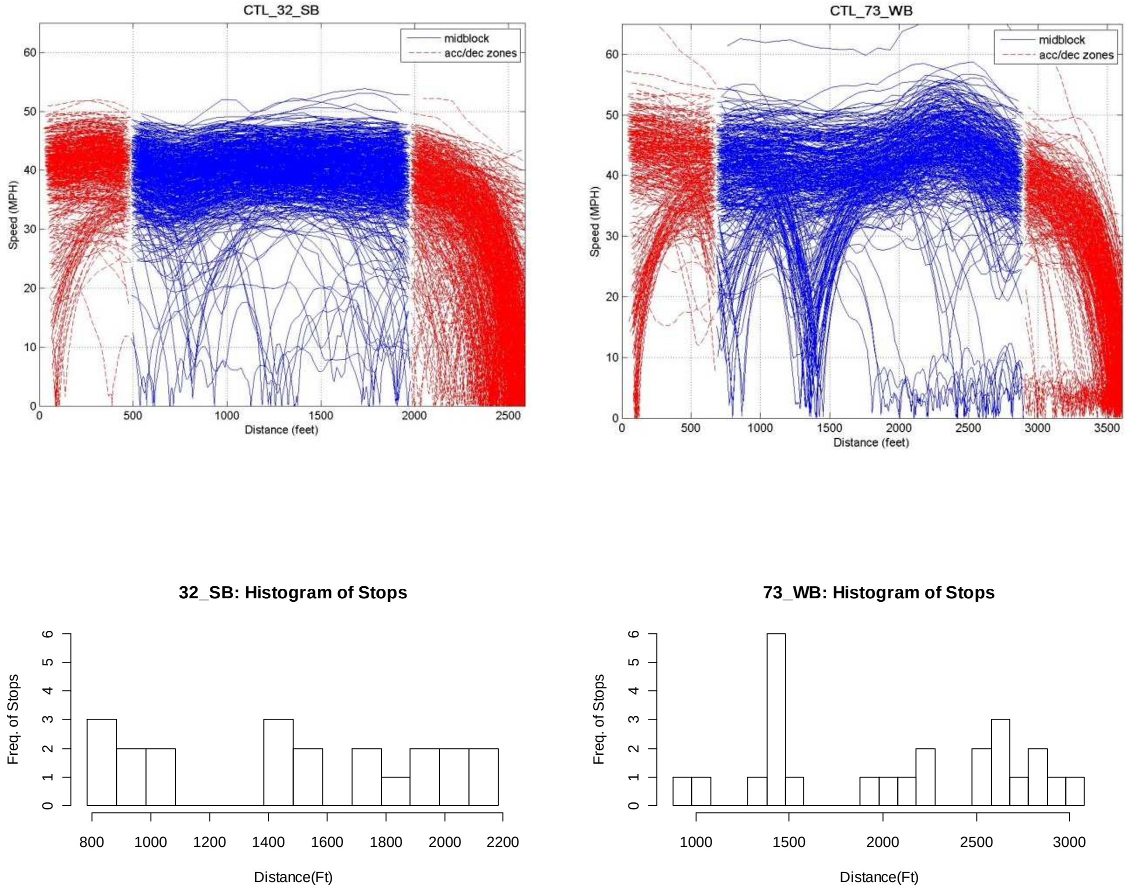


Figure 49: Speed Profile of a Non-Free-Flow Trip Traversing the Corridor 30SB

5.3.2 Coefficient of Variation of stops within 100-ft block (CV_S100)

While $STOPS$ measures the stop frequency per trip per length of the corridor, another measure was desired to indicate the dispersion of stop locations. Figure 50 compares speed profiles and stop location distributions of two corridors, namely Roberts Drive Southbound and Westside Parkway Westbound. The $STOPS$ values for the corridors Roberts Drive Southbound and Westside Parkway Westbound are 0.33 and 0.35 stops/trip-mile, respectively. Even though the $STOPS$ values are similar, the speed profiles and histograms show that the distributions of the stop locations are quite different for the two corridors. That is, stop locations on Roberts Drive Southbound are more evenly distributed compared with those on Westside Parkway Westbound. The difference in stop location distribution might result in different crash distribution.



(a) STOPS=0.33/trip-mile, CV_S100=0.73 **(b) STOP=0.35/trip-mile, CV_S100=1.19**

Figure 50: Speed Profiles and Histogram of Stop Frequency in each 100-ft block along the corridor (a) Roberts Drive Southbound and (b) Westside Parkway Westbound

Therefore, the measure CV_S100 or the coefficient of variation of stops within each 100-ft block is developed to capture the difference in distribution of stop locations. First, the road segment is divided into 100-ft intervals. Next, the number of stops within each interval is determined. This is similar to creating a histogram of stop location frequency with a 100-ft bin size. Mean and standard deviation of frequency in each bin are calculated. Finally, we obtain the parameter, CV_S100, as the ratio of mean to standard deviation of the frequency of stop locations. It can be formulated as:

$$CV\_S100 = \frac{SD\_S100}{M\_S100}$$

where M_S100 and SD_S100 are the mean and standard deviation of stop frequency in the corridor intervals.

5.3.3 The 90th Percentile Count of Stops within 100-ft Intervals (P90_S100)

Side streets or driveways with high traffic volumes may create higher numbers of conflicts with the main street traffic and, in turn, increase the crash risk. The 90th percentile highest stop frequency within 100-ft interval represents degree of conflict between the main street and the side street or driveway on the corridor.

5.3.4 Moran's Index of Number of Stops within 100-ft Block (MI_S100)

The Moran's Index (*49*) provides a measure of the spatial auto correlation of a variable. For this study, it is desired to measure the correlation between the value of stop frequency in one location (interval) and values in neighboring interval. When the high stop frequency interval are located close to each other in one area and low stop frequency intervals are also located close to each other in another area, it indicates that majority of vehicles tend to stop in the same location, which could be a major driveway with high conflict. These areas have the potential to be high crash locations. The Moran's Index can be calculated as:

$$I = \frac{N \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\left(\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}\right) \sum_{i=1}^{n}(x_i - \bar{x})^2}$$

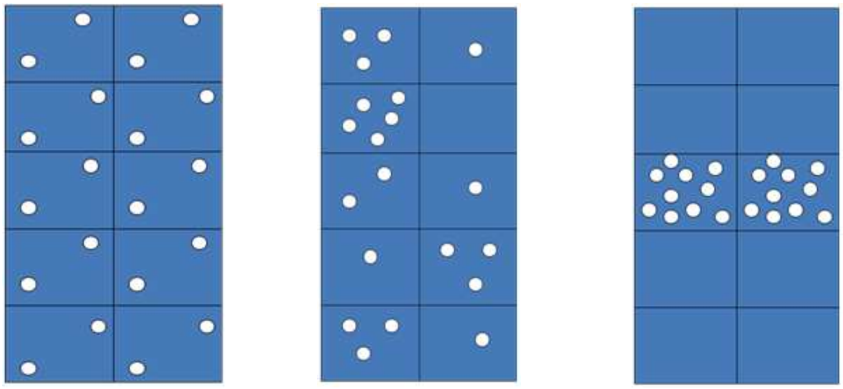
where N = Number of intervals along the corridor

x_i = Stop frequency of interval $i, i = 1, \dots, N$

x_j = Stop frequency of the neighbor interval $j, j = 1, \dots, N$

w_{ij} = A weight indexing location of i relative to j

Moran’s Index varies between -1 and +1. The value close to +1 indicates high clustering pattern of stop locations, i.e., high and high frequency intervals or low and low frequency intervals are close to each other. On the other hand, the value close to -1 indicates a highly dispersed pattern or uniform distribution of stop locations, i.e., high and low frequency interval are alternately next to each other. When Moran’s Index is zero, the stop locations are distributed randomly. Figure 51 illustrates the example of dispersed, random, and clustered spatial patterns. As an example in this study, the Moran’s Index for Robert Drive Southbound is -0.03 while the value for Westside Parkway Westbound is 0.74.



(a) Dispersed
(Moran's I → -1)

(b) Random
(Moran's I → 0)

(c) Clustered
(Moran's I → +1)

Figure 51: Spatial Distribution Patterns (*50*)

5.4 OTHER MEASURES

5.4.1 Percent Through-Traffic Trips (P_THRU)

The ratio of through traffic trips to all trips is used to determine degree of activity of land use along the corridor during daytime, dry road surface conditions. The term "all trips" include through traffic as well as the entering/exiting trips to/from the side streets or driveways along the corridor. A high percentage of through traffic indicates low activity land use, which also indicates low traffic on side street/ driveways. The percent through traffic trips can be formulated as:

$$P\_THRU = \frac{No.of\ Through\ Traffic\ Trips}{No.of\ Total\ Trips}$$

5.4.2 Percent Free-Flow Trips (P_FF)

The proportion of potentially free-flow trips (determined using the algorithm in Section 4.1.7) to all trips is used to determine degree of traffic congestion during the daytime and on dry road surface condition. The percent free-flow trips can be formulated as:

$$P\_FF = \frac{No. of Free - Flow\ Trips}{No. of\ Total\ Trips}$$

5.5 SUMMARY

The traffic attributes can be divided into three categories including speed-related measures, stop-related measures, and other measures. The speed-related measures capture the speed characteristics of the roadway, including speed variation, mean of the 85th percentile speed, coefficient variation of the 85th percentile speed, interquartile range of the 85th percentile speed, variation from the speed limit, mean and variation of the speed band, and acceleration noise. The stop-related measures capture the conflict movement characteristics of the roadway including stop frequency, coefficient of variation of stops within each 100-ft long interval, the 90th percentile count of stops within each 100-ft long interval, and the Moran's Index. The other measures capture the intensity of land use activities, including percent through trips and percent free-flow trips. The traffic attributes to be considered in the model development effort are summarized in Table 10.

Table 10: Summary of Traffic Attributes

Traffic Variable	Description
SD85	Variation of 85th percentile speed along corridor
M85	Mean of 85th percentile speed along corridor
CV85	Coefficient of Variation of 85th percentile speed along corridor
IQ85	Interquartile range (Q3-Q1) of 85th percentile speed along corridor
SVLIM	Deviation of 85th percentile speed from speed limit along corridor
M_BND	Mean of speed band (95th-5th percentile speed) along corridor
SD_BND	Deviation of speed band (95th-5th percentile speed) along corridor
AN_FF	Acceleration noise of free-flow trips
AN_AF	Acceleration noise of non-free-flow trips
STOPS	Stop frequency per trip per mile
CV_S100	Coefficient of variation of stops within 100-ft interval
P90_S100	The 90th percentile count of stops within 100-ft interval
MI_S100	Moran’s Index
P_THRU	Percent through-traffic trips
P_FF	Percent free-flow trips

Chapter 6. SENSITIVITY OF SPEED MEASURES

6.1 INTRODUCTION

This chapter explores the effects of data filters, developed in Chapter 4, on the speed measures, developed in Section 5.2. The analysis investigates the sensitivity of the measures to the application of sequential filters and the effect of individual filters. The former analysis, discussed in Section 6.2, determines the changes in speed measures after each data filter is applied sequentially. The latter analysis, discussed in Section 6.3, tries to answer the "what-if" type of questions. For instance, one might wonder what if we do not apply the weather filter because we do not have weather data, how it is going to affect our speed measures given other conditions remain the same.

6.2 SENSITIVITY TO SEQUENTIAL FILTERS

6.2.1 Methodology

As stated, the speed measures in Section 5.2 were calculated using likely free-flow data during the non-inclement daylight conditions. The identification of a trip as likely free-flow during non-inclement conditions is based on a series of identified filters. In this section, the speed measures are calculated after each filtering step to see how the values of speed measures vary as each data filter is sequentially applied.

Table 11 lists the seven data filters being used in the analysis. Table 12 identifies the eight incremental steps in the application of the filters starting from run number 0, which does not have any filter applied, to run number 7, which has all seven filters applied. The "plus" symbol in front of the filtering codes in Table 12 indicates that one additional filter is included to the previous filter set. Only the trips that pass all filtering criteria in each run will be used to calculate the speed measures for the corridor. Table 13 shows the data structure of the data file used to calculate the speed measures for each run. Each line represents individual trip’s speed data sampled every 200 ft, tagged by nine filter values. The filter values are determined for every trip according to the discussion in Section 4.1. The distance sampling interval for the speed measures follows Section 5.2. The first record states that trip No. 1 (TRP01) of driver No. 1 (DVR01) occurred during daylight (L=1), inclement weather (R=1), free-flow traffic condition (Q=0,F1=4, F2=1, and D=0), and had a high GPS signal quality. This trip was made on Friday (W=5) between 9:00 and 11:00 time period (O=2).

Table 11: Data Filter Code Description

Filter Number in Section 4.1.	Description	Code
5	Daylight Filter	L
6	No Rain Filter	R
7A	Queue Filter	Q
7B	Free-Flow Filter Type I	F1
7C	Free-Flow Filter Type II	F2
9	Highly Deviated Trips Filter	D
10	GPS Signal Quality Filter	S
-	Weekday/Weekend	W
-	Peak/ Off-Peak Period	O

Table 12: Sequence of Filters Applied to the Speed Data

Run	Filters Applied	Code
0		RAW
1	L	+L
2	LR	+R
3	LRQ	+Q
4	LRQF1	+F1
5	LRQF1F2	+F2
6	LRQF1F2D	+D
7	LRQF1F2DS	+S

Table 13: Data Structure of GPS Speed Data with Tagged Filtering Information

	Filter Value									Speed Data		
TRIP ID	L	R	Q	F1	F2	D	S	W	O	@0 ft	@200ft	…
DVR01_TRP01	1	1	0	4	1	0	1	5	2	34.00	37.09	38.28
DVR01_TRP02	1	0	0	2	4	0	1	6	2	38.83	38.47	39.43
DVR01_TRP03	1	0	0	4	4	1	1	4	2	35.04	33.15	34.77
DVR01_TRP04	1	0	0	2	2	1	1	1	4	36.28	32.28	30.21
DVR01_TRP05	1	0	0	2	4	1	1	2	5	40.33	39.06	39.49
DVR02_TRP01	1	0	1	0	0	1	1	5	3	32.06	30.05	28.73
DVR02_TRP02	1	0	0	4	1	1	1	5	5	41.31	41.53	39.71
DVR02_TRP03	1	0	1	0	0	1	1	5	4	34.59	34.65	32.82
DVR02_TRP04	1	0	1	0	0	1	1	1	4	35.68	32.63	35.56
DVR02_TRP05	1	0	0	2	4	0	1	4	2	38.05	38.99	40.68
DVR03_TRP01	1	0	0	2	3	1	1	1	3	37.08	32.25	27.59
DVR03_TRP02	1	1	0	2	2	0	1	1	3	41.70	39.69	40.01

6.2.2 Results and Discussions

Six corridors were selected for sensitivity analysis. Two of the six corridors have relatively low variation (26 Eastbound and 71 Westbound), two corridors have relatively high variation (40 Eastbound and 71 Westbound), and two corridors have high stop frequency (03 Eastbound and 35 Southbound). The degree of variability was determined from the calculated SD85 and the frequency of stops were determined from the speed profile plots.

The results are presented in bar chart format (Figure 52 to Figure 57) where each figure reports eight traffic attributes ($SD85$, $M85$, $CV85$, $IQ85$, $SVLIM$, AN, M_BND, and SD_BND) and two data availability attributes (No. of drivers and No. of trips) for the individual corridor. An individual bar chart represents the sensitivity of one measure, where each bar shows the value of the speed measure of a single scenario. The denotation of filter code is listed in Figure 9. Filters are sequentially added from left to right, i.e., the left most bar does not incorporate any data filters while the right most bar has seven filters applied.

In general, the speed measures derived from the 85th percentile speed (V_{85}) profile such as $SD85$, $M85$, $CV85$, and $IQR85$ are minimally affected by the filter set. For example, the V_{85} profiles before appending any filters and after applying all seven data filters on South Atlanta Road Southbound is illustrated in Figure 58. The before and after speed profiles show that the profile pattern remains relatively the same except a slightly offset (approximately 1 mph higher) after the filtering process. The small change in speed profile is due to the fact that the filters Q, F1, F2, and D are designed to remove potentially non-free-flow trips, which normally impact trips with lower speeds. Therefore, V_{85} tends to increase slightly after removing potentially non-free-flow trips. This results in a stable $SD85$ and a small increase of $M85$.

The variation from speed limit of V_{85} $(SVLIM)$ is more sensitive to the data filters than the measures $SD85$, $M85$, $CV85$, and $IQR85$. That is, the design consistency parameter shows a gradually increasing trend with the sequence of filters. The likely reason is that the measure $SVLIM$ squares the difference between V_{85} and the speed limit. Therefore, even a small increase in mean speed along the corridor would noticeably increase the sum of squares of the difference between V_{85} and speed limit.

Acceleration noise (AN) decreases as more data filters are applied. This result is reasonable as the data filters increasingly removed additional potentially non-free-flow, which are the trips that that tend to have higher acceleration noise than the free-flow trips.

The speed band $(V_{95} - V_5)$ indicates the difference between the high and low speeds at a point location, which represents speed variation at a specific point on the road. The two speed band measures, namely, M_BND and SD_BND determine the magnitude and variation of the speed band along the corridor, respectively. The data filters seem to have the highest influence on these two measures. Both M_BND and SD_BND have a decreasing trend as more data filters apply.

The decreasing trend of M_BND implies that the size of speed band decreases as more potentially non-free-flow trips are removed in the filtering process. The measure

SD_BND also has a decreasing trend. The results shows that the variation of the speed band decreases as more of potentially non-free-flow trips are removed.

To visualize the sensitivity of the speed band by data filter, the speed band profiles of South Atlanta Road Southbound are plotted in Figure 59. The speed band is bounded by the 95th (V_{95}) and 5th (V_5) percentile speed profiles. The "before" speed band is denoted by two black solid lines and the "after" speed band is denoted by two red dashed lines. Interestingly, this figure shows that the data filters influence mostly the low speed percentile and rarely to the high speed percentile. The V_5 speed profile changes dramatically after the data filtering process while the V_{95} speed profile remains mostly the same. Additionally, the V_5 pattern became similar to the V_{95} after the data filtering process. In conclusion, the size of speed band (M_BND) decreases after the data processing because V_5 increases while V_{95} remains relatively constant. Furthermore, the speed band has less variability (SD_BND) along the corridor because the V_5 pattern becomes similar to the V_{95}.

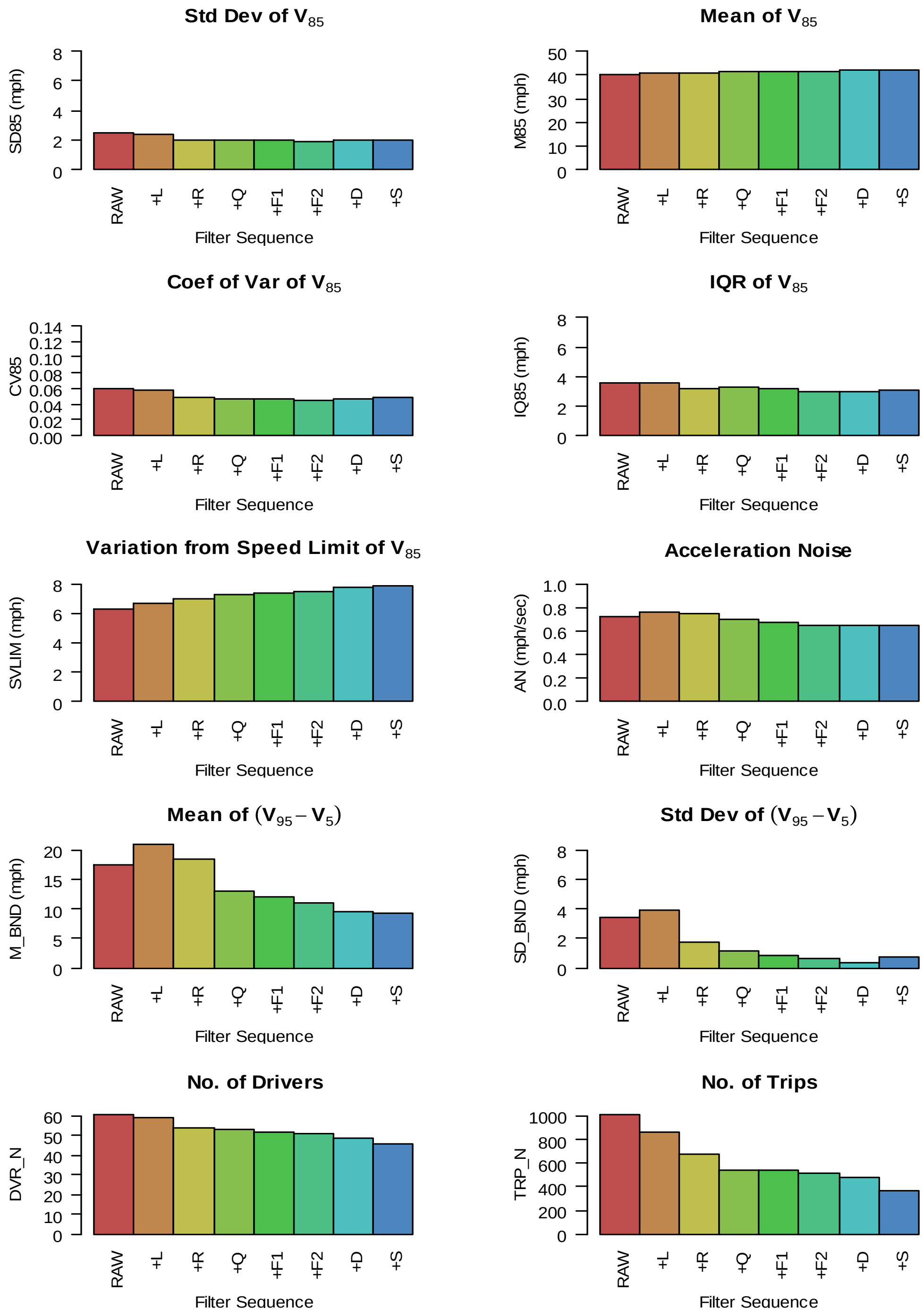


Figure 52: Sensitivity of Sequential Data Filters to the Traffic Attributes on Corridor 03 Eastbound (L=Light, R=No Rain, Q=Queue, F1=Free-Flow Type I, F2=Free-Flow Type II, D=Deviated Trips, S=GPS Signal)

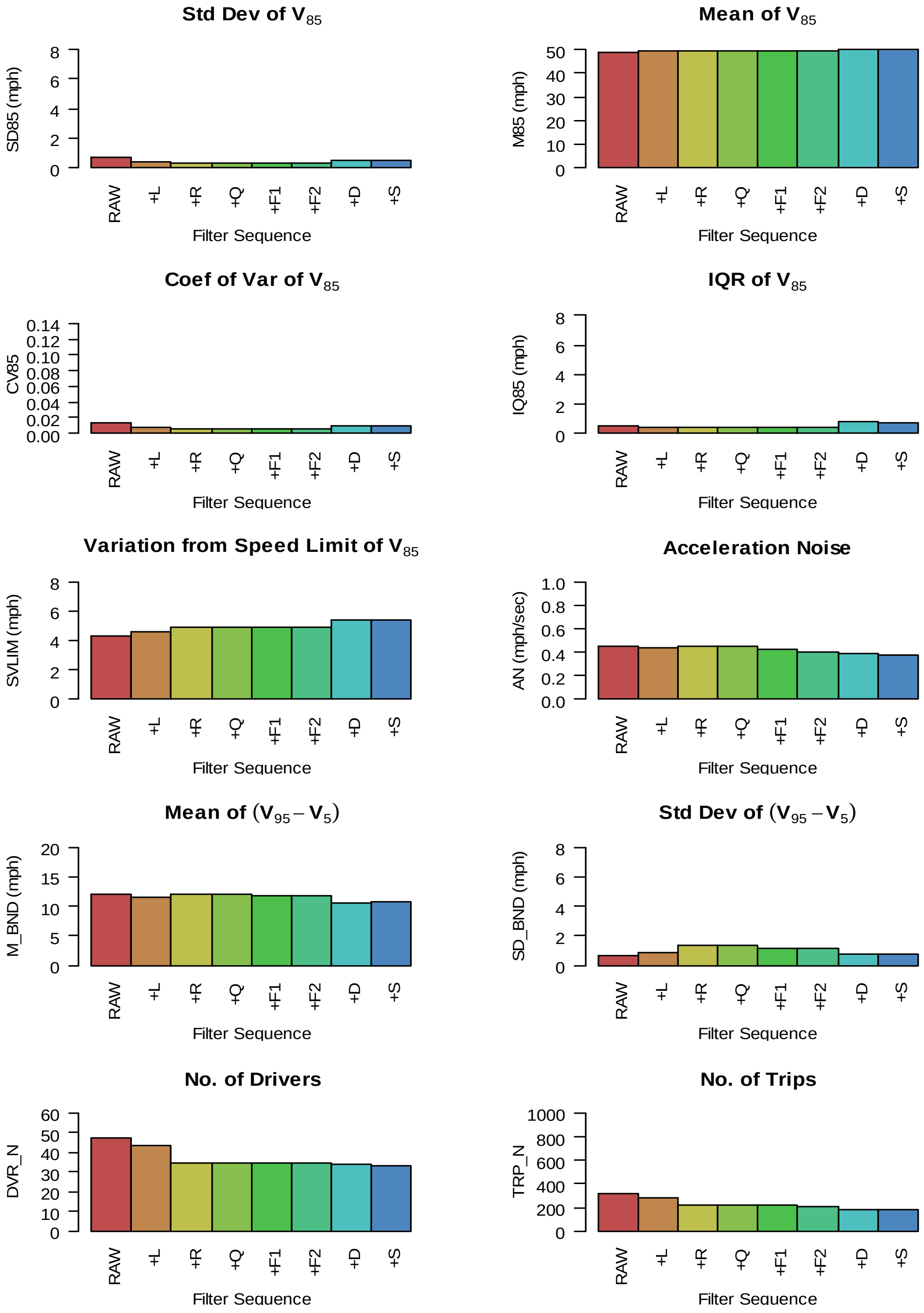


Figure 53: Sensitivity of Sequential Data Filters to the Traffic Attributes on Corridor 26 Westbound (L=Light, R=No Rain, Q=Queue, F1=Free-Flow Type I, F2=Free-Flow Type II, D=Deviated Trips, S=GPS Signal)

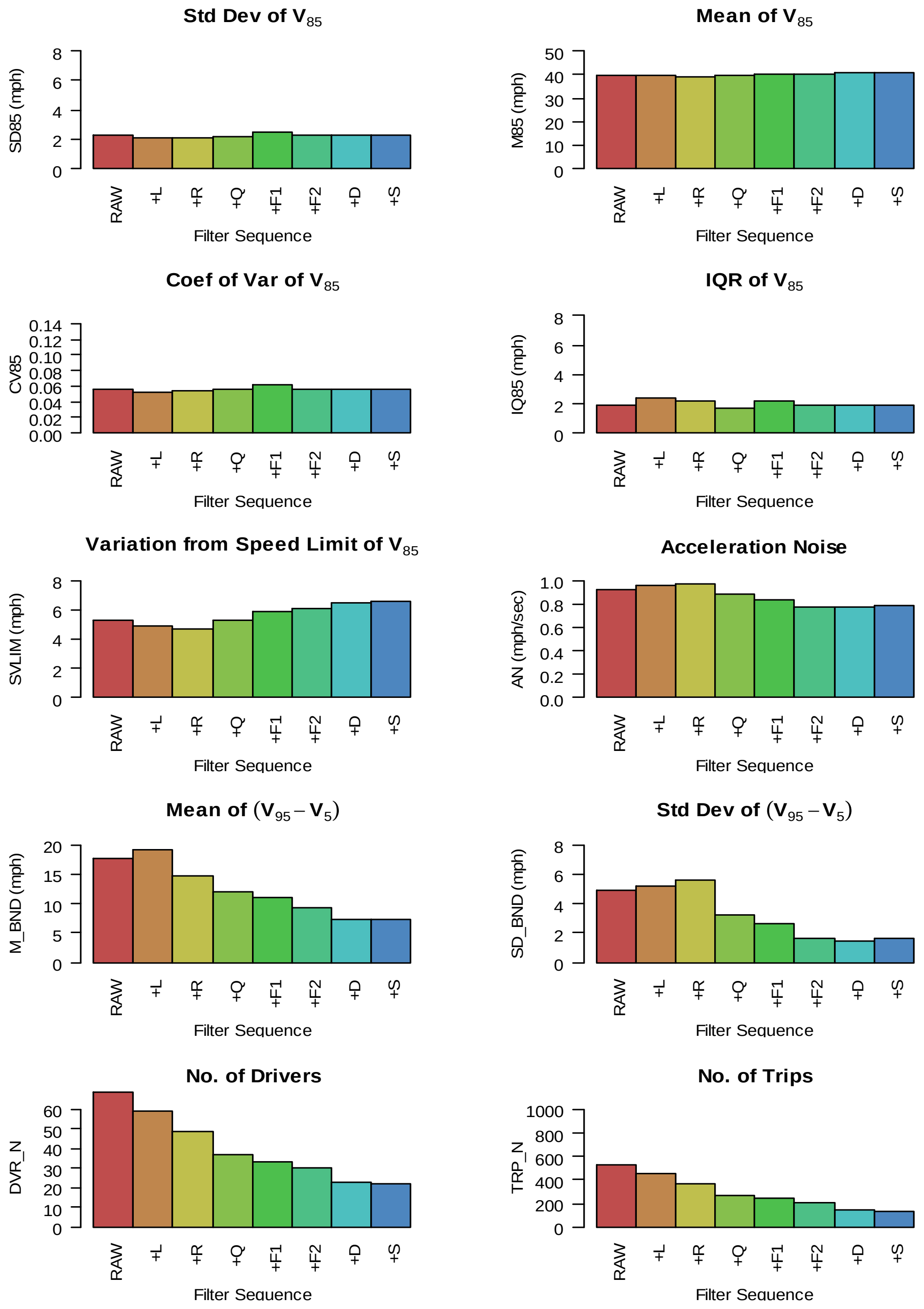


Figure 54: Sensitivity of Sequential Data Filters to the Traffic Attributes on South Atlanta Road Southbound (L=Light, R=No Rain, Q=Queue, F1=Free-Flow Type I, F2=Free-Flow Type II, D=Deviated Trips, S=GPS Signal)

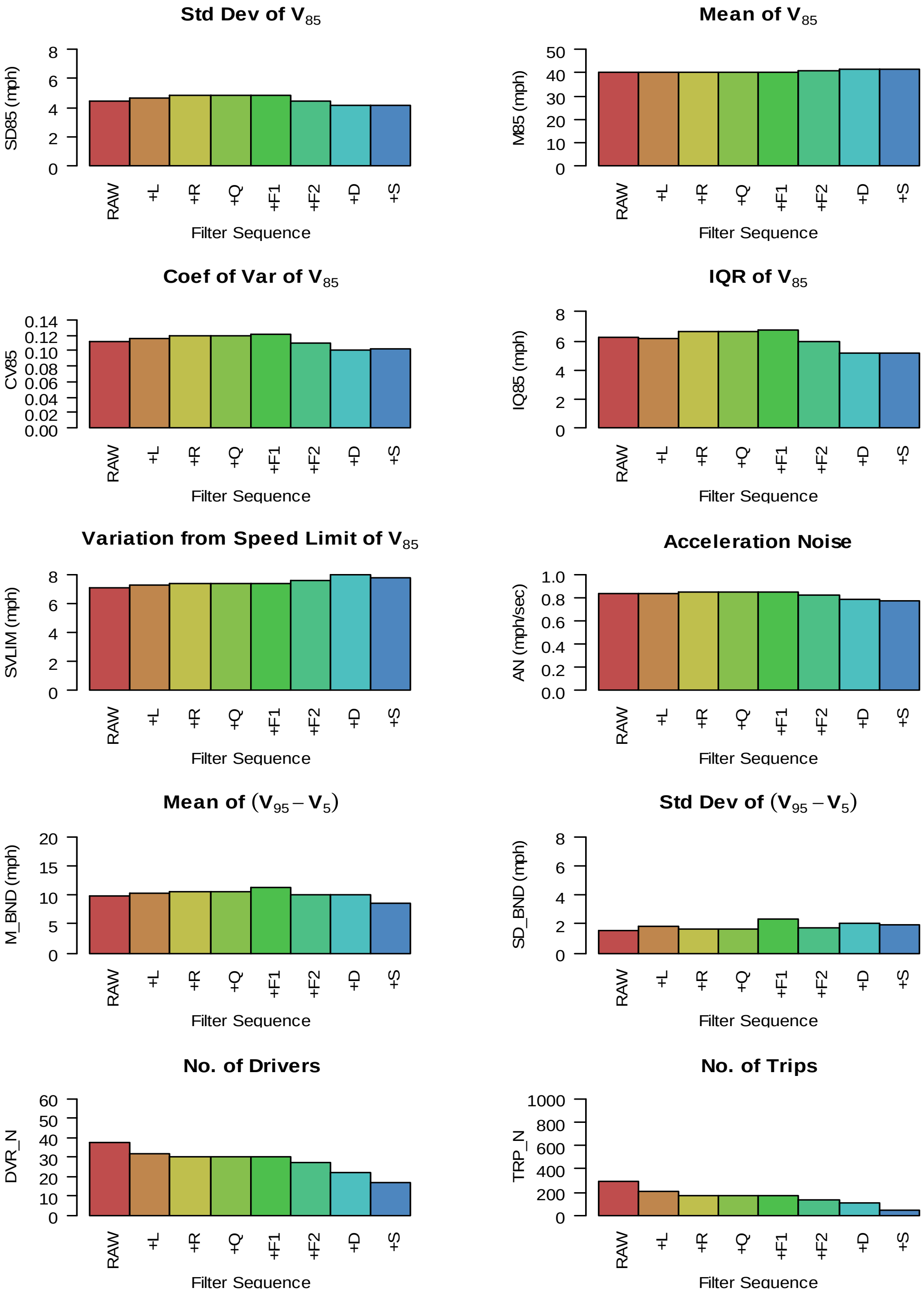


Figure 55: Sensitivity of Sequential Data Filters to the Traffic Attributes on Corridor 40 Eastbound (L=Light, R=No Rain, Q=Queue, F1=Free-Flow Type I, F2=Free-Flow Type II, D=Deviated Trips, S=GPS Signal)

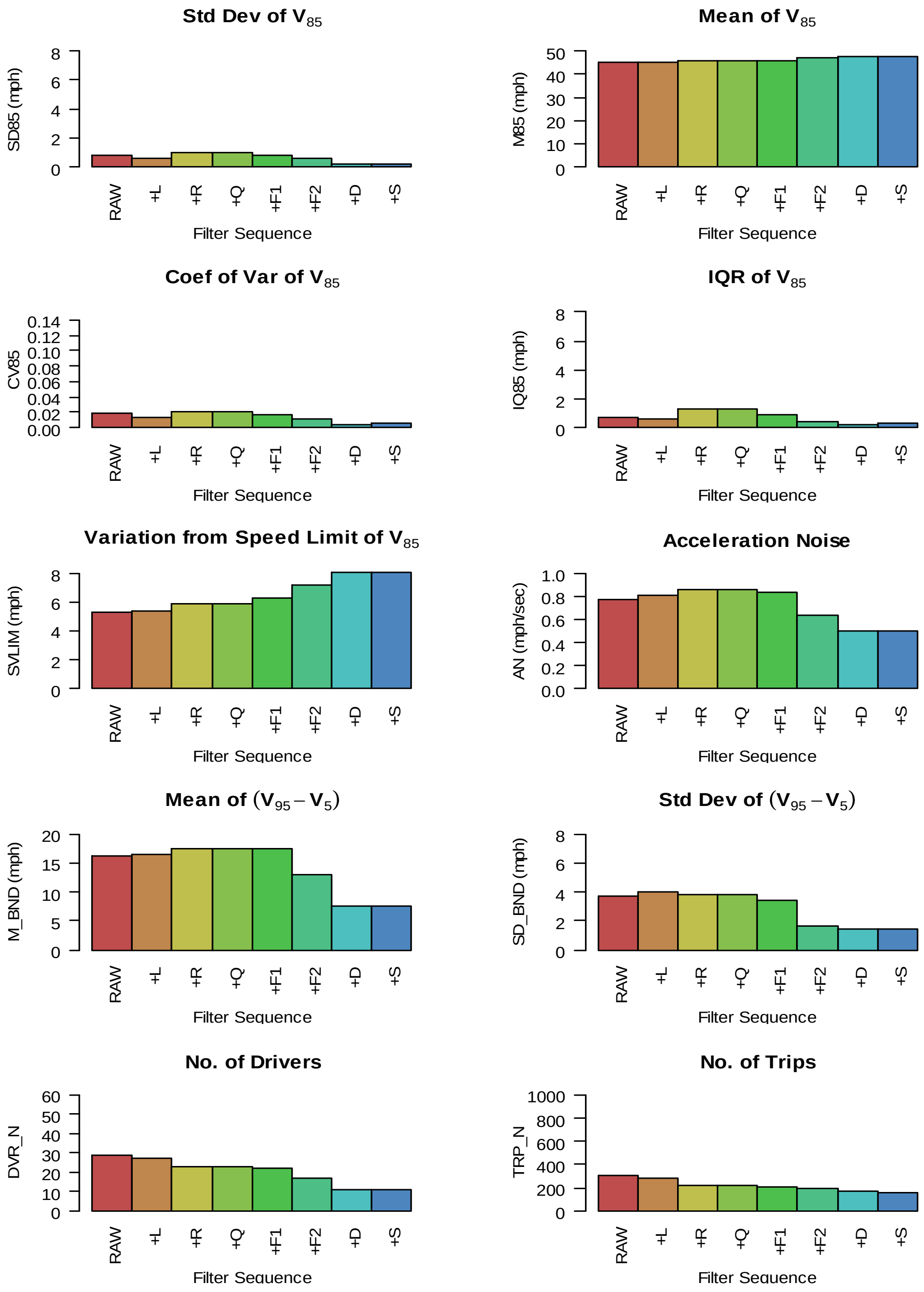


Figure 56: Sensitivity of Sequential Data Filters to the Traffic Attributes on Corridor 71 Westbound (L=Light, R=No Rain, Q=Queue, F1=Free-Flow Type I, F2=Free-Flow Type II, D=Deviated Trips, S=GPS Signal)

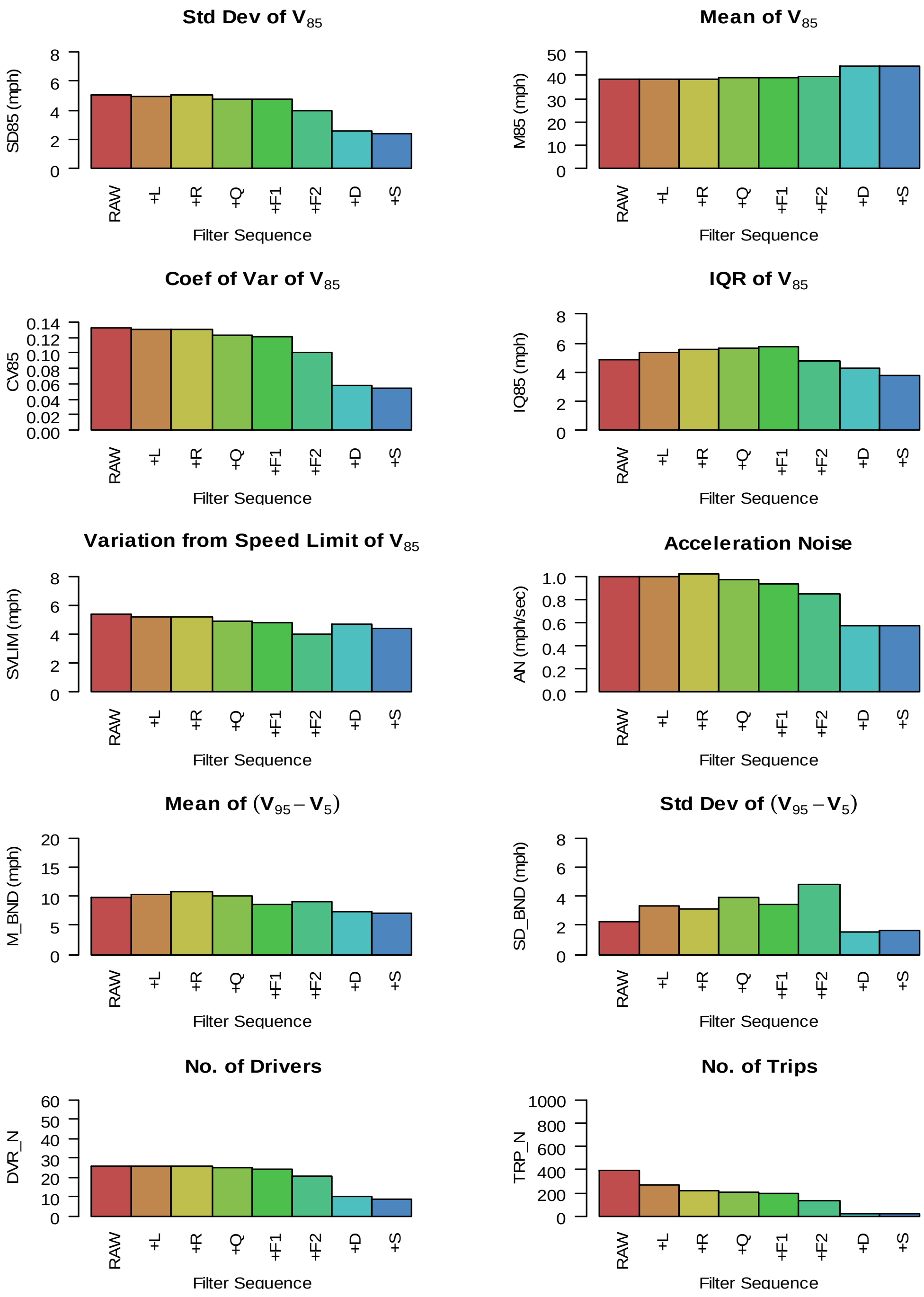


Figure 57: Sensitivity of Sequential Data Filters to the Traffic Attributes on Corridor 92 Southbound (L=Light, R=No Rain, Q=Queue, F1=Free-Flow Type I, F2=Free-Flow Type II, D=Deviated Trips, S=GPS Signal)

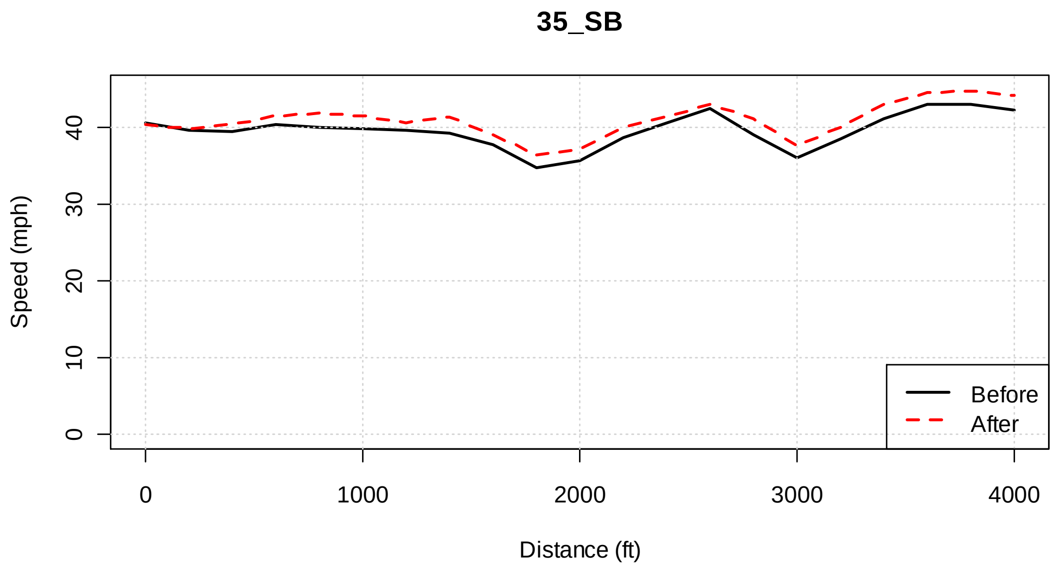


Figure 58: The 85th Percentile Speed Profile of South Atlanta Road Southbound Before (Solid Line) and After (Dashed Line) Applying All Filters

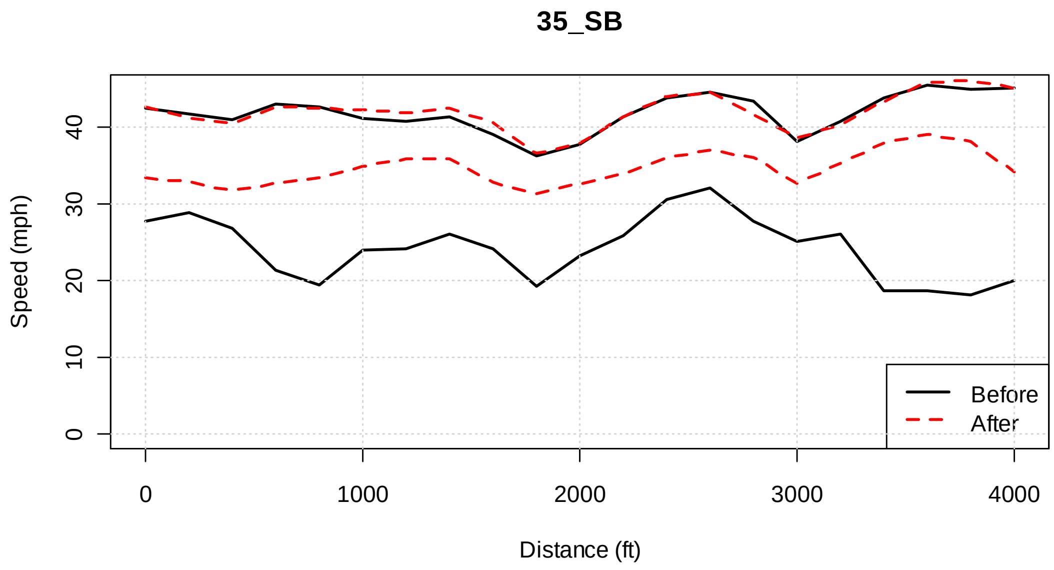


Figure 59: The Speed Band (V_{95}-V_5) of South Atlanta Road Southbound of Before (Bounded by Solid Lines) and After (Bounded by Dashed Lines) Applying All Filters

6.3 SENSITIVITY TO INDIVIDUAL FACTORS

In this section, the filters of interest are switched on and off from the filter set to determine how each filter affects the speed parameters.

6.3.1 Methodology

Eleven scenarios (numbers 0, 1,2…10) were modeled to gain an understanding of the influence of individual filters. The testing factors and filter combinations are described in Table 14 and the filter codes are depicted in Table 15. The “Code” column in Table 14 indices what filters are turned on or off compared with the base scenario “LRFS”. The plus sign indicates the filter is turned on and the minus sign indicates the filter is turned off. For example, the code “-F+O” denotes that the scenario number 10 has the free-flow filter turned off and the off-peak filter turned on, compared with the base case filter sequence, LRFS. Therefore, the filters applied in this combination are L, R, O, and S.

In Table 14, the base scenario (No.1) has the filters L, R, F and S on while other filters, namely, W, O, and D are turned off. Scenario No.0 is when there are no filters applied to the speed data, accounting for a calculation of the value of speed measures when no data filtering processes are involved. Scenarios 2 to 8 test the influence of the daylight, weekday/weekend, inclement weather, free-flow condition, off-peak period, highly deviated trip pattern, and GPS signal quality, respectively. Scenario 9 tests the sensitivity of speed measures when the GPS signal filter is applied first, instead of being the last filter. The last scenario, number 10, tests whether the off-peak period can be used as a surrogate of the free-flow pattern algorithm since, in some situations, the speed profile data may not be available.

Note that in this section we combined three free-flow filters, namely, Q, FF1, and FF2 as free-flow filter (F). This is necessary as these filters are designed to be run together and in sequences. In addition, the off-peak filter (O) does not appear in the previous analysis. The off-peak period is defined in Table 16. The off-peak periods exclude the pre-defined peak periods which are the morning peak (7:00-9:00), the midday peak (11:00-13:00), and the evening peak (16:00-19:00). The Weekday (W) filter is also included to determine the difference between a full week (Monday to Sunday) and weekday only (Monday to Friday) traffic characteristics.

Table 14: Planning Matrix to Determine Individual Factor Effects, 11 Runs

No.	Effect	Filter Sequence	Δ Base	L	W	R	F	O	D	S
0	Raw Data		-LRFS							
1	Base	LRFS	LRFS	x		x	x			x
2	Daylight	RFS	-L			x	x			x
3	Weekday	LWRFS	+W	x	x	x	x			x
4	Rain	LFS	-R	x			x			x
5	Free-Flow	LRS	-F	x		x				x
6	Off-peak	LRFOS	+O	x		x	x	x		x
7	Highly-Deviated Trips	LRFDS	+D	x		x	x		x	x
8	GPS Signal	LRF	-S	x		x	x			
9*	GPS Signal Sequence	SLRF	SLRF	x		x	x			x
10	Off-peak, No F Filter	LROS	-F+O	x		x		x		x

*Note that Run 9 applied the same set of filters as Run 1; however, GPS Signal filter was used at the first step in Run 9 to determine the sensitivity of the speed measures due to the sequence of filters.

Table 15: Filter Code Description

Code	Filter
L	Daylight Filter
W	Weekday Filter
R	No Rain Filter
F	A Set of Free-Flow Filters (Q+F1+F2)
O	Off-Peak Filter
D	Highly Deviated Trip Filter
S	GPS Signal Check Filter

Table 16: Off-Peak Periods

Off-Peak Period	Time Period
1	Midnight-7:00
2	9:00-11:00
3	13:00-16:00
4	19:00-midnight

6.3.2 Results and Discussion

The sensitivity analysis results for the six selected corridors are presented in the bar chart format from Figure 60 to Figure 65. Each figure contains ten bar charts: eight are the results of speed measures and the other two charts show number of drivers and number of trips left from the filtering process. An individual bar chart contains 11 bars, which report the values of the same speed measure from 11 filtering combinations listed in Table 14. The denotation of the filter code is listed in Table 15. The values of every bar should be compared with the base case, LRFS to determine the effect of the factor that has been turned on or off.

The followings are the results and comparisons between each of the ten filter combinations and the base scenario.

No Filter (-LRFS)

In most cases, the speed measures related to V_{85} changed only slightly after all the filters were removed. The measures, $SD85$, $CV85$, and $IQ85$ increase a little. This means that V_{85} only slightly more variable if we do not use the filters. The average of V_{85} along the corridor ($M85$) remains stable with or without the filters.

The design consistency parameter ($SVLIM$) has a decreasing trend due to lower V_{85} when all the filters are removed.

The higher value of acceleration noise (AN) is consistent among the six corridors. This is a reasonable result as the acceleration noise is expected to be lower under free-flow traffic condition (*42*).

Regarding the speed band measures, the size (M_BND) and variability (SD_BND) of speed band ($V_{95} - V_5$) increases significantly after removing the data filters for most of the corridors.

In summary, the “-LRFS” and the “LRFS” provide similar results for $SD85$, $M85$, $CV85$, and $IQ85$ as these variables are not highly affected by the non-free-flow traffic. However, the “-LRFS” scenario yields significant higher speed band measures and acceleration noise while giving a lower value of $SVLIM$. Therefore, it is reasonably acceptable to derive $SD85$, $M85$, $CV85$, and $IQ85$ from the raw speed data.

The followings are the summary of the effects of the individual filters to the free-flow data set. Note that since every corridor has different speed characteristics, the summary is the findings from the overall results, rather than corridor-by-corridor descriptions.

Light Condition (-L)

The comparison between daylight (LRFS) and all day (RFS) speed data show that the speed measures have similar values during the day and night. This implies that the potentially free-flow trips during the day and night have similar characteristics.

Weekday Only (+W)

The variation measures seem to be a little lower when including only the weekday free-flow trips in the data. This is likely because weekday and weekend trips have different characteristics. Combining the two distributions increase the variability in the dataset.

Weather Condition (-R)

Removing the weather filter does not significantly affect the speed measures of the likely free-flow trips. This is likely due to small number of trips being made during the rain condition.

All Traffic Condition (-F)

The result is in line with expectation. That is, the free-flow filters remove the speed variability due to the traffic condition. The measures of speed variability increases as the free-flow filters are removed.

Off-Peak Only (+O)

Adding the fixed off-peak filter does not affect the speed measures of the already likely free-flow trips. This means the free-flow filters effectively remove the non-free-flow trips during the fixed peak time.

Highly Deviated Trips (+D)

The highly deviated trip filter does not significantly affect the speed measures of the likely free-flow trips.

No Signal Quality Check (-S)

With or without signal check filter does not affect the speed measure results significantly.

Check GPS Signal Quality First (SLRF)

The GPS signal quality check filter is used to determine the percent time an individual trip contains “good” quality GPS signal data (the criteria are described in 4.1.10). If more than 20 percent of the signal data is identified as low quality, the trip is identified as having low quality data and removed from analysis. This filter is strategically placed after the acceleration/deceleration zone filter so that low quality data in these zones would not be included. That is, as the low quality data in the acceleration/deceleration zones is removed before the signal check filter, this trip has a higher chance to pass the 80 percent good signal criteria. The assumption is that significant portion of low quality data tend to be found under stop conditions, which are more likely to occur at intersections. As this study is concerned with midblock performance, intersection data, and thus the source of much of the low quality data, will be removed as part of the standard filters.

The SLRF filter combination places the signal check filter as the first filter. The SLRF case shows very similar results to the LRFS case.

Off-Peak Period as a Free-Flow Filter (-F+O)

This scenario tests the possibility to use the off-peak period as a free-flow filter.

The results show that the -F+O filter does not reduce the variation as significantly as the free-flow filter. In some corridors, there is even a higher speed variation than the raw data case. One reason for this behavior may be that the fixed off-peak period might not match with the real off-peak time on those corridors. Therefore, the off-peak filter might remove free-flow condition trips while leaving non-free-flow condition in the dataset. As a result, the speed data become more variable. In summary, the fixed off-peak filter should not be used as a surrogate for the free-flow filter.

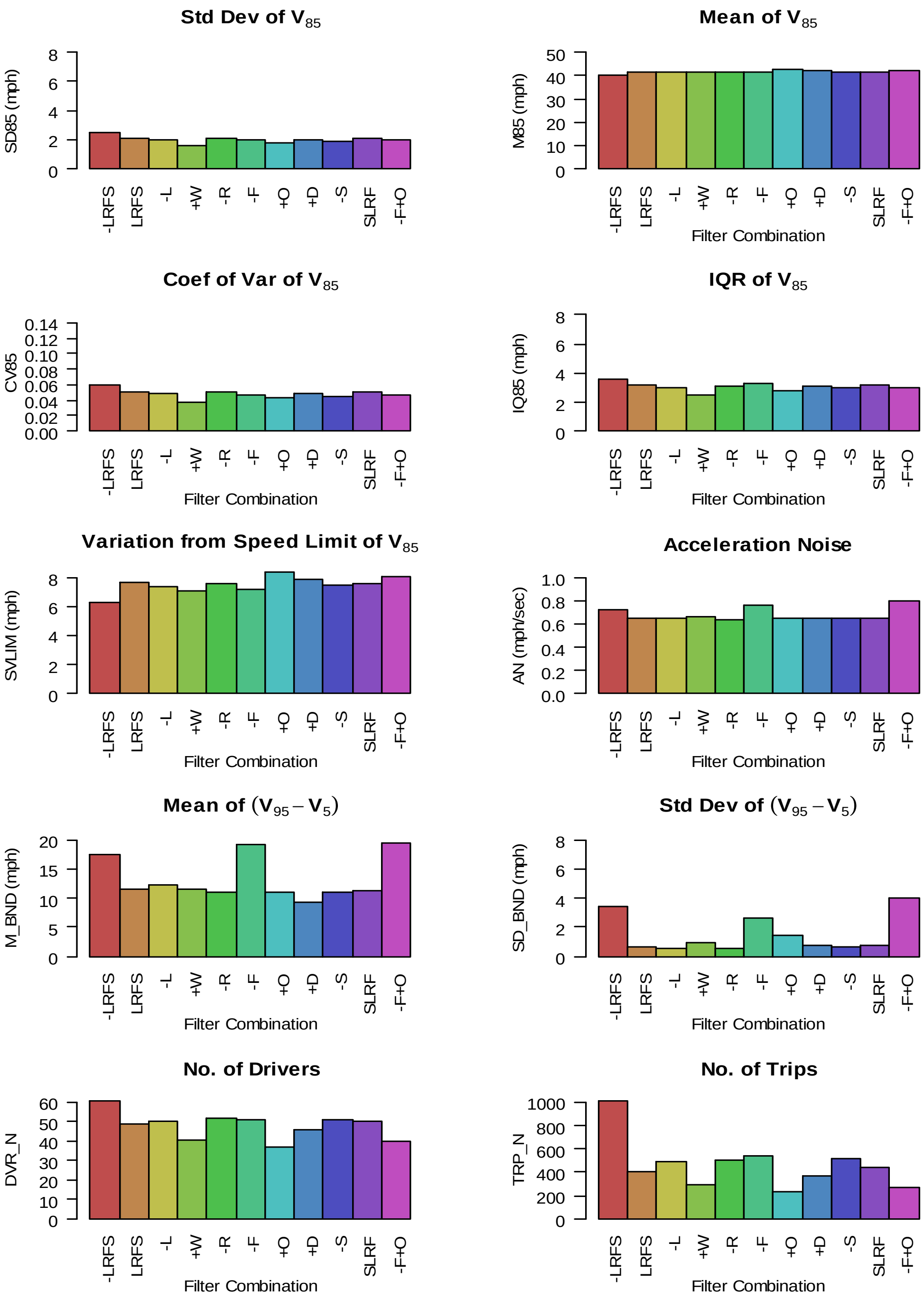


Figure 60: Sensitivity of Data Filters to the Traffic Attributes on Corridor 03 Eastbound (L=Light, W=Weekday, R=No Rain, F=Free-Flow, O=Off-Peak, D=Deviated Trips, S=GPS Signal)

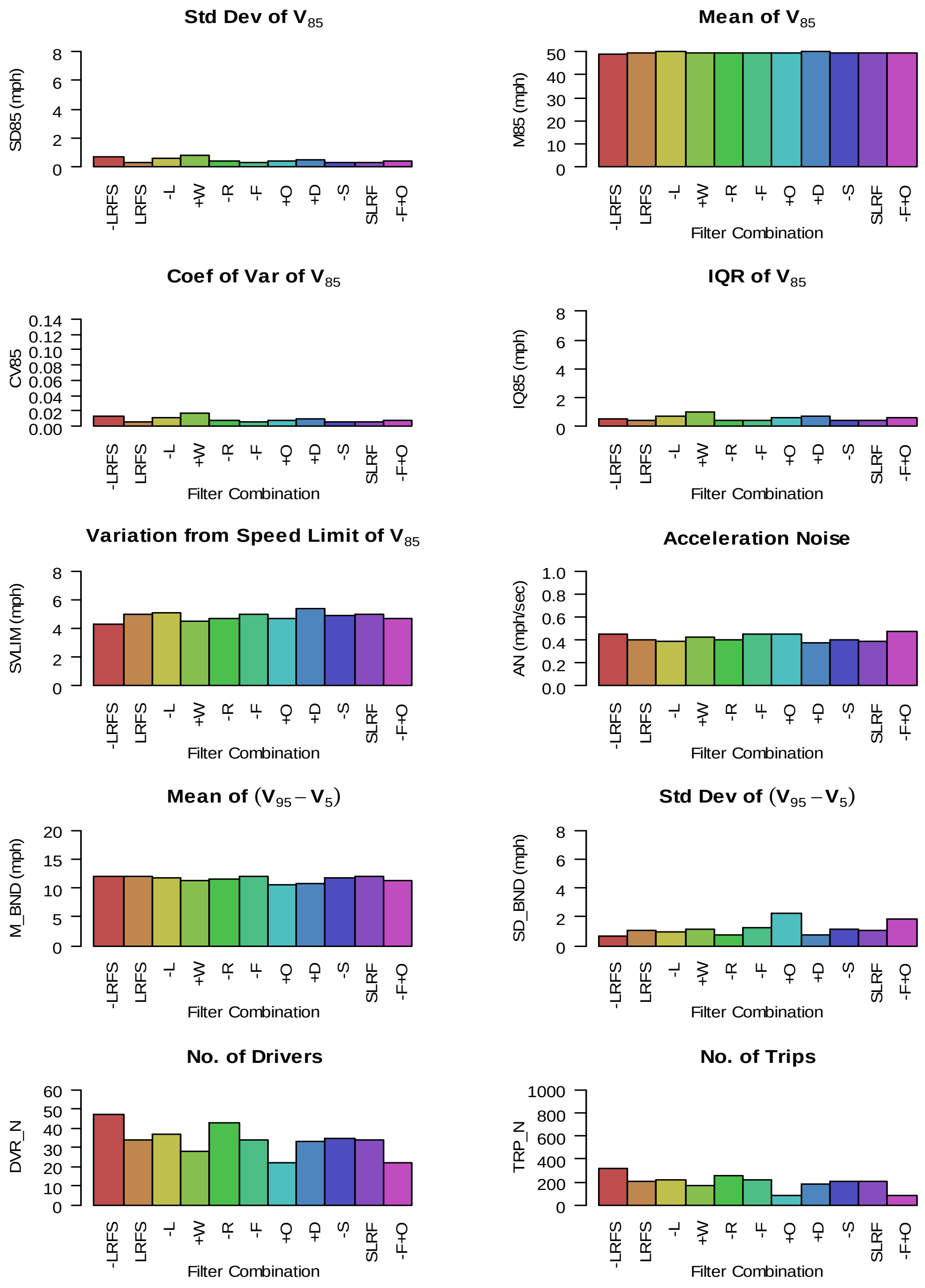


Figure 61: Sensitivity of Data Filters to the Traffic Attributes on Corridor 26 Westbound (L=Light, W=Weekday, R=No Rain, F=Free-Flow, O=Off-Peak, D=Deviated Trips, S=GPS Signal)

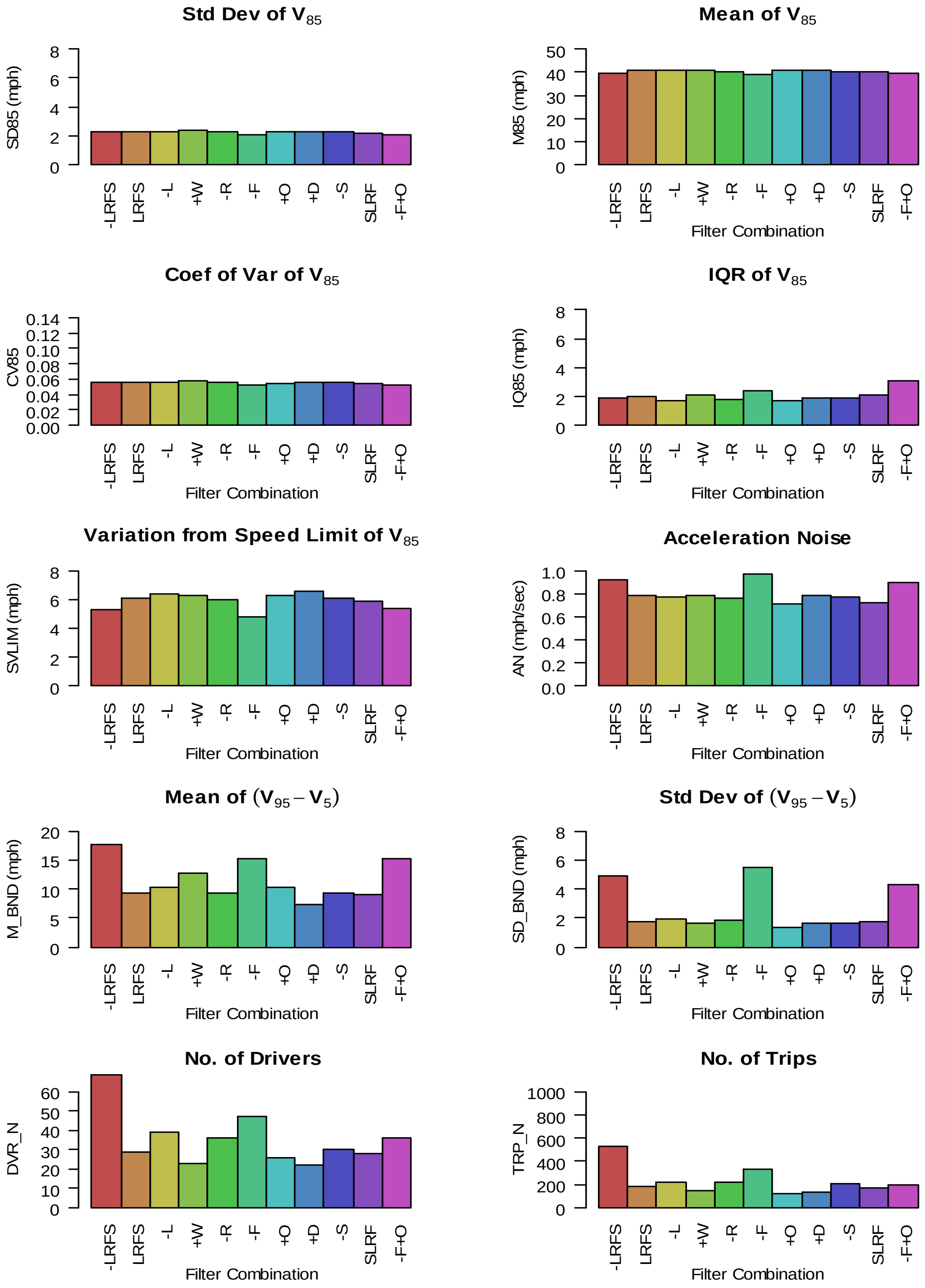


Figure 62: Sensitivity of Data Filters to the Traffic Attributes on South Atlanta Road Southbound (L=Light, W=Weekday, R=No Rain, F=Free-Flow, O=Off-Peak, D=Deviated Trips, S=GPS Signal)

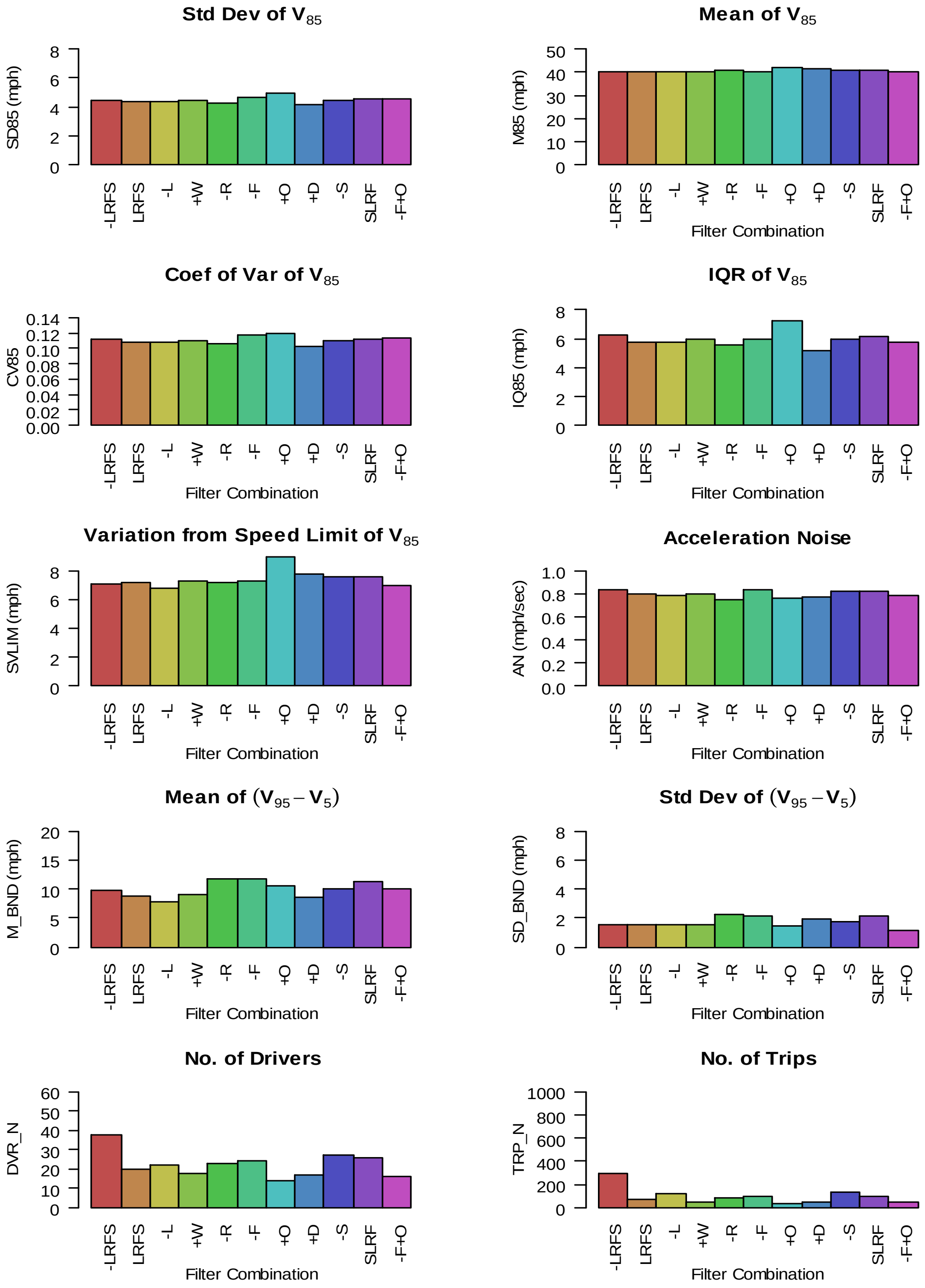


Figure 63: Sensitivity of Data Filters to the Traffic Attributes on Corridor 40 Eastbound (L=Light, W=Weekday, R=No Rain, F=Free-Flow, O=Off-Peak, D=Deviated Trips, S=GPS Signal)

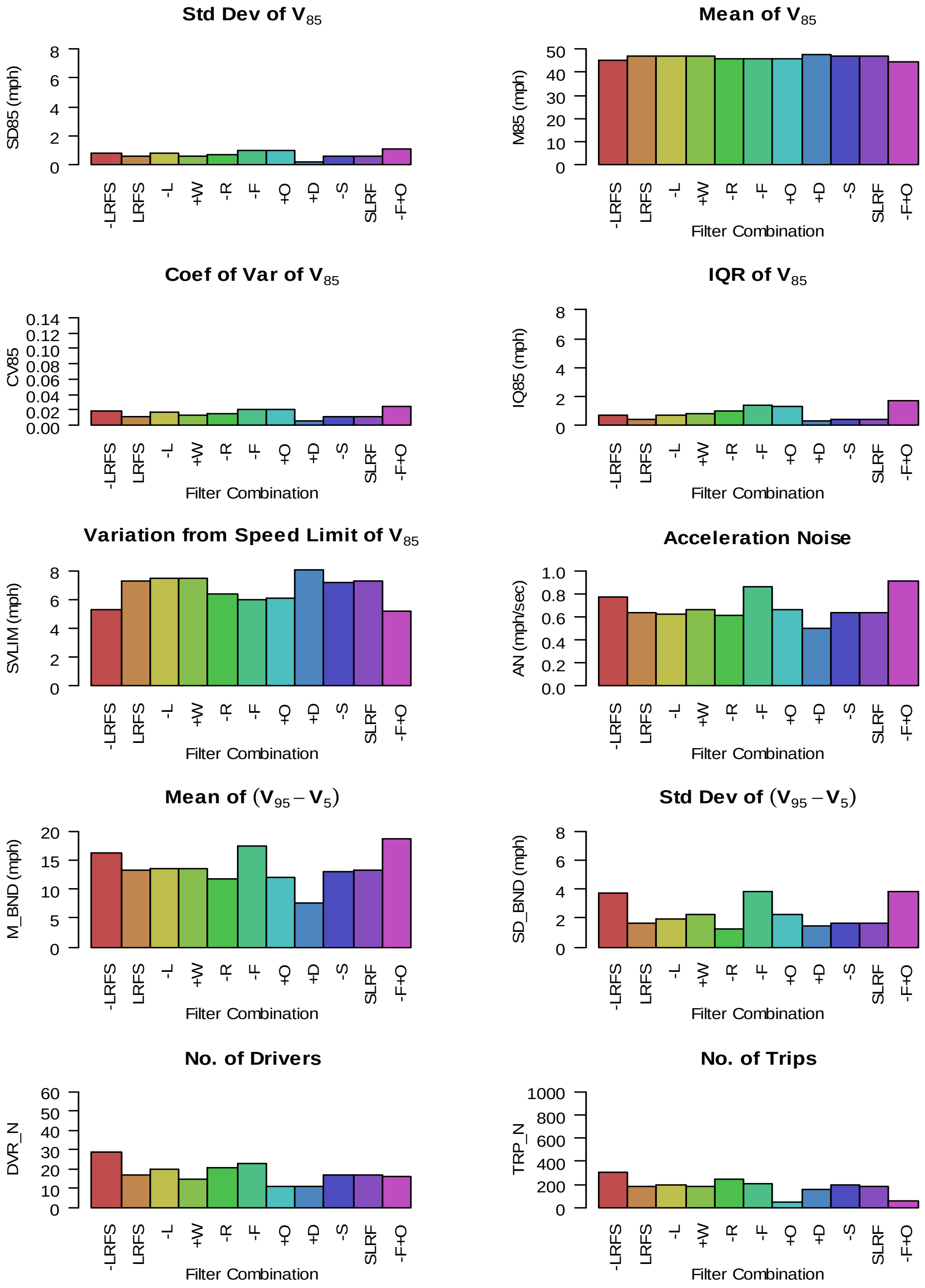


Figure 64: Sensitivity of Data Filters to the Traffic Attributes on Corridor 71 Westbound (L=Light, W=Weekday, R=No Rain, F=Free-Flow, O=Off-Peak, D=Deviated Trips, S=GPS Signal)

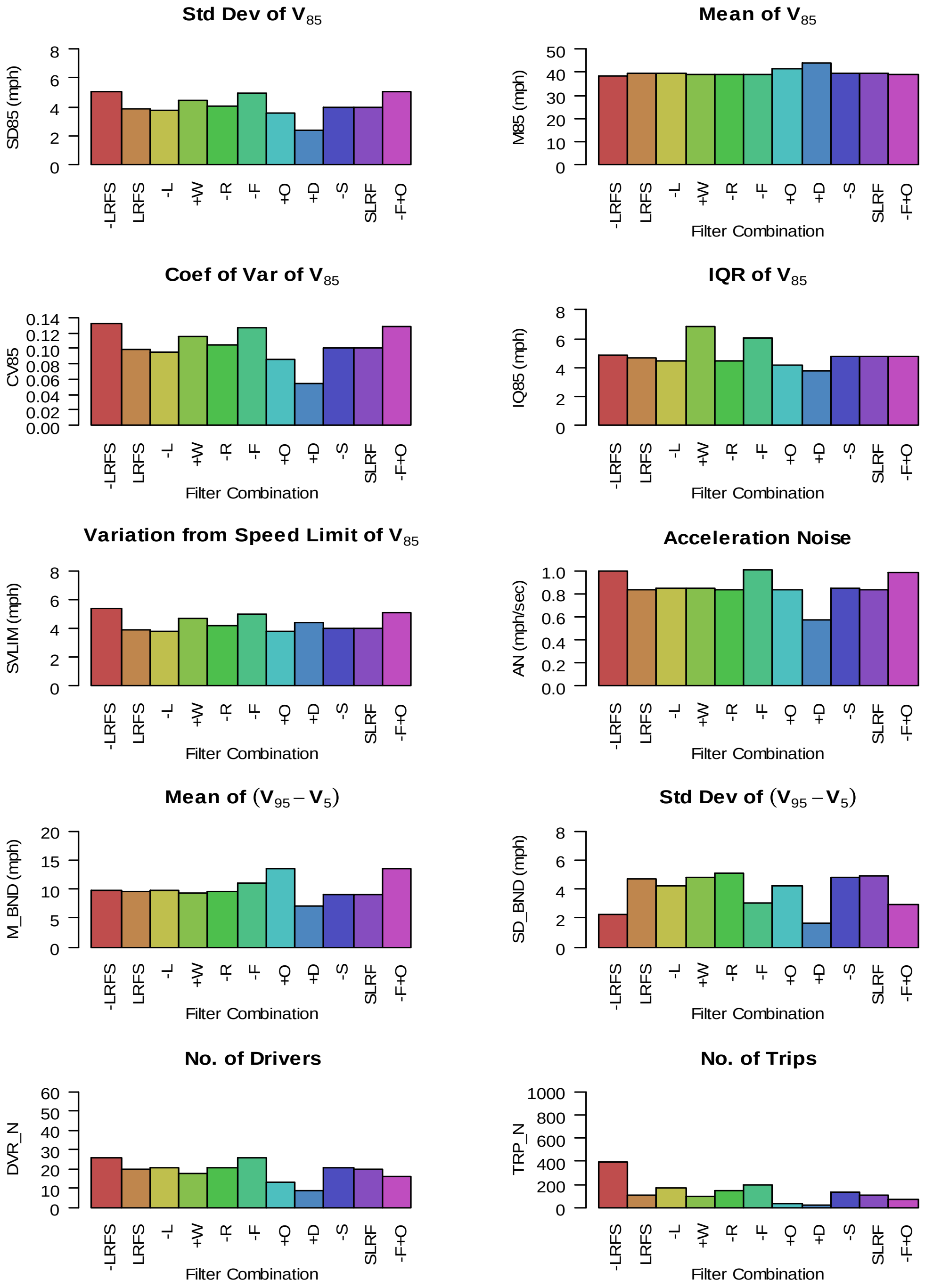


Figure 65: Sensitivity of Data Filters to the Traffic Attributes on Corridor 92 Southbound (L=Light, W=Weekday, R=No Rain, F=Free-Flow, O=Off-Peak, D=Deviated Trips, S=GPS Signal)

6.3.3 Summary

This chapter discussed the sensitivity of speed measures to the data filters used in the data processing. The analyzes were performed in two manners: the sensitivity of the speed measures as filters are sequentially added the sensitivity of the speed measures to individual filters.

The findings from the *sequential filtering analysis* are summarized below:

- The speed measures derived from V_{85} seem to have little sensitivity to the sequential data filters. This is because the filters were designed to remove non-free-flow trips, which usually contain low speed data points. Therefore, the filters have little impact on the V_{85} profile.

- The measure AN tends to decrease as additional filters are applied. This is because each subsequent filter removes additional the variability from the data. However, the variability reduction due to removing the non-free-flow trips seems to be insignificant. It is possible that the method by which the AN is calculated in this study smoothes out the noise from traffic congestion. That is, acceleration noise from multiple trips made at different times of day and different days of the week made by the same driver were averaged to obtain the representative AN for that driver. Driver's acceleration noises were then further averaged to obtain the representative AN of the corridor. Therefore, the AN in this study is not substantially affected by the non-free-flow condition.

- The measures derived from the speed band ($V_{95} - V_5$), such as M_BND and SD_BND, were more significantly impacted by the filters than the above measures. This is because the filters substantially influence the low speed element, V_5, of the speed band. More specifically, the filters increase V_5 but not V_{95} along the corridor, therefore, the size of speed band (M_BND) reduces substantially. In addition, the filters remove the variability of the V_5 due to potential traffic congestion resulting in the V_5 profile along the corridor similar to the V_{95} profile. As a result, the size of speed band becomes more consistent along the corridor, which in turn reduces the variance of the band, SD_BND, along the corridor.

- The measure SVLIM tends to increase as more filters are applied. This is because V_{85} slightly increases along the corridor after the potentially non-free-flow trips were removed. Therefore, the V_{85} is further away from the speed limit.

- When comparing the traffic attributes of different corridors, the variance measures seem to be consistent. More specifically, a corridor with relatively high values of $SD85$ tends to have high $CV85$, $IQR85$, and AN as well.

The findings from the *individual effect analysis* are summarized below. The comparisons are made between the LRFS case and the other test cases.

- The individual effect analysis showed a similar result as the sequential analysis for the speed measures derived from V_{85}. That is, these speed measures seem to have little sensitivity to the sequential data filters. This is because the filters were designed to remove non-free-flow trips, which usually contain low speed data points.

- The daylight effect does not have a significant impact on the speed measures. It is unknown if the study corridor had street lighting that may be impacting performance. Also, the sample size of the night time trips are relatively small, e.g., 20 percent of the trips on corridor 03 Eastbound were made during the night time.

- The variation measures seem to be a little lower when only weekday free-flow trips are considered. This is likely because weekday and weekend trips have different characteristics.

- When removing the free-flow filter set, the measures SD85, CV85, IQR85 and SVLIM seems to be only slightly impacted while the speed band measures, namely M_BND and SD_BND increase significantly.

- The sequence of applying GPS signal filter does not affect the result.

- The off-peak filter does not typically yield the same result as the free-flow filter. This is likely because different corridors have different peak period which can vary daily and seasonally even on the same corridor.

- It is seen that the +D filter (removing highly deviated trip) might overly reduce the variation of the speed data. For example, the SD85 of corridor 71 Eastbound and 92 Southbound were reduced by 56 percent and 40 percent, respectively. The variation from the road geometries might be lost due to this filter. Therefore, the highly deviated trip filter should be removed from the filtering process.

Chapter 7. MODEL DEVELOPMENT

This chapter discusses the model development effort. In this effort, crash frequency per unit of roadway length is the predicted variable and the traffic attributes in Chapter 5 along with roadway classification, corridor length, and traffic volume are used to construct the prediction model. In the first section of this chapter, road facilities are classified using a combined road classification and traffic volume criteria. Section 7.2 explores the distribution of the dependent variable, crash frequency. The relationship between crash frequency and ADDT is investigated in Section 7.3. The regression tree technique used to determine the predictor variables is discussed in Section 7.4. The model development methodology is described in Section 7.5. and the results are summarized and in Section 7.6.

Thus far in this research effort, speed measures have been calculated by direction of travel as significantly different speed characteristics were often observed in the opposing traffic directions on many corridors. The different speed characteristics are likely a result of differences in road features such as land uses, driveway density, the direction of horizontal curvature, etc. This suggests that the crash prediction model should be constructed by direction. However, ninety-three percent of the study corridors are undivided roads and the direction of travel at crash impact cannot be accurately determined without the police crash reports. At the time of this research, the crash reports are available only from 2004-2005.

Given this limitation, the model considers the crash data from the two travel directions aggregated together. The speed measures are also combined using a weighted average by number of drivers in each direction. Further research, when crash reports become available, will seek to explore modeling power gained by considering each corridor direction separately.

7.1 ROAD CLASSIFICATION FOR SAFETY

The literature on crash prediction models often group study locations by categorical variables such as traffic control type, divided/undivided, and functional classification prior to the model calibration (*32*). One possible grouping explored for this research was by function classification. The 61 final corridors can be classified into three groups based on the GDOT's road functional classification system, namely, minor arterial, collector, and local street (note that the data is originally draw from the FHWA project, the Effects of Urban Street Environment on Operating Speeds, which include no major arterial data). However, the road characteristics might not be accurately represented by the GDOT's current functional classification due to changes in land use and traffic volume over time. Therefore, this section investigates whether the study roadways grouped by GDOT's classification demonstrate similar speed and safety characteristics.

Table 17 describes the study corridor characteristics in terms of average traffic volume, and corridor length, and road classifications. On average, the minor arterial corridors have the highest traffic volume, followed by collectors and local streets. In terms of corridor length, the three road classifications have approximately the same section length.

Table 17: Road Characteristics by Functional Classification

Functional Classification	No. of Corridors	AADT			Corridor Length (ft)		
		Min	Max	Average	Min	Max	Average
16	25	12,465	38,325	23,801	1,746	5,575	3,319
17	23	5,160	21,660	13,059	1,992	5,143	3,295
19	13	1,096	19,557	9,811	2,408	5,672	3,349
Total	61	1,096	38,325	16,770	1,746	5,672	3,316

7.1.1 Definitions

Because the information on the GDOT's functional classification is limited, the definitions from multiple sources are considered in this study. Functional classification is defined by the FHWA (*51*) as "the process by which streets and highways are grouped into classes, or systems, according to the character of service they are intended to provide." NCHRP Report Number 504 (*45*) also described the characteristics of each class in detail. Figure 16 illustrates the excerpt from this report.

Table 18: Typical Characteristics for Urban Road Classifications, Excerpt from (*45*)

Functional Classification	Anticipated Speed	Service	Typical Cross Section
Minor Arterial	35-55 mph	Balances between mobility and access	Multilane divided or undivided
Collector	30-50 mph	Connects local roads to arterial	2-3 lanes with curb and gutter
Local Streets	25-35 mph	Permits access to abutting land	Two lanes with curb and gutter

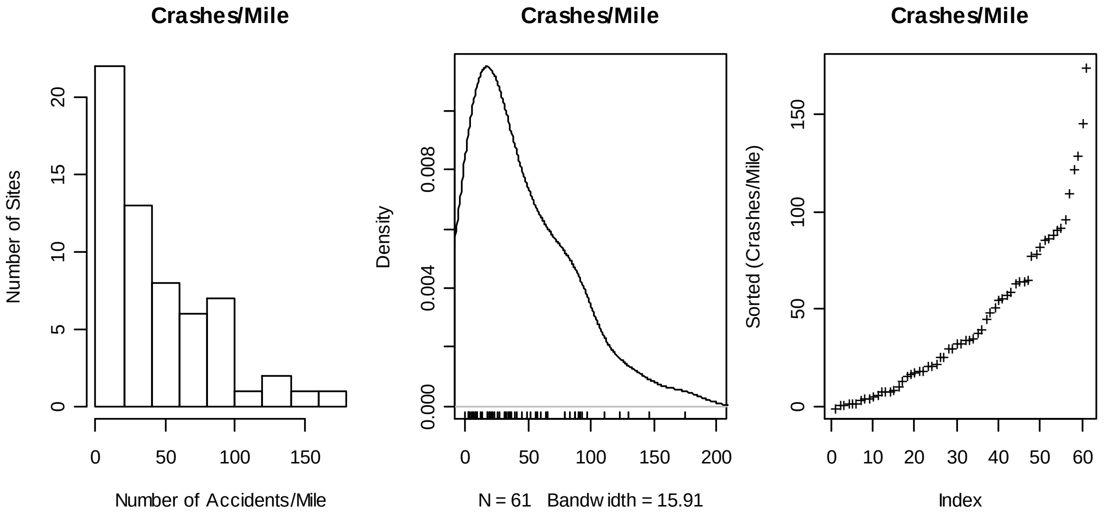
7.1.2 Data Exploration

The objective of this section is to understand the distribution of the crash data among the different facility types. The examination starts with plots of crashes per mile over all road classes. The histogram of the crashes per mile variable in Figure 66(a) shows a rough estimate of the density of the dependent variable, crashes/mile, which is clearly not normally distributed. Since the histogram plot is sensitive to the bin size, the kernel density estimation was also generated in Figure 66(b). The density estimator for a sample set X_1,..., X_n is of the form:

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h_n}K\left(\frac{x - x_i}{h_n}\right)$$

where K represents the kernel function and h_n is the size of the bandwidth analogous to bin size of the histogram plot. The optimal bandwidth can be calculated from the formula: $h_n^* \approx 1.06\hat{\sigma}n^{-1/5}$ (*52*).

The kernel density function suggests that the crashes/mile distribution is not unimodal and might have a few different distributions residing in this dataset, i.e., one having mode at approximately 20 crashes/mile and the other having its mode at approximately 80 crashes/ mile in the four-year time period. This separation might also be due to the difference in traffic and geometric characteristics of different facilities. The right panel of Figure 66 shows the individual data points by plotting sorted value of crashes/mile against its index. It is seen that the data points for many corridors have low crash frequency and about 4-5 corridors apparently have a higher crash frequency than other corridors.



(a) **(b)** **(c)**

Figure 66: Distribution of Number of Crashes per Mile for All Corridors (a) Histogram, (b) Kernel Density Estimate, and (c) Index Plot of the Sorted Values.

Next, we investigated road characteristics by their road functional classifications, namely, minor arterial, collector, and local street. The boxplot of number of accidents in four years of different road classifications is showed in the left panel of Figure 67. Apparently, one minor arterial has an extremely high number of crashes relative to the other corridors during the four-year period. There are also two outliers on the local street group while the collector group does not show any outliers. In addition, the crash distribution of minor arterial class is obviously different from the collector and the local classes, i.e., the median number of accidents on minor arterials is three-fold that of the collectors and six-fold that of the local streets.

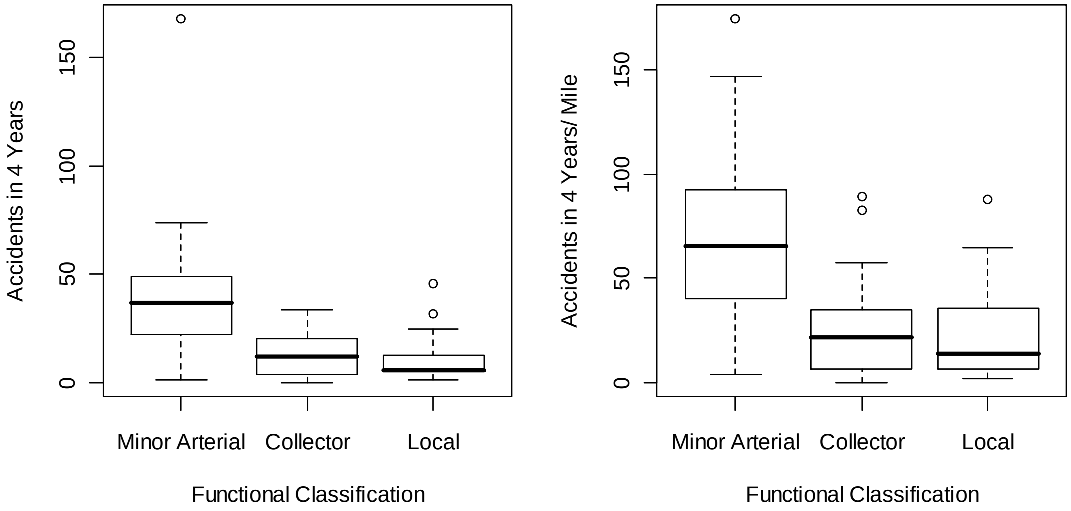


Figure 67: Boxplots of Number of Crashes (Left) and Number of Crashes per Mile across Different Road Classifications (Right)

When number of crashes is standardized by the corridor length, the distributions of crash frequency across classes remain proportionally constant (See Figure 67, right panel). That is, the median of the minor arterial is approximately three-fold that of the collectors and five-fold that of the local streets consistent with the volume differences. In addition, the first quartile of the minor arterial is only slightly larger than the third quartile of collector group. Unlike the total crash count variable, the variations of the crashes/mile variable between collector and local groups are more alike, i.e., the interquartile ranges of both groups are similar. Unlike the previous plot, two outliers from the collector are revealed and only one outlier is shown in the local group.

Other exposure measures such as traffic volume and corridor length might be helpful in distinguishing the separation in the crash distribution. The left panel of Figure 68 depicts the AADT distribution for the different road classifications.

Based on the AADT distribution, some collectors have AADT values similar to the minor arterial group and other collectors have AADT values similar to the local street group. The boxplot of corridor lengths in the right panel of Figure 68 shows that median lengths of the three road classifications are similar. The minor arterial group seems to have a wider range of corridor length than the collector and local street groups.

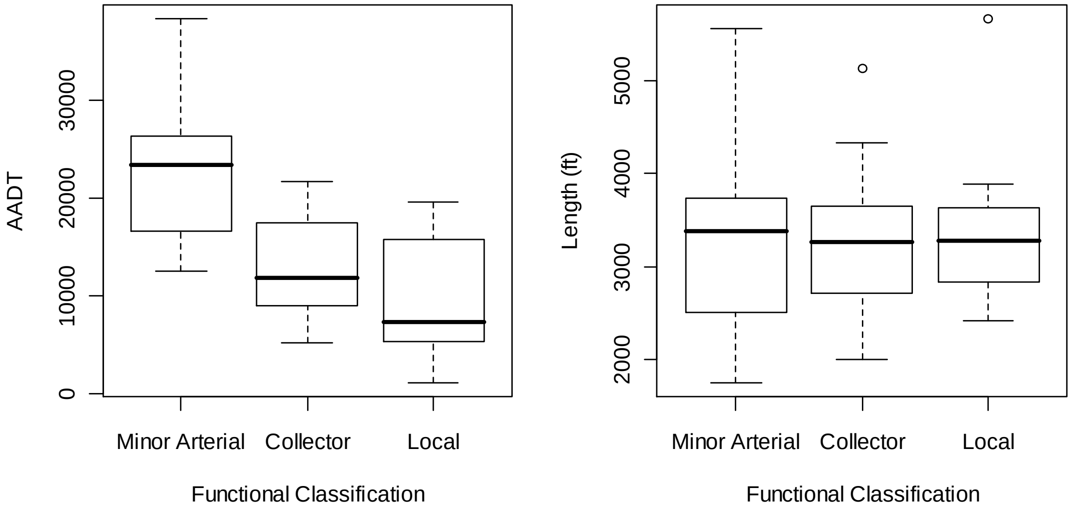


Figure 68: Boxplots of Traffic Volume (Left) and Corridor Length (Right)

In the next section (Section 7.1.3), we deployed a statistical method to classify the bimodal crash frequency distribution based on the roadway functional classification.

7.1.3 Technical Approach

The regression tree technique was used to stratify the distribution of crash frequency by the road functional classification variable and other traffic characteristic variables. This technique is often used in data exploratory analysis to determine how the data might be grouped, and which variables are important, what model form should be used (*32, 53*). The *rpart* (**R**ecursive **Part**itioning) package, which can be implemented in the R statistical analysis software, was used to construct tree models. The regression tree is built by first searching for a single predictor variable and its split point to obtain the "best fit", then separating the data into two groups (for binary partitioning) according to the criterion found in the first step. The process continues recursively until some stopping rule is satisfied (*54*). The tree model then simply uses an average of the responses at each node to represent its predicted value. In the *rpart* module, the split variable and its split value are chosen when the residual sum of squares (RSS) is minimized. The RSS is computed as:

$$RSS(partition) = RSS_L + RSS_R$$

where $RSS = \sum(y_i - \bar{y})^2$ and $\bar{y}$ is the mean of the response in each partition (*55*).

There are a few stopping rules that can be controlled in the *rpart* package (*56*):

- *minsplit*: the recursive binary partitions will stop when the number of observations in a node is less than the *minsplit* value.

- *minbucket*: the algorithm will not split if either the left or right branch will have the number of observations less than the *minbucket* value after the split.

- *cp* (the threshold complexity parameter): the algorithm will not split if the overall R-squared does not increase by at least *cp*. The R-squared is of the form (*39*, *53*, *55*):

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\sum_j \sum_{\mathrm{i}} \left(y_{i,j} - \hat{y}_j\right)^2}{\sum_j \sum_{\mathrm{i}} \left(y_{i,j} - \bar{y}\right)^2}$$

where $y_{i,j}$ is the observation i at the terminal node *j*

$\hat{y}_j$ is the predicted value of the response at the terminal node *j*

$\bar{y}$ is the grand mean of all the observations

7.1.4 Results

The regression tree results are showed in Figure 69. To grow a large tree, the cost complexity, *cp*, is set to be as low as 0.001. The minimum number of data points before splitting the node, *minsplit*, is 10 and the minimum number of data points after splitting, *minbucket*, is 5. Sixteen traffic attributes, traffic volume, and functional classification variable were included as input variables. Sixty-one observations are used as the input.

As a result, the variables used in this tree include acceleration noise during all traffic condition (AN_A) and free-flow condition (AN_F), traffic volume (AVG_AADT), stop frequency per trip per mile ($STOPS$), functional classification (FC), interquartile range of the 85th percentile speed ($IQ85$), and Moran's Index ($MORANS$). Among these variables, FC, AVG_AADT, and AN_F are the first few splitting variables in the tree. The root node error is the ratio of the total sum of squares of the dependent variable to the sample size.

The first split is on the FC variable, 36 corridors are either collector or local street with a mean response value of 26 whereas 25 corridors are minor arterials with the mean response value of 71. This supports our boxplot observation in Figure 67 that crash frequency on minor arterial is higher than the other two groups. After the first split, the total sum of squares reduced from 97,200 to 22,000+45,000 = 67,000.

Figure 70 visualizes the results in the tree diagram format. The tree diagram illustrates quite an interesting result. For local and collector group, the safety is determined by AVG_AADT, AN_A, $MORANS$, $IQ85$, and $STOPS$. All the mentioned variables, except $IQ85$, represents the level of traffic activity along the corridors. On the other hand, the safety on minor arterial is influenced by AN_F and AVG_AADT. The variable AN_F is the acceleration noise measured under the potentially free-flow condition; therefore, only the noises caused by driver and road geometries are included. The finding from this result (for the corridor included in the study) is that the safety on the minor arterial class is influenced by road geometries while the safety on the collector and local classes is influenced by traffic activity along the roadway. This is reasonable as the collectors and local streets have higher accessibility than the minor arterials.

The cost complexity table shows the value of cp parameter indicating how well the tree improves as the number of nodes or splits increases. The R^2 is improved by 31 percent after the first split, followed by 18 and 7 percent after the second and third splits, respectively. The relative error (*rel error*) is the ratio of SSE to SST, or equivalently $1 - R^2$, where R^2 is the usual coefficient of determination in linear regression. *xerror* and *xstd* denote the cross validation error and cross validation standard error. The plot on the left in Figure 71 shows that the first split (FC) offers the most information. The figure on the right also shows the 10-fold cross validation error with number of splits and suggests that the tree should include only the two splits.

```
Regression tree:
rpart(formula = CPL ~ P_THRU + P_FF + STOPS + MI_S100 + SD85 +
    SVLIM + M85 + SD_BND + M_BND + AN_AF + AN_FF + IQ85 + CV85 +
    CRAWL_P + FC + AVG_AADT, data = dat, method = "anova", control =
rpart.control(minsplit = 10,
    minbucket = 5, cp = 0.01, maxcompete = 4, maxsurrogate = 5,
    usesurrogate = 2, xval = 10, surrogatestyle = 0, maxdepth = 30))

Variables actually used in tree construction:
[1] AN_AF    AVG_AADT CV85     FC       M_BND    STOPS

Root node error: 97201/61 = 1593

n= 61

       CP nsplit rel error xerror   xstd
1 0.30846      0    1.0000 1.0223 0.2139
2 0.16555      1    0.6915 0.9843 0.2309
3 0.06587      2    0.5260 0.9429 0.2036
4 0.05290      3    0.4601 0.9728 0.1984
5 0.02765      4    0.4072 0.9512 0.1762
```

```
6 0.02113      5    0.3796 0.9854 0.1680
7 0.01648      6    0.3584 1.0096 0.1684
8 0.01000      7    0.3420 1.0563 0.1865
node), split, n, deviance, yval
      * denotes terminal node
node), split, n, deviance, yval
      * denotes terminal node

 1) root 61 97200.7400  44.691150  
   2) FC=Collector,Local 36 22252.9300  26.216130  
     4) AVG_AADT< 11511 20  4847.8720  14.287600  
       8) M_BND< 11.6122 14   640.0568   8.428776 *
       9) M_BND>=11.6122 6  2605.9430  27.958200 *
     5) AVG_AADT>=11511 16 11002.0200  41.126780  
      10) STOPS< 0.1195111 7  1991.3090  26.431320 *
      11) STOPS>=0.1195111 9  6323.2530  52.556580 *
   3) FC=Minor Arterial 25 44965.6900  71.295180  
     6) CV85< 0.03466276 14  9082.6310  48.806540  
      12) AN_AF< 0.670254 5  4920.0990  32.556410 *
      13) AN_AF>=0.670254 9  2108.6780  57.834400 *
     7) CV85>=0.03466276 11 19791.3700  99.917090  
      14) AN_AF< 0.8014879 6  9895.4850  80.180710 *
      15) AN_AF>=0.8014879 5  4754.1620 123.600700 *
```

Figure 69: Regression Tree Results of All Corridors

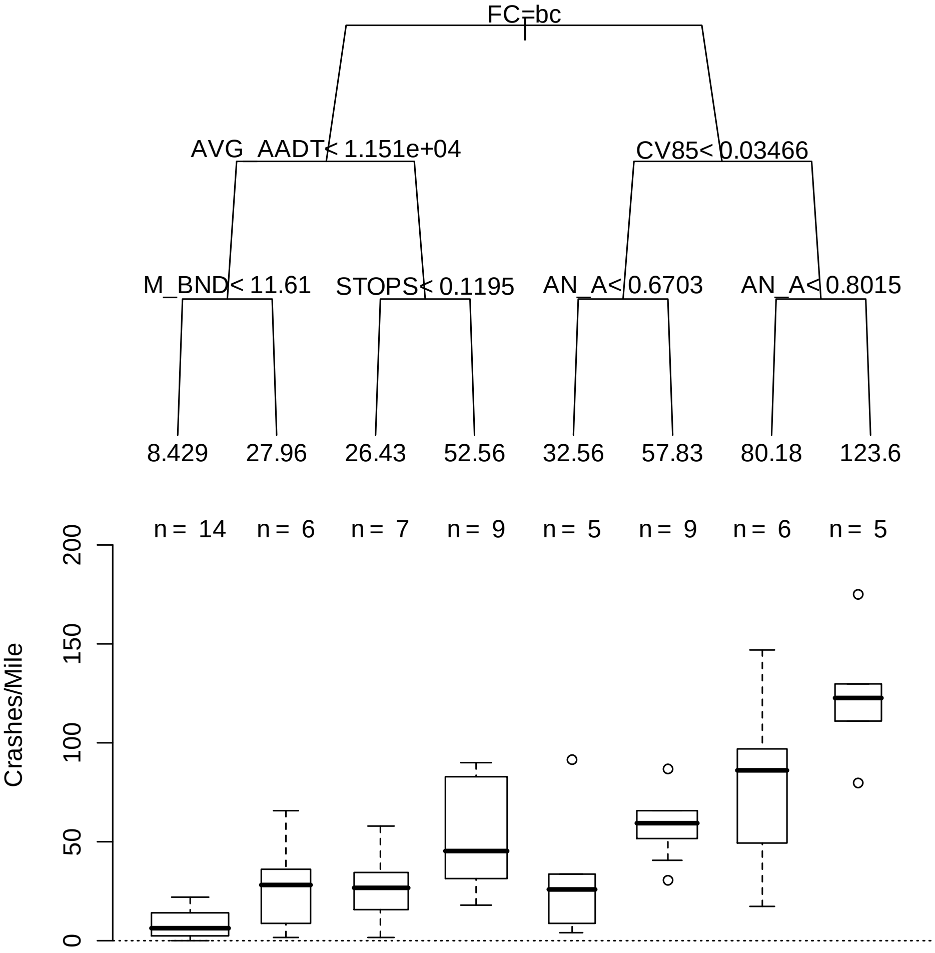


Figure 70: Regression Tree with Functional Classification (The Left Branch "bc" is the Collector and Local Street Group and the Right Branch is the Minor Arterial Group)

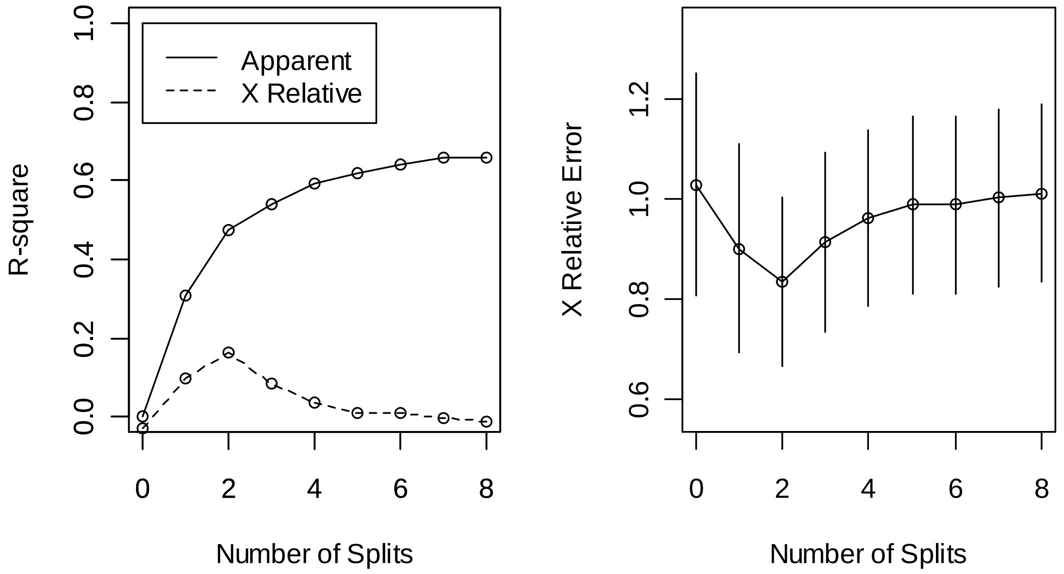


Figure 71: Plots of the R-Squared (Left) and the Relative Error from Cross-Validation (Right) for the Different Splits of All Corridors

According to the regression tree results, the minor arterials appear to be very different from the collectors and local streets. Using the functional classification as a split, we might be able to re-arrange the study corridors into two groups, namely, Higher Classification (HC) and Lower Classification (LC). The HC group would contain all the minor arterial corridors and some of the collectors while the LC group would contain all the local street corridors and the remaining collectors. The reasons of the rearrangement into the HC and LC group are as follows: (1) the minor arterial usually has a clear distinction from the local street; (2) the collector is harder to be distinguished from the local or minor arterial as the collector is the transition between the minor arterial and the local street. Some collectors might have characteristics more similar to the minor arterial while the other might have characteristics more similar to the local streets. To classify the collector corridors into HC or LC, a split variable and its split value was determined by fitting the regression tree using only the collector corridors. The results in Figure 72 show that the R-squared is improved by 37 percent using AVG_AADT<11,500 vehicles/day as the first split. Therefore, the split variable along with its split value of 11,500 vehicles/day will be used as a split criteria. More specifically, collectors with AADT less than 11,500 would be classified into LC group and the collector corridors with AADT at least 11,500 would be classified into HC group.

```
Regression tree:
rpart(formula = CPL ~ P_THRU + P_FF + STOPS + MI_S100 + SD85 +
    SVLIM + M85 + SD_BND + M_BND + AN_AF + AN_FF + IQ85 + CV85 +
    CRAWL_P + FC + AVG_AADT, data = subset(dat, FC == "Collector"),
    method = "anova", control = rpart.control(minsplit = 5, minbucket =
3,
        cp = 0.001, maxcompete = 4, maxsurrogate = 5, usesurrogate = 2,
        xval = 10, surrogatestyle = 0, maxdepth = 30))

Variables actually used in tree construction:
[1] AVG_AADT M_BND    SD_BND   STOPS    SVLIM

Root node error: 13878/23 = 603.4

n= 23


        CP nsplit rel error  xerror    xstd
1 0.380570      0   1.00000 1.09579 0.35736
2 0.342674      1   0.61943 1.14338 0.38116
3 0.078719      2   0.27676 0.67320 0.24323
4 0.033785      3   0.19804 0.69743 0.18464
5 0.031562      4   0.16425 0.71897 0.18298
6 0.011033      5   0.13269 0.72232 0.18611
7 0.001000      6   0.12166 0.74804 0.18958
n= 23

node), split, n, deviance, yval
```

```
     * denotes terminal node

1) root 23 13878.220000 27.345000
  2) AVG_AADT< 11511 11  1215.826000 11.517430
    4) M_BND< 8.643337 5    27.842320  4.365471 *
    5) M_BND>=8.643337 6   719.104300 17.477390
     10) STOPS< 0.0518867 3   168.840300  8.933114 *
     11) STOPS>=0.0518867 3   112.236200 26.021670 *
  3) AVG_AADT>=11511 12  7380.764000 41.853610
    6) SD_BND>=1.22708 9  2019.974000 30.360000
     12) SVLIM< 6.707976 3   261.921400 14.778850 *
     13) SVLIM>=6.707976 6   665.578100 38.150570
       26) SD_BND>=1.891415 3     9.195342 33.098950 *
       27) SD_BND< 1.891415 3   503.269400 43.202200 *
    7) SD_BND< 1.22708 3   605.079200 76.334440 *
```

Figure 72: Regression Tree Results of Collector Corridors

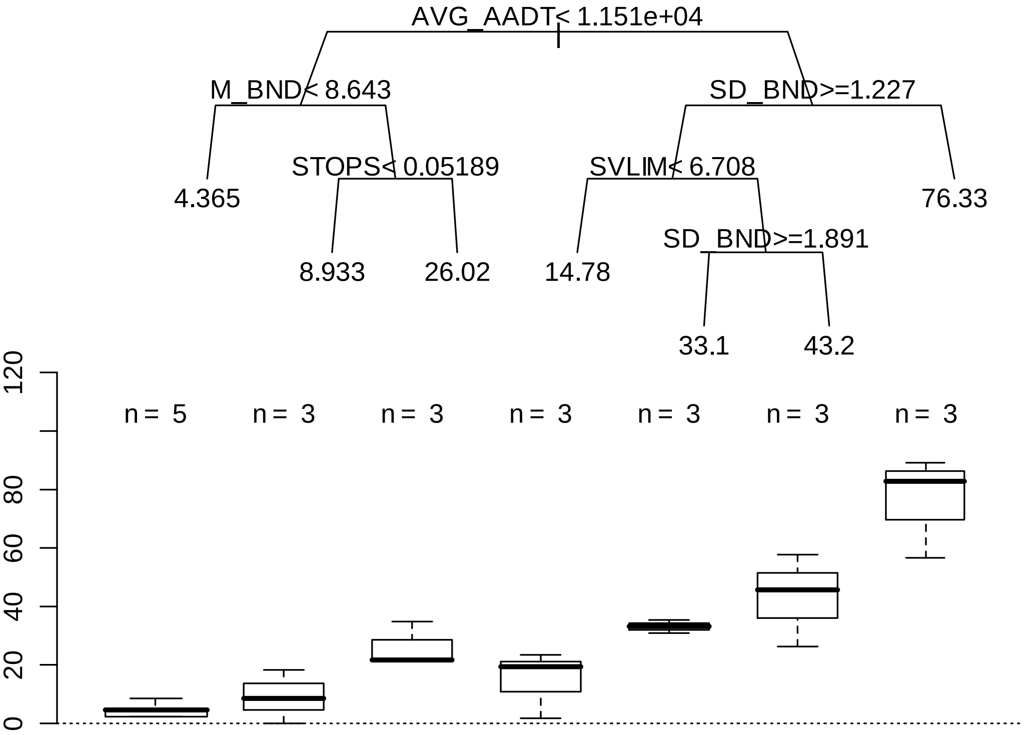


Figure 73: Regression Tree for Collector Corridors using All Traffic Characteristic Variables

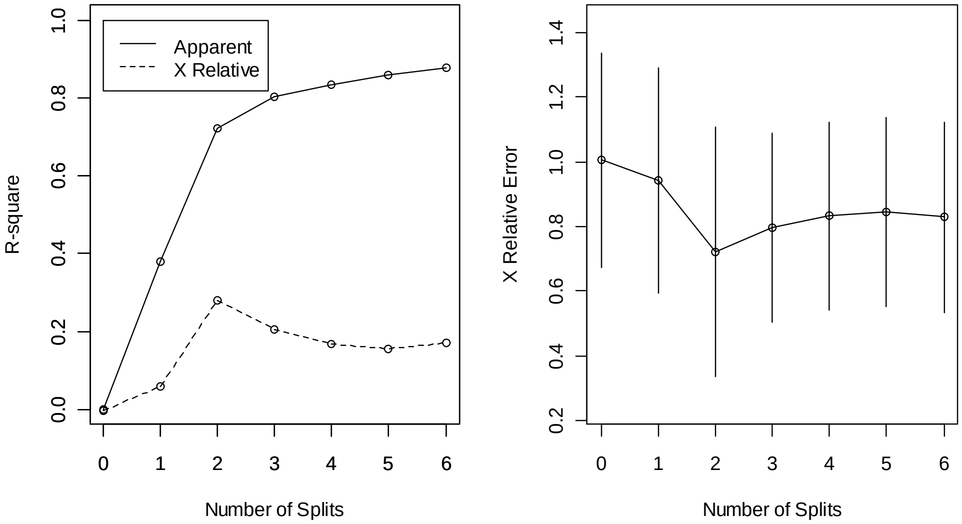


Figure 74: Plots of the R-Squared (Left) and the Relative Error from Cross-Validation (Right) for the Different Splits of Collector Corridors

The scatter plot in Figure 75 supports the regression tree result. The number of crashes per mile tends to be higher when AADT is greater than 11,500 vehicles/day (separated by a vertical dashes line).

The regression tree shows that there is a distinct difference in the crash characteristics between the facility with AADT less than and greater than 11,500 vehicles/day. That is, the collector with AADT less than 11,500 has a mean response value of 11 while the facility with AADT greater than 11,500 has a mean response value of 41. Consequently, a new functional classification system was proposed. The collectors with AADT less than 11,500 vehicles/day and all local streets were grouped together and called the “Lower Classification” (LC). The Minor Arterials and the collectors with AADT greater than 11,500 vehicles/day are defined as the “Higher Classification” (HC).

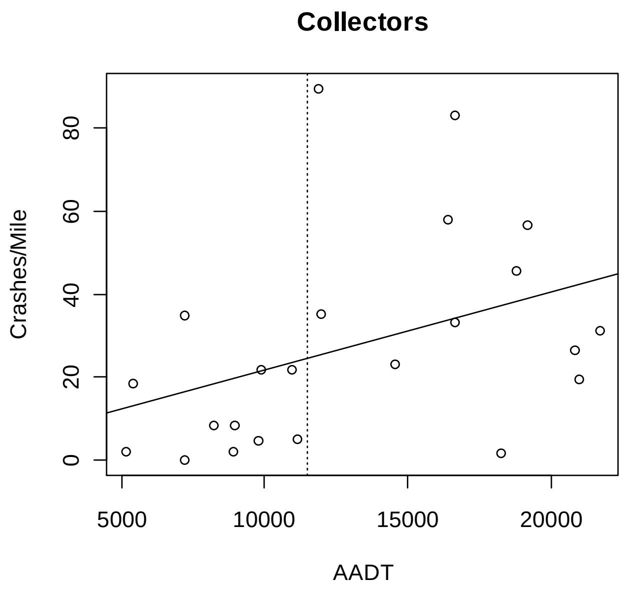


Figure 75: Scatter plot of Crashes/Mile vs. AADT on Collector Corridors with Regression Fit (R-Squared = 0.157), and the dotted line separating data at AADT = 11,500 vehicles/day

From Table 19, it is seemed that the average AADT of the HC group is more than twice of the LC group. In addition, the average corridor length for the HC group is similar to that of the LC group.

Table 19: Sample Size, AADT, and Corridor Length by New Road Classification

Functional Classification	No. of Corridors	AADT			Corridor Length (ft)		
		Min	Max	Average	Min	Max	Average
HC	37	11,873	38,325	21,692	1,746	5,575	3,328
LC	24	1,096	19,557	9,179	2,408	5,672	3,299
Total	61	1,096	38,325	16,769	1,746	7,856	3,316

There are total of 1,217 accidents over the 4 year period on the 20 miles of HC road segments and 232 accidents on the 13 miles of LC road segments. The crashes per mile of the HC group is approximately three times that of the LC group.

When divided by manner of collision, rear end type crashes account for more than fifty percent for both HC and LC groups, followed by angle type with approximately a quarter of the total accidents. It is observed that the percentages of head on and opposite direction sideswipe accidents are higher on the LC than the HC. One potential reason is that the LC roads generally do not feature TWLTL or raised medians. The same direction sideswipe for HC roads likely have a higher percentage as the HC roads generally have higher traffic volume, allowing for more opportunities for this type of incident. Compared with the state roads (urban and rural) during the same time span, the non motor vehicle collisions for the urban streets are much lower. This implies that the surrogate measures related to traffic congestion might be the better explanatory variables for the urban street data set in this study. The statistics are summarized in Table 20.

Table 20: Crash Distribution by Manner of Collision

Road Class	Crash Counts	Angle	Head on	Rear End	Sideswipe-Same Dir	Sideswipe-Opposite Dir	Not a Collision with Mother Vehicle
HC	1,217	26%	2%	52%	12%	2%	6%
LC	232	26%	4%	54%	4%	3%	9%
Total	1,449	26%	2%	53%	10%	2%	6%
State of Georgia	1,004,675*	27%	2%	36%	9%	3%	24%

*Crash counts from 2002-2005 without traffic signal in the vicinity of collision

7.1.5 Discussions

Figure 76 illustrates boxplots of various variables. With the number of incidents per mile (CPL) variable, many of the observations are well separated by HC/LC classification system. The CPL distribution of the LC group is much lower than that of the HC group in general. The median of the LC group (10 crashes/mile) is less than one-fifth of the HC group (57 crashes/mile). The two outliers of the LC group have much higher crashes/mile values than the rest of the LC group. The CPL values for these two corridors reside between the mean and the 75th percentile value in the HC group, and would therefore not have been detected as high crash locations if the HC and LC corridors were combined. This functional classification alone reduces the variation in the crash data by 28 percent.

The traffic volume of the HC group has a higher median than that of the LC group, by approximately 2.5 times. This indicates that the crash frequency is not proportional to the traffic volume. Regarding the corridor length, the HC group has a higher variability of the length than that of the LC group.

In general, the variation of speed profile ($SD85$) of the HC group is lower than the LC group. This is reasonable as the road geometries on the HC group tend to have a higher design standard. The mean of 85th percentile speed of the HC group is only slightly greater than that of the LC group.

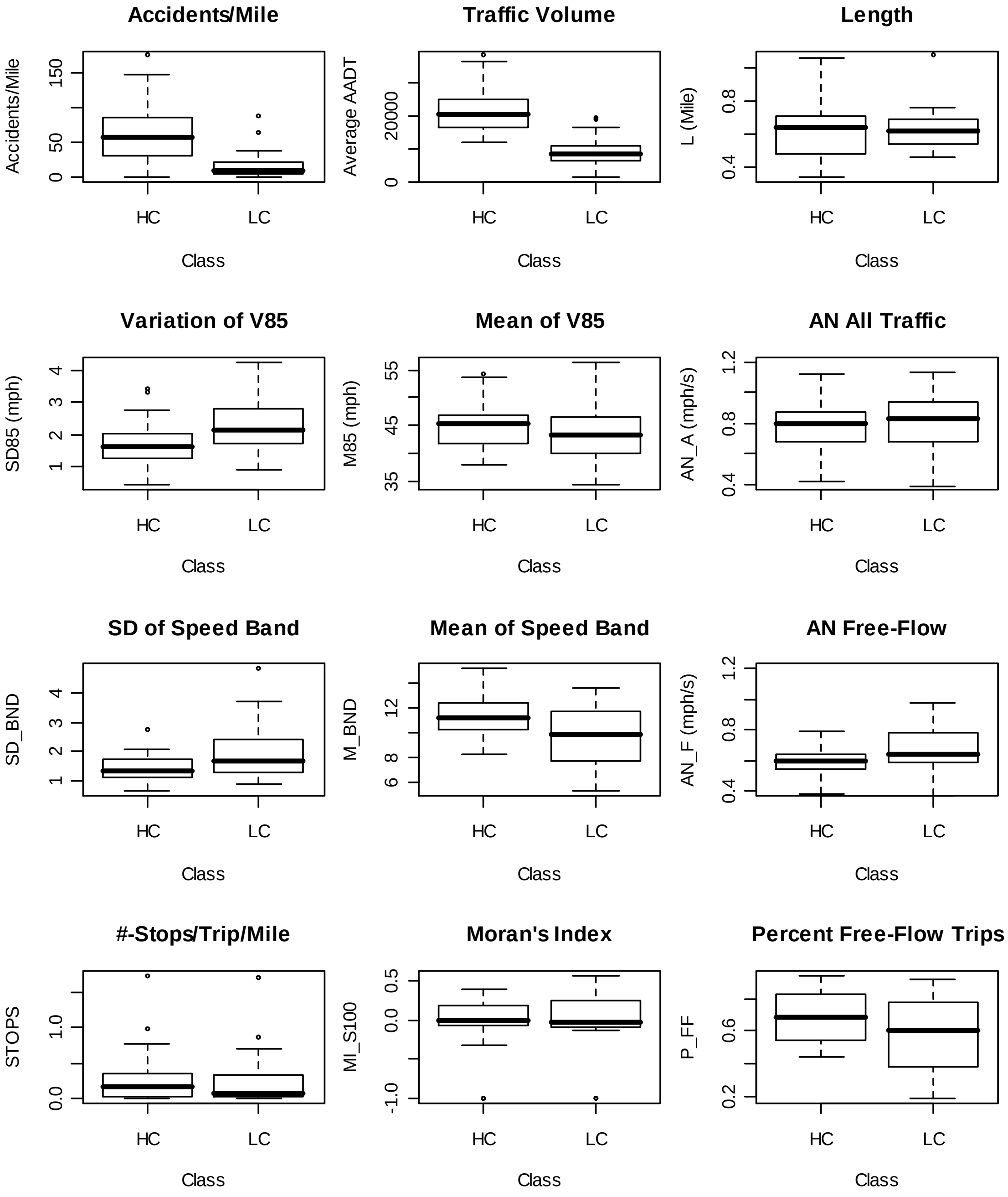
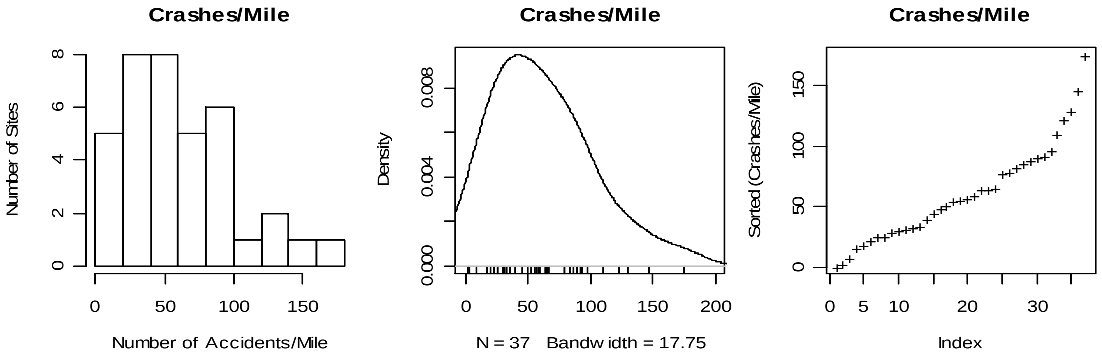


Figure 76: Box Plots of Roadway Parameters for Higher and Lower Road Classification Corridors

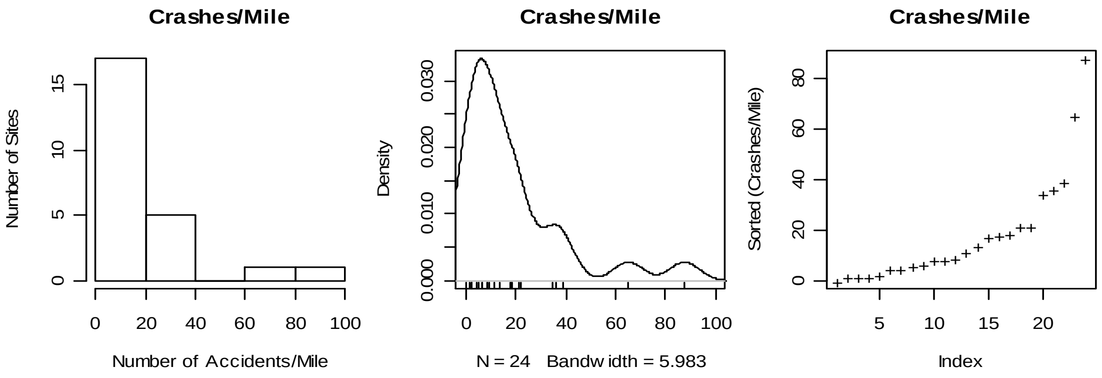
7.2 DISTRIBUTION OF CRASH FREQUENCY

Various distribution plots for the HC and LC groups are illustrated in Figure 77 and Figure 78, respectively. It is seen that crash frequency is not normally distributed. Based on the shape of the crash frequency distribution, Poisson and negative binomial distributions were used. A number of studies have shown that crash data fitted well with the Poisson and negative binomial distributions (*17, 30, 32, 57-60*).



(a) **(b)** **(c)**

Figure 77: Distribution of Number of Crashes per Mile for Higher Classification Corridors (a) Histogram, (b) Kernel Density Estimate, and (c) Index Plot of the Sorted Values.



(a) **(b)** **(c)**

Figure 78: Distribution of Number of Crashes per Mile for Lower Classification Corridors (a) Histogram, (b) Kernel Density Estimate, and (c) Index Plot of the Sorted Values

7.3 CRASH FREQUENCY AND AADT

Many crash prediction models have been developed using AADT as a predictor variable to represent traffic exposure of the roadways (*17*, *18*, *32*, *34*, *35*, *37*, *48*, *58*, *60*, *61*). This section investigates the relationship between the crash frequency per mile and the traffic volume on different road classifications. In Figure 79, the scatter plot of the crash frequency per mile vs. the AADT and the regression fit shows that the two variables have a positive relationship. The R-square of 0.29 indicates that traffic volume alone explains 29 percent of the total variation of the crash data. It can be seen that in the lower traffic volume range, the data points are more clustered around the fitted line and becomes more varied when the traffic volume increases.

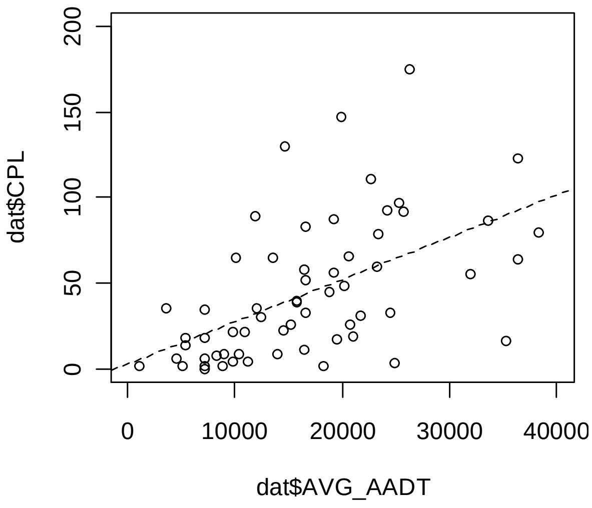
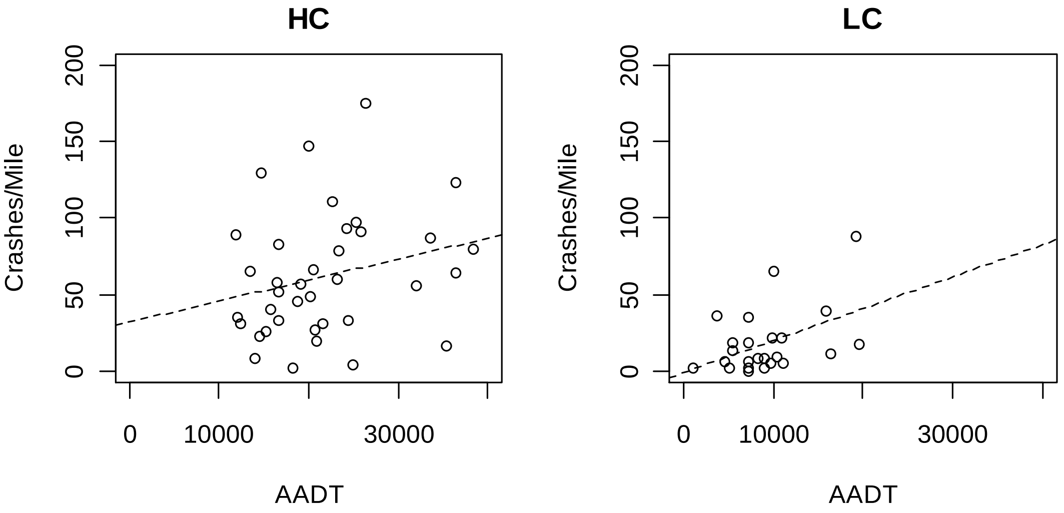


Figure 79: Scatter Plot of Crashes per Mile vs. AADT with Regression Fit (R^2 = 0.292) of All Classifications

Previous studies (*30, 32, 34, 35*) reported that crash frequency and AADT have a non linear relationship, that is, the ratio of crash frequency to the AADT is not constant throughout the AADT range. This implies that the relationship of crash and traffic volume might not be the same for different road classifications. Regression lines were fitted for the HC and LC roads separately in Figure 79. For the HC group, the fit shows a low R-square and the AADT is not a significant variable for predicting crash frequency. For the LC group, the fit shows that the AADT is a significant variable in the model, explaining 20 percent of the variation in the model.



(a) R^2 = 0.061

(b) R^2 = 0.203

Figure 80: Scatter Plot of Number of Crashes per Mile vs. AADT with Regression Fit of (a) Higher Road Classification and (b) Lower Road Classification

These results provide insight regarding the impact of traffic volume to the crash trend on different road classifications. That is, the traffic volume has a higher impact on the LC group and has little impact on the HC group. Other traffic attributes of the HC roads might have a higher influence on the accidents than merely the traffic volume.

7.4 ANALYSIS OF INFLUENTIAL FACTORS

The first step of statistical analysis involved the use of statistical tools to understand the predictor variables that might have relationships with the response variable, i.e., crash frequency per unit length (CPL). Note that only speed-profile variables are considered in this study to keep the crash prediction model practical for a road network screening process. Including road geometries and roadside features might have improved the prediction power but would require demanding data collection efforts and hence makes the model costly for an initial screening tool.

The regression tree technique is used to explore the importance of each variable. The HC and LC groups are analyzed separately as shown in Figure 81. All the potential explanatory variables were first supplied to both the HC and LC models.

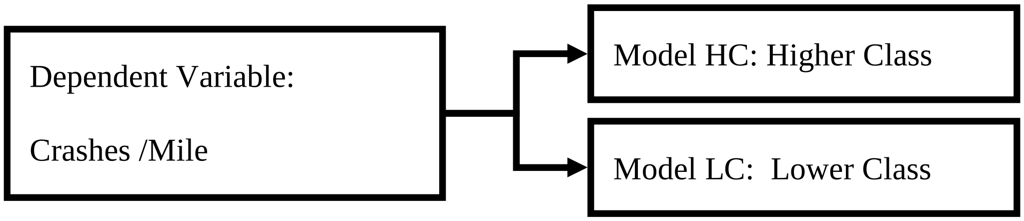


Figure 81: Final Model Development Structure

7.4.1 Results

Model HC: Crashes per Mile for Higher Classification

The results for the Model HC are shown in Figure 82. The corridors are split into six groups using four variables including M_BND, SD_BND, $STOPS$, and $SVLIM$. As seen in Figure 83, the variable $STOPS$ is used as the first split and it reduces the variation by 19 percent. The left node with STOPS < 0.36 has the corresponding mean response of 53 accidents per mile over a four year period. The right node with STOPS > 0.36 has the corresponding mean response of 95 accidents per mile over a four year period. The R-squares of further splits are much smaller. The cross validation error in Figure 84 suggests that the error increases as more splits are added.

```
Regression tree:
rpart(formula = CPL ~ P_THRU + P_FF + STOPS + MI_S100 + SD85 +
    SVLIM + M85 + SD_BND + M_BND + ANT_A + ANT_F + IQ85 + CV85 +
    FC + AVG_AADT, data = hc, method = "anova", control =
rpart.control(minsplit = 9,
    minbucket = 3, cp = 0.01, maxcompete = 4, maxsurrogate = 5,
    usesurrogate = 2, xval = 10, surrogatestyle = 0, maxdepth = 30))

Variables actually used in tree construction:
[1] M_BND  SD_BND STOPS  SVLIM

Root node error: 59375/37 = 1604.7

n= 37

        CP nsplit rel error xerror    xstd
1 0.189199      0   1.00000 1.0309 0.25367
2 0.076698      1   0.81080 1.3403 0.28601
```

```
3 0.056138      3   0.65740 1.4006 0.31329
4 0.001000      5   0.54513 1.3870 0.32625
> dat_rpart
n= 37

node), split, n, deviance, yval
      * denotes terminal node

 1) root 37 59374.6200 61.74656
   2) STOPS< 0.3567338 29 26588.9200 52.59479
     4) M_BND< 13.89791 24 17087.9200 46.91634
       8) SD_BND>=1.473479 8  1158.8040 27.29626 *
       9) SD_BND< 1.473479 16 11309.7500 56.72638
        18) SVLIM< 9.146291 11  7930.2020 48.12635
          36) STOPS< 0.1608791 5  1351.7350 27.07342 *
          37) STOPS>=0.1608791 6  2515.5620 65.67046 *
        19) SVLIM>=9.146291 5   776.1402 75.64643 *
     5) M_BND>=13.89791 5  5012.5150 79.85138 *
   3) STOPS>=0.3567338 8 21552.0700 94.92173 *
```

Figure 82: Regression Tree Results for the Model HC

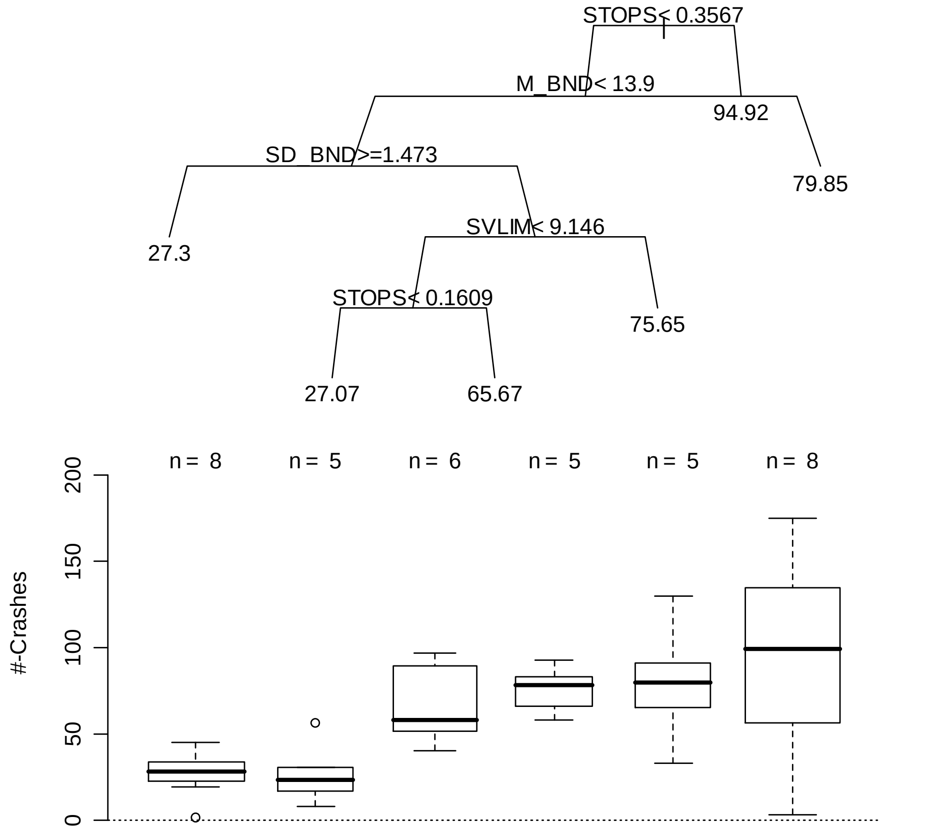


Figure 83: Regression Tree Diagram and its Corresponding Box Plot for Model HC

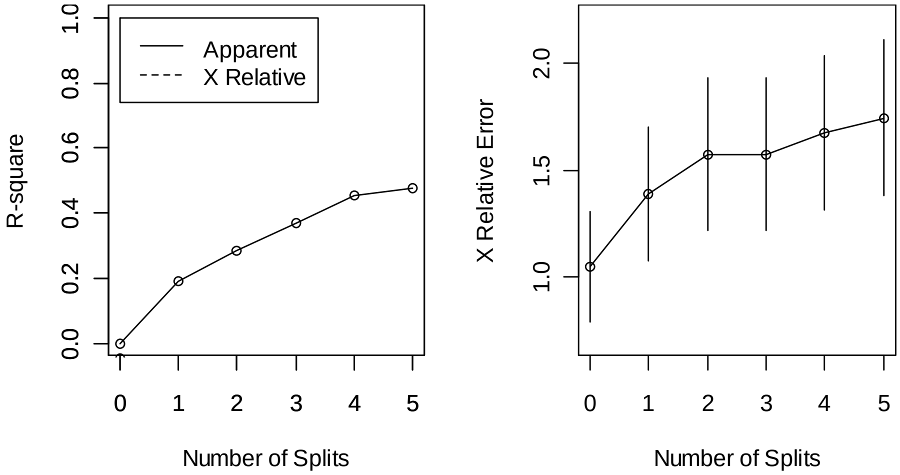


Figure 84: Plots of the R-Squared (Left) and the Relative Error from Cross-Validation (Right) for the Different Splits for Model HC

Model LC: Crashes per Mile for Lower Classification

The results for the Model LC are shown in Figure 85. The corridors are split into five groups using four variables including AN_A, P_FF, P_THRU, and $STOPS$. As seen in Figure 86, the variable $STOPS$ is used as the first split and it reduces the variation by 59 percent. The left node with STOPS < 0.59 has the corresponding mean response of 12 accidents per mile over four years. The right node with STOPS > 0.59 has the corresponding mean response of 63 accidents per mile over four years. Note that there are only three observations on the right node. The R-squares of further splits are very small. The cross validation error in Figure 84 suggests that only the first split reduces the cross validation error and the error will increase as number of splits increases.

```
Regression tree:
rpart(formula = CPL ~ P_THRU + P_FF + STOPS + MI_S100 + SD85 +
    SVLIM + M85 + SD_BND + M_BND + ANT_A + ANT_F + IQ85 + CV85 +
    CRAWL_P + CRAWL_I + BLK_P_01 + FC + AVG_AADT, data = lc,
    method = "anova", control = rpart.control(minsplit = 9, minbucket
= 3,
        cp = 0.001, maxcompete = 4, maxsurrogate = 5, usesurrogate =
2,
        xval = 10, surrogatestyle = 0, maxdepth = 30))

Variables actually used in tree construction:
[1] ANT_A  P_FF   P_THRU STOPS

Root node error: 10471/24 = 436.28

n= 24

        CP nsplit rel error xerror    xstd
1 0.638042      0   1.00000 1.1301 0.52269
```

```
2 0.069542      1   0.36196 1.0955 0.54058
3 0.014496      3   0.22287 0.9335 0.34044
4 0.001000      4   0.20838 0.9620 0.33745
> dat_rpart
n= 24 

node), split, n, deviance, yval
      * denotes terminal node

 1) root 24 10470.6000 18.397380  
   2) STOPS< 0.5923704 21  2370.9720 12.091350  
     4) ANT_A< 0.8457284 12   500.1113  7.687239  
       8) P_THRU>=0.8615558 7    75.8281  4.681509 *
       9) P_THRU< 0.8615558 5   272.5050 11.895260 *
     5) ANT_A>=0.8457284 9  1327.7660 17.963510  
      10) P_FF< 0.6067029 6   241.9371 10.840800 *
      11) P_FF>=0.6067029 3   172.6345 32.208930 *
   3) STOPS>=0.5923704 3  1418.9460 62.539580 *
```

Figure 85: Regression Tree Results for the Model LC

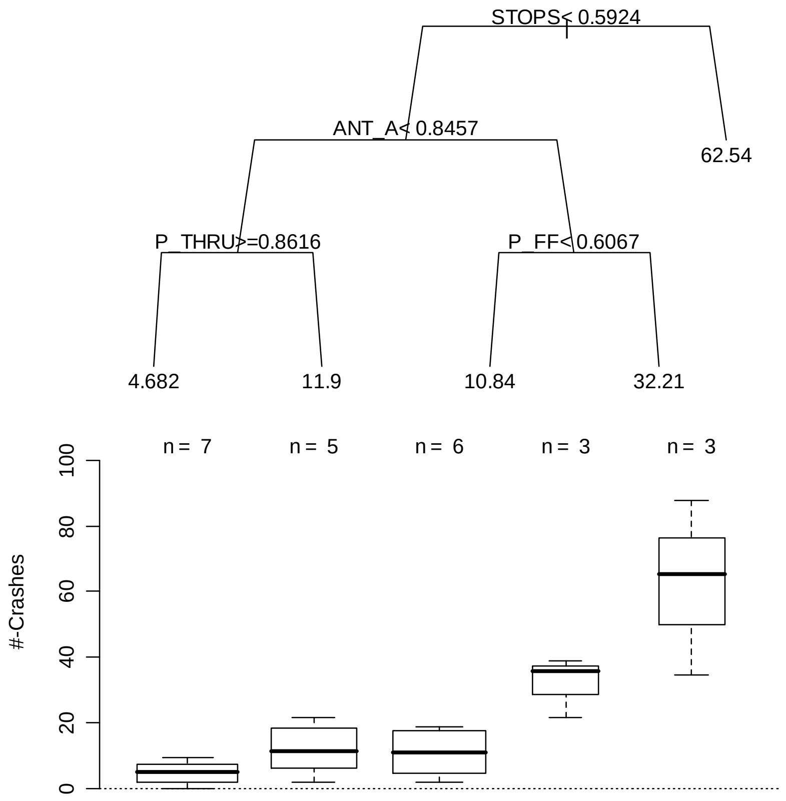


Figure 86: Regression Tree Diagram and its Corresponding Box Plot for Model LC

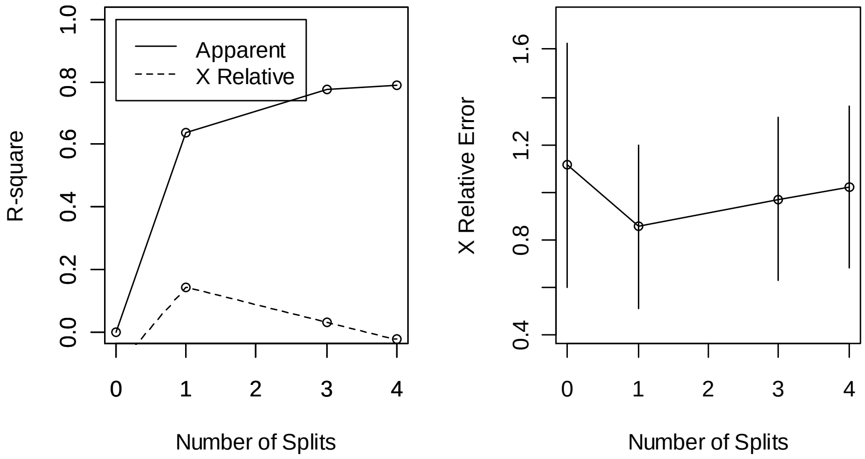


Figure 87: Plots of the R-Squared (Left) and the Relative Error from Cross-Validation (Right) for the Different Splits for Model LC

7.4.2 Discussions

Model HC: Crash Frequency per Mile for Higher Classification

The cross validation error in Figure 84 estimates how this tree model will perform in practice. The figure shows that the model might not perform well with other datasets. Since STOPS is the most influential variable in the model, the relationship between STOPS and CPL is illustrated by the scatter plot in Figure 88. The STOPS variable does not appear to be a good classifier between the low and high crash road groups. The splitting line is very tight with five observations almost on the line. Corridors 35, 33, and 21 are on the right side while the corridors 17 and 32 are on the left side of the line.

Despite the low R-square and high cross validation error, the model results can be used to investigate the relationship of the response to each variable. The direction of relationships between the response (CPL) and the predictor variables in the tree model were generally as expected with the exception of SD_BND. For instance, the crash frequency increases as the stop frequency increases. The regression tree shows that when the $STOPS$ is greater than 0.36 stops/mile/trip, the number of accidents increases two-fold. For the variable M_BND, the number of accidents also increases as the speed band widens. The high variation from the speed limit along the roadway ($SVLIM$) associates with high crash frequency. The direction of variation of speed band (SD_BND) to the response is counter-intuitive, i.e., the crash frequency decreases as the variation of speed band increases.

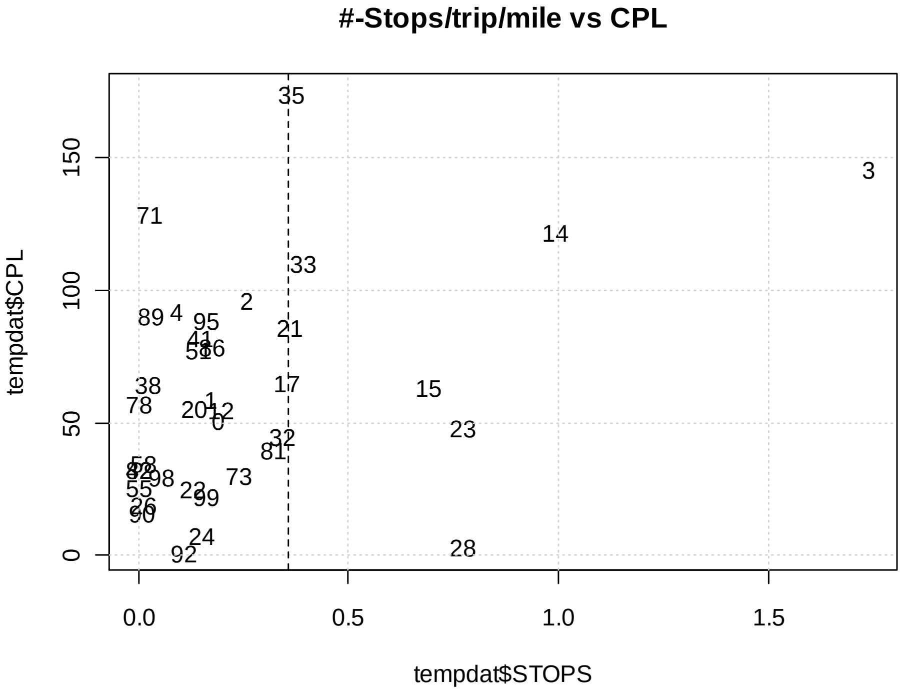


Figure 88: Scatter Plot between Stop Frequency and Number of Crashes per Mile. The Vertical line shows the data separation at STOPS=0.36 Stops/Trip/Mile.

Model LC: Crash Frequency per Mile for Lower Classification

The LC model indicates the $STOPS$ variable being the most influential variable. The cross validation error in Figure 87 shows that the error increases for a number of splits greater than one. This means that only the $STOPS$ variable has a predictive power in practice. The scatter plot of $STOPS$ versus CPL is shown in Figure 89. It is seen that the splitting line at $STOPS$ = 0.59 separates the high crash and low crash corridors quite well. The three corridors with stop frequency per trip per mile greater than 0.59 have relatively high crashes while the others have relatively low crashes. This is in line with expectation as a corridor with high number of stops indicates likely higher traffic congestion. Note that high traffic congestion does not necessarily imply high traffic volume. For example, corridor number 40 has twice the traffic volume as corridor number 30 but half the number of stops. This means reducing number of stops on the road may also reduce number of accidents on the road, depending on the method used.

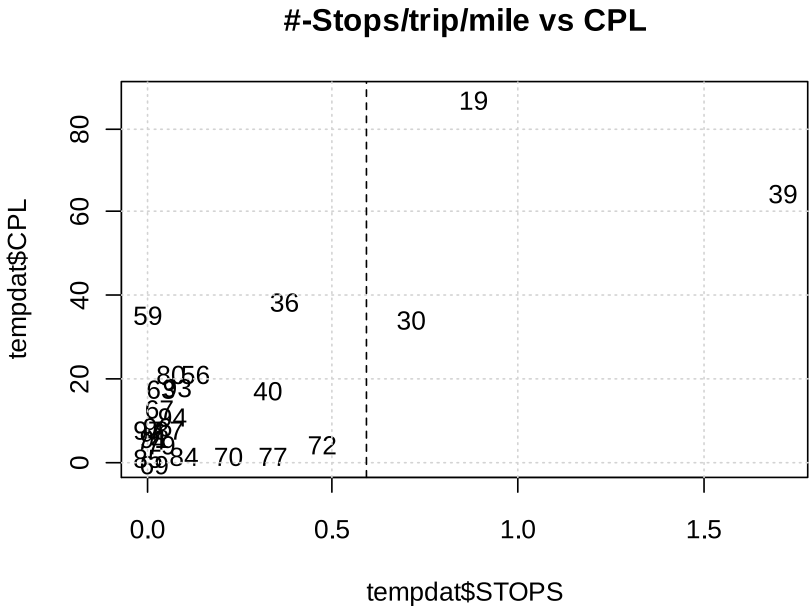


Figure 89: Scatter Plot between Stop Frequency and Number of Crashes per Mile for the LC Group. The Vertical line shows the data separation at STOPS=0.59 Stops/Trip/Mile.

Regarding the direction of the relationship to the response, two variables are in line with expectation. For instance, high acceleration noise associates with a higher number of accidents. Also, a high percentage of through traffic (or low percentage of turning movement to/from the driveways) associates with lower crash frequency. The last split is counter-intuitive. It indicates that high percentage of free-flow traffic associates with high crash frequency.

7.5 MODEL DEVELOPMENT

Regression tree and the linear regression techniques are deployed in the model development effort. The regression tree results from the previous section are compared with the results from the linear models constructed in this section. The model with a better R-square is selected.

Using multiple linear regression might not be appropriate with the crash data because the response can take only positive integer values and the crash counts are unlikely to have a normal distribution. A Generalized Linear Model (GLM) approach was applied with a log link function as described in Section 2.5.3. This approach ensures that the fitted values are positive and does not require the dependent variable to be normally distributed (*62*). The followings are model development for the HC and LC corridor groups.

Model HC: Crash Frequency for Higher Classification Roadways

HC: Tree Model with Stop Frequency Variable

The tree model with $STOPS$ as the most important explanatory variable is used. The result is shown in Figure 90. The mean responses (accidents in 4 years per mile) are 53 for the left node and 95 for the right node. The R-square for this model is 0.19. The further splits are not shown because the tree model could not be significantly improved by adding more variables.

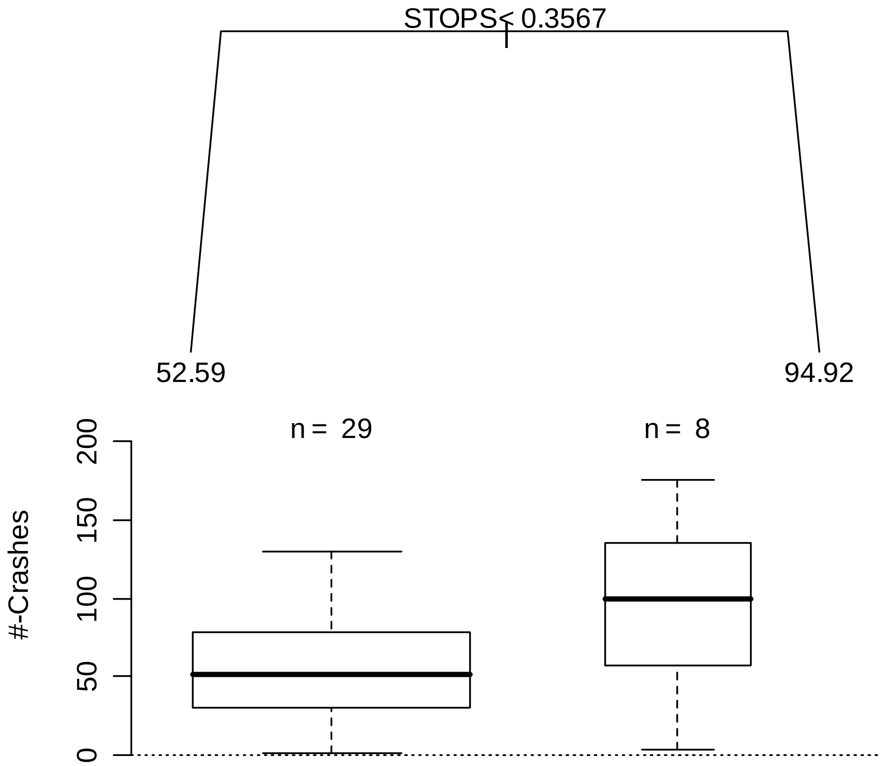


Figure 90: Regression Tree Diagram for Model HC with R-square of 0.19

HC: Generalized Linear Model

In this section, we try to fit the GLM model. Most of the crash prediction models were constructed using the Poisson and Negative Binomial error distribution (*18*). The following model form is used:

$$CPL = e^{(\beta_0+\beta_1 X_1+\beta_2 X_{2.\dots})}$$

Since the Poisson distribution require the random variable to be discrete, the variable corridor length (L) is moved to right hand side with its power assumed to be one. That is, number of accidents increases proportionally to the length of the corridor. To make parameter estimation simple, the log transformation is used and the model form becomes

$$\log(ACC/L) = \beta_0 + \beta_1 X_1 + \beta_2 X_{2.} \dots, \text{ or}$$

$$\log(ACC) = \log(L) + \beta_0 + \beta_1 X_1 + \beta_2 X_{2.} \dots$$

The log form of length with a fixed coefficient of one is used because it is expected that the corridor length has a proportional effect on the crash frequency.

HC: Poisson vs. Negative Binomial Distribution

Since we do not know whether the Poisson or negative binomial error structures should be used, we first estimate the model parameters $(\beta_0, \beta_1, \beta_2, \dots)$ with the Poisson error structure and calculate the deviance. The deviance (D) for the Poisson regression can be calculated as (*17, 38, 39*):

$$D = 2\sum_{i=1}^{n}\left(y_i \log(y_i/\hat{\mu}_i) - (y_i - \hat{\mu}_i)\right)$$

where y_i is the response value of observation i and μ_i is the fitted value for the corresponding y_i. If the model deviance is significantly greater than its corresponding degrees of freedom (n-p), it indicates that the data have greater dispersion than could be captured by the Poisson distribution and the negative binomial distribution is suggested.

The Poisson regression model is fitted to the HC dataset and the results are shown in Figure 91.

```
Call:
glm(formula = ACC ~ STOPS + offset(I(log(LEN))), family = poisson,
    data = hc)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-7.4778  -3.1957  -0.3224   2.0051  10.9754

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.95709    0.03630 109.017   <2e-16 ***
STOPS        0.61062    0.07382   8.272   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 546.95  on 36  degrees of freedom
Residual deviance: 489.24  on 35  degrees of freedom
AIC: 677.98

Number of Fisher Scoring iterations: 5
```

Figure 91: Results of the Poisson HC Model with the Predictor Variable STOPS

The fitted GLM model shows that the estimated parameter for $STOPS$ is statistically significant beyond the 0.01 level of significance, even though we cannot observe an obvious trend between the crash frequency and stop frequency in the scatter plot of Figure 88.

Also included in the results are the null deviance and residual deviance. The null deviance is the deviance for a model with no predictor variable and the deviation is calculated merely from the intercept term. The residual deviance measures how much the data deviates from the current model.

It can be seen that the residual deviance (489) is much larger than the corresponding degrees of freedom (35). This is likely to be the symptom of overdispersion, i.e., the residual deviance is much larger than the degrees of freedom. The Pearson chi-square test can be used to statistically check the goodness of fit of the model. The null hypothesis is that the Poisson regression model has a good fit to the data. The residual deviance of 489, to be compared with a chi-square distribution with 35 degrees of freedom. The p-value of less than 0.001 rejects that the Poisson regression model fits well.

The property of Poisson distribution that mean equals the variance is too restrictive for the empirical variance in this data. Therefore, the negative binomial should be considered to remedy the overdispersion phenomenon of the crash data.

HC: Negative Binomial Regression with Stop Frequency

The Negative Binomial family is used to fit the crash prediction model with STOPS as a predictor. The detailed results are not reported since the $STOPS$ variable is not significant at 0.05 level of confidence. However, the residual deviance (41) for the negative binomial model is closer to its corresponding degrees of freedom (35) indicating a better fit than using the Poisson model.

HC: Negative Binomial Model with Acceleration Noise, All Traffic Condition

To select other important predictor variables, scatter plots of the crash frequency per mile and the predictor variables were investigated. Out of the 15 predictor variables, only acceleration noise under all traffic condition, (AN_A) seems to have a relationship with the response (Figure 92). The GLM approach is used to construct the model shown in Figure 93.The AN_A is significant at 0.05 confidence level with the R-square of 1-(19,392/28,992) = 33 percent, which is greater than the R-square from the regression tree with the $STOPS$ variable. The residual deviance for this model is 40.8 to be compared with the chi-square distribution with 35 degrees of freedom. The p-value for the chi-square test is 0.23, therefore we do not have to reject the null hypothesis that the negative binomial model has a good fit with the data.

The mathematical equation for this negative binomial regression model is as follows:

$$\hat{y}_i = \mu_i = L^1 * e^{(2.626+1.862(AN\_A_i))} \text{ ,or}$$

$$\widehat{CPL}_{HC} = {}^{\hat{y}_i}/_{L} = e^{(2.626+1.862(AN\_A_i))}$$

where $\hat{y}_i$ is expected number of accidents in four years for the road section i with length L (mi); $\widehat{CPL}_{HC}$ is the expected number of accidents per mile over four years for the same road section; and AN_A is the acceleration noise under all traffic condition (mph/sec).

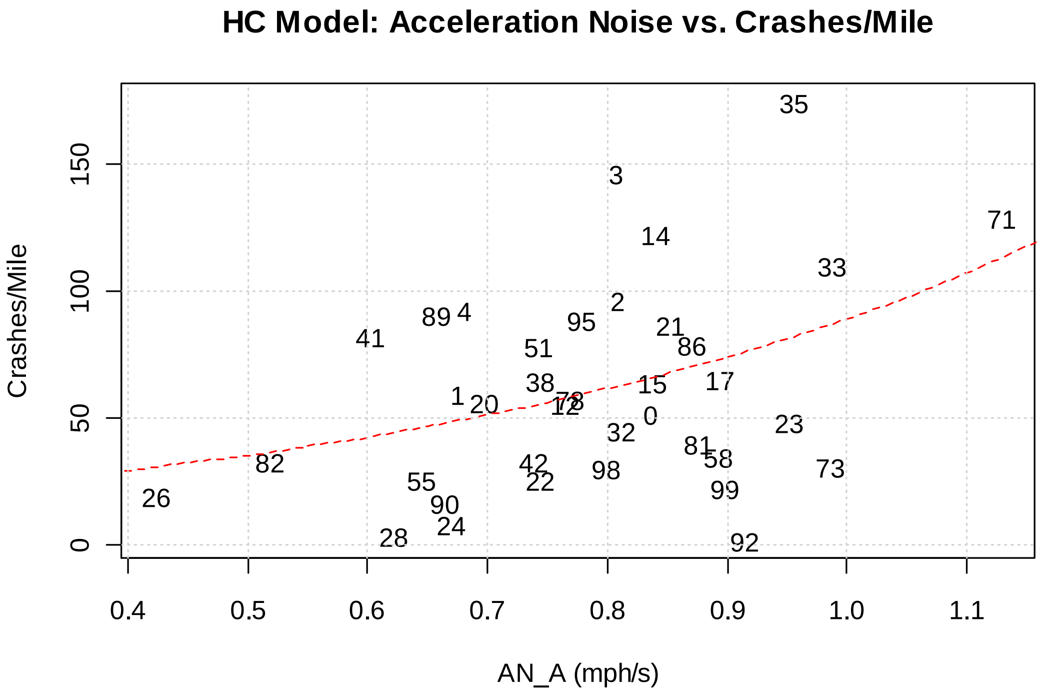


Figure 92: Scatter Plot between Acceleration Noise under All Traffic Condition and Number of Crashes per Mile for HC dataset. The dotted line represents the fitted model with R-square of 0.33.

```
Call:
glm.nb(formula = ACC ~ ANT_A + offset(I(log(LEN))), data = tempdat,
    init.theta = 2.45167823371255, link = log)

Deviance Residuals:
     Min         1Q     Median         3Q        Max
-3.22218  -0.77174  -0.04393   0.40073   1.50348

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)    2.626      0.638   4.116 3.86e-05 ***
ANT_A          1.862      0.796   2.339   0.0193 *
---
```

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


(Dispersion parameter for Negative Binomial(2.4517) family taken to
be 1)


    Null deviance: 46.013  on 36  degrees of freedom
Residual deviance: 40.800  on 35  degrees of freedom
AIC: 322.85


Number of Fisher Scoring iterations: 1



              Theta:  2.452
          Std. Err.:  0.623


 2 x log-likelihood:  -316.853
```

Figure 93: Results of the Negative Binomial HC Model with the AN_AF Variable with R-square of 0.33

HC: Negative Binomial Model with Acceleration Noise and Multiplicative Form of Length

The assumption of proportional relationship between segment length and crash frequency is tested by relaxing the fixed coefficient of the length variable. When the coefficient of the log of corridor length is not fixed to 1.0, the relationship becomes multiplicative form and the results are shown in Figure 94.

In this model, the power of the corridor length is less than one (0.706). This means that if there are N accidents on 1 mile section, it is expected to have 1.6*N accidents on the 2 mile section with the same traits. The proportion of deviance explained by this model, 1-20,076/28,992= 31 percent, indicates a similar fit to the previous model.

To test the goodness of fit of this model, the Pearson chi-square test is used. The residual deviance for this model is 40.819, to be compared with the chi-square distribution with 34 degrees of freedom. The p-value of 0.20 indicates that the model appears to be adequate.

The mathematical equation for this negative binomial regression model is as follows:

$$\hat{y}_i = \mu_i = L^{.706} * e^{(2.2923+2.0295(AN\_A_i))}$$

where $\hat{y}_i$ is expected number of accidents in four years for the road section i with length L (mi); and AN_A_i is the acceleration noise under all traffic condition (mph/sec).

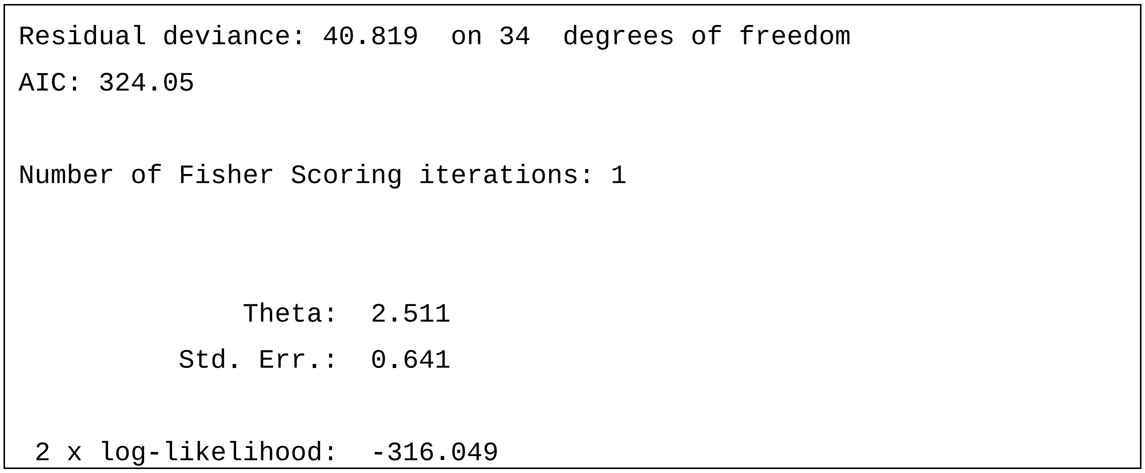
```
Call:
glm.nb(formula = ACC ~ ANT_A + (I(log(LEN))), data = tempdat,
    init.theta = 2.51080164865263, link = log)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-3.23781  -0.76959  -0.06308   0.46183   1.69620

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   2.2923     0.7299   3.141  0.00169 **
ANT_A         2.0295     0.8079   2.512  0.01200 *
I(log(LEN))   0.7060     0.3349   2.108  0.03506 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(2.5108) family taken to
be 1)

    Null deviance: 55.361  on 36  degrees of freedom
```



```
Residual deviance: 40.819  on 34  degrees of freedom
AIC: 324.05

Number of Fisher Scoring iterations: 1



              Theta:  2.511
          Std. Err.:  0.641

 2 x log-likelihood:  -316.049
```

Figure 94: Results of the Negative Binomial HC Model with the Predictor Variables STOPS and LEN with the R-square of 0.31

Model LC: Crash Frequency for Lower Classification Roadways

LC: Tree Model with Stop Frequency Variable

From the analysis in Section 7.4, the measure $STOPS$ is the most important variable for the lower functional classification. The splitting point is at $STOPS$ = 0.59 stops per trip per mile. The mean responses are 12 and 63 accidents in 4 years per mile for the left and right nodes, respectively. There are 21 corridors on the low crash frequency group and only 3 corridors on the high crash frequency group. The R-square of the first split is 0.64.

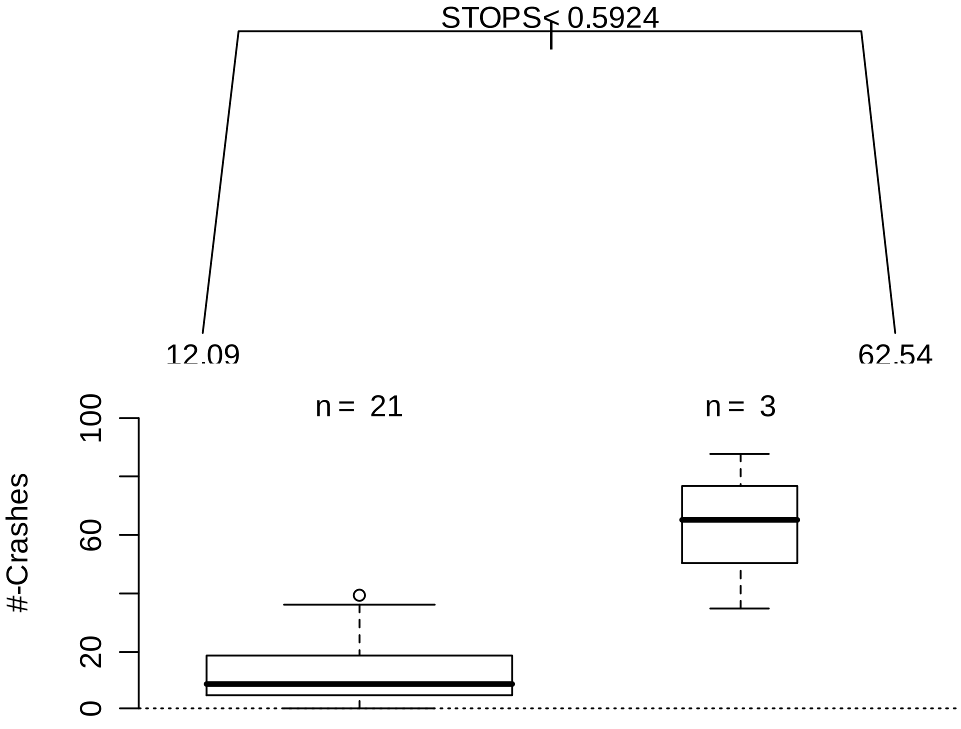


Figure 95: Regression Tree Diagram for Model LC with R-square of 0.64

LC: Negative Binomial Model with Stop Frequency Variable

Next, we check if the GLM model could have a better fit than the regression tree model. The two variables that are significant in the negative binomial model include $STOPS$ and AN_A. The $STOPS$ model reports the R-square of 1-2346/2927= 20 percent and its scatter plot is shown in Figure 96. The residual deviance of 26.537 is compared with the chi-square distribution with 22 degrees of freedom. The p-value of 0.23 is large enough to conclude that the negative binomial model has a good fit.

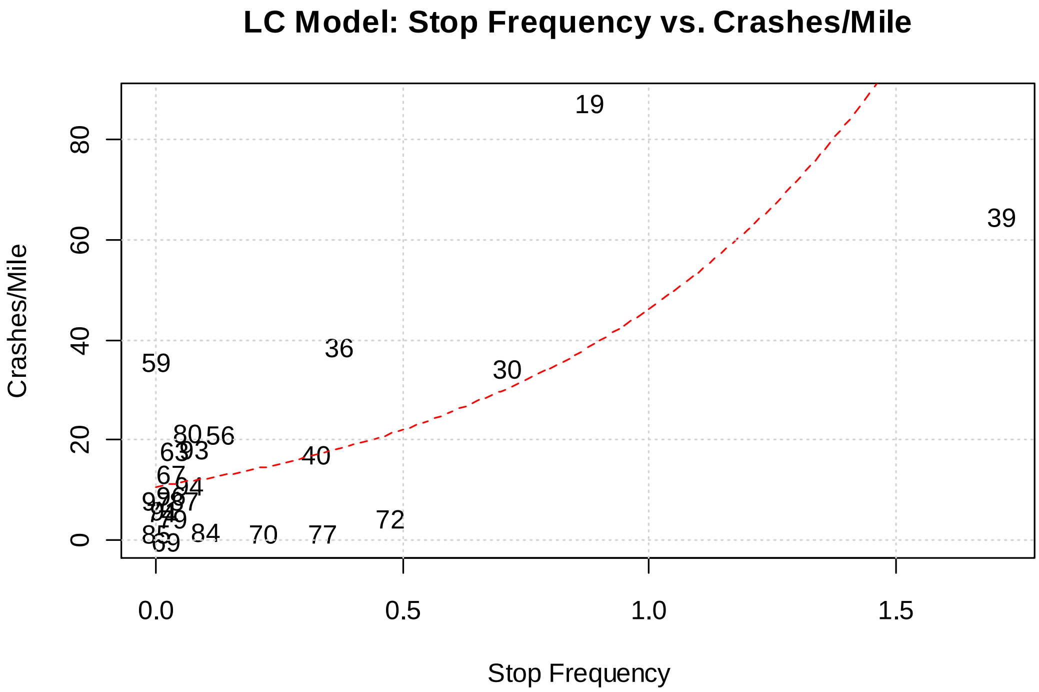


Figure 96: Scatter Plot between Stop Frequency and Number of Crashes per Mile for LC Dataset. The dotted line represents the fitted model (R-square = 0.20).

LC: Negative Binomial Model with Acceleration Noise, All Traffic Condition

Furthermore, the acceleration noise (AN_A) model indicates the R-square of 1-1755/2927=0.4 and its scatter plot is shown in Figure 97. The residual deviance of 25.8 is to compared with the chi-square distribution with 22 degrees of freedom. The p-value of 0.26 is large enough to conclude that the negative binomial model has a good fit.

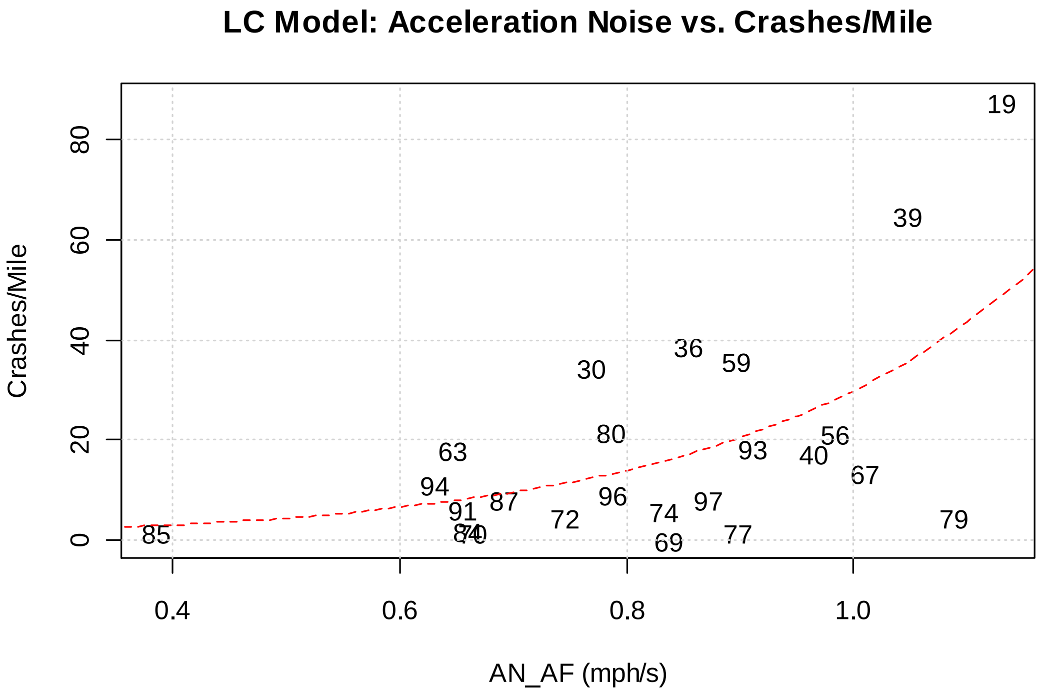


Figure 97: Scatter Plot between Acceleration Noise under All Traffic Condition and Number of Crashes per Mile for LC Dataset. The dotted line represents the fitted model (R-square= 0.40).

Since the regression tree model is superior to the two GLM models, we classify the LC dataset using the STOPS criterion. The mean response of 63 accidents per mile is used as a predicted value for the right node since there are only three observations. The left node has 21 observations and we might be able to improve the prediction power by adding splitting nodes or constructing a GLM model. According to the regression tree results in Figure 85, the next split would improve the R-square by 7 percent, which is not a sufficient improvement to justify the additional model complexity.

LC [STOPS<0.59]: Negative Binomial Model with Acceleration Noise, All Traffic Condition

For the STOPS< 0.59 dataset, the measure AN_AF is used to construct the negative binomial regression model. The parameter estimate is barely significant with the p-value of 0.04 and the R-square for this model is 1-620/684=0.09. The Pearson chi-square test shows a p-value of 0.25 indicating that the negative binomial model has a good fit.

The mathematical equation for this negative binomial regression model is as follows:

$$\hat{y}_i = \mu_i = L^1 * e^{(0.4677+2.4581(AN\_A_i))} \text{ ,or}$$

$$\widehat{CPL}_{LC1} = {}^{\hat{y}_i}/_{L} = e^{(0.4677+2.4581(AN\_A_i))}$$

where $\hat{y}_i$ is expected number of accidents in four years for the road section i with length L (mi), $\widehat{CPL}$ is the expected number of accidents in four years per mile for the same road section, and AN_A is the acceleration noise under daylight and dry conditions (mph/sec). Note that the coefficient for the corridor length is not restricted to one, the parameter estimate is not statically significant and therefore the model with multiplicative effect of corridor length is not considered in the final model.

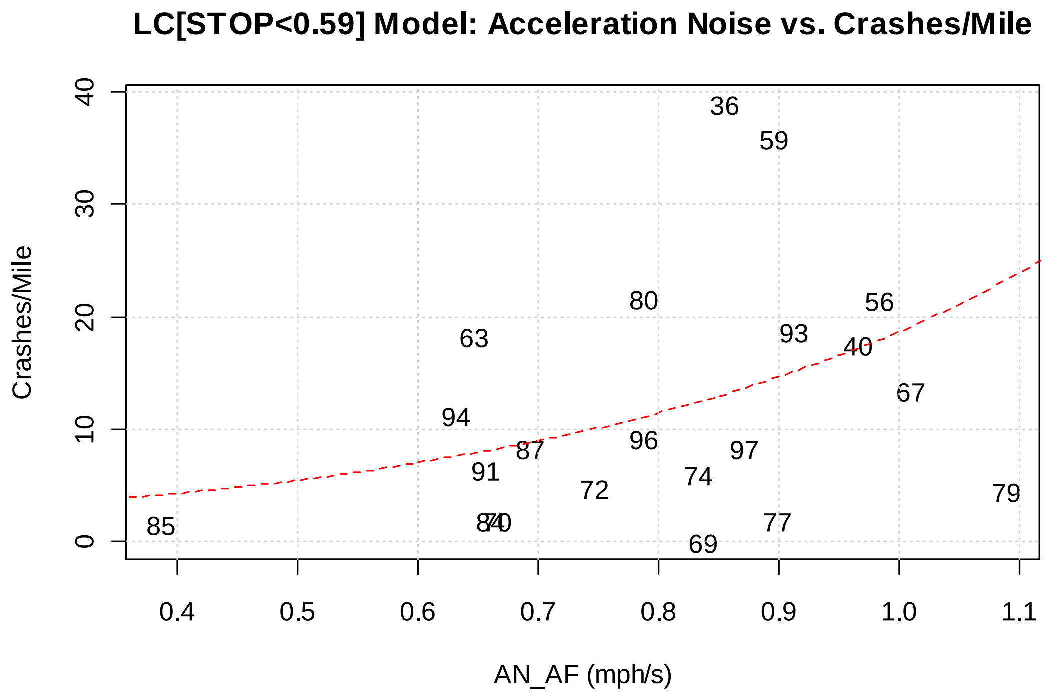


Figure 98: Scatter Plot between Acceleration Noise under All Traffic Condition and Number of Crashes per Mile for LC[STOP<0.59] Dataset. The dotted line represents the fitted model (R-square= 0.09).

```
Call:
glm.nb(formula = ACC ~ ANT_A + offset(I(log(LEN))), data = tempdat,
    init.theta = 1.94540428001504, link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2223  -0.7984  -0.2272   0.2825   1.7236

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.4677     0.9836   0.475   0.6344
ANT_A         2.4581     1.1875   2.070   0.0385 *
```

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


(Dispersion parameter for Negative Binomial(1.9454) family taken to
be 1)


    Null deviance: 26.057  on 20  degrees of freedom
Residual deviance: 22.611  on 19  degrees of freedom
AIC: 123.10


Number of Fisher Scoring iterations: 1



              Theta:  1.945
          Std. Err.:  0.816


 2 x log-likelihood:  -117.103
```

Figure 99: Results of the Negative Binomial Model with the AN_AF Variable for the LC with STOPS Less Than 0.59. The R-Square is 0.09

7.6 SUMMARY

The final crash prediction model is a combination of the regression tree and the generalized linear models. First, the proposed functional classification system is used as a splitting variable to separate lower and higher functional classifications. The higher classification is a combination of the minor arterials and high traffic volume (greater than 11,500 vehicles per day) collectors while the lower functional classification includes the lower volume (less than 11,500 vehicles per day) collectors and local streets. This classification improved the R-square by 28 percent.

For the lower functional classification, the variable $STOPS$ is used to separate high stop frequency (> 0.59 stops/trip/mile) and low stop frequency (<0.59 stops/trip/mile) segments. The R-square is improved by 64 percent by this split. For the low crash frequency corridors, the crash frequency is predicted by the GLM model (R-square = 0.09):

$$ACC_{LC1} = L^{1} * e^{(0.4677+2.4581(AN\_A_i))}$$

The high stop frequency LC segments will have an expected crash frequency of 63 accidents per mile in four years and can be formulated as:

$$ACC_{LC2} \ = \ L * 63$$

For the HC corridors, the crash frequency is predicted by the GLM models (R-squares = 0.33 and 0.31 for the L^{1} and $L^{0.7}$ models, respectively):

$$ACC_{HC} = L^{1} * e^{(2.626+1.862(AN\_A_i))},$$

or

$$ACC_{HC} = L^{.706} * e^{(2.2923+2.0295(AN\_A_i))}$$

The overall R-square of the final model is 0.48, which indicates that almost half of the variation in the crash data can be explained using functional classification, stop frequency and acceleration noise variables. The model is potentially be used as a screening tool for road safety improvement program. The final model structure is illustrated in Figure 100.

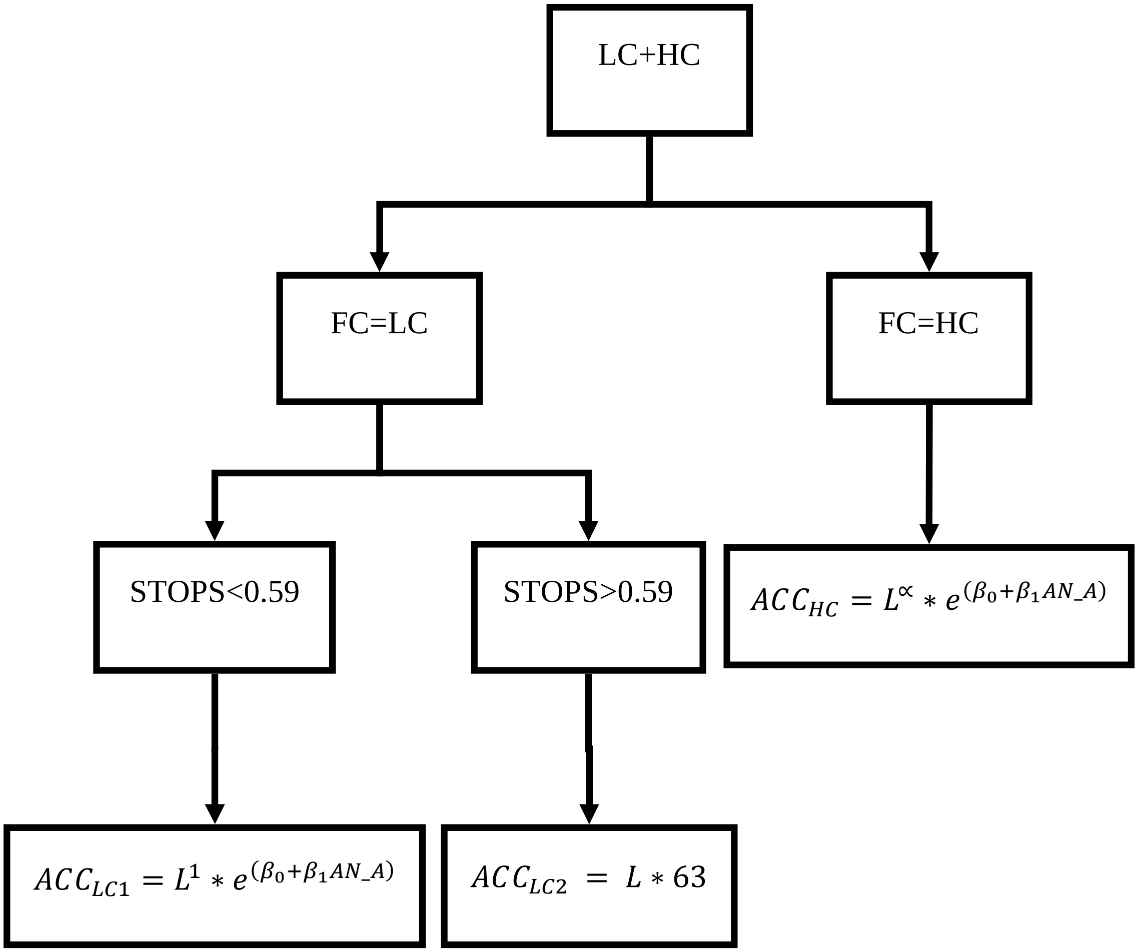

LC+HC
FC=LC
FC=HC
STOPS<0.59
STOPS>0.59
$ACC_{HC} = L^{\propto} * e^{(\beta_0+\beta_1 AN\_A)}$
$ACC_{LC1} = L^1 * e^{(\beta_0+\beta_1 AN\_A)}$
$ACC_{LC2} = L * 63$

Figure 100: Final Model Structure R-square = 0.48 (LC=Lower Functional Classification, HC=Higher Functional Classification, STOPS= Stops Frequency, AN_AF= Acceleration Noise under All Traffic Condition)

Chapter 8. CONCLUSIONS

This chapter summarized the research tasks and findings from this dissertation in Section 8.1. Major contributions are listed in 8.2. Limitations and future works are suggested in Section 8.3.

8.1 SUMMARY OF FINDINGS

This dissertation is the first study developing a crash prediction model for low speed urban streets using continuous speed data from the GPS-equipped vehicle data. The developed model is intended to be used as a screening tool in which traffic engineers can use GPS traces from vehicles in concert with basic roadway information such as road classification to identify the sections on the urban street network that might be expected to exhibit a higher than normal number of crashes.

While the previous researchers constructed similar models using point-specific speed measures, this study proposed several measures using the speed profile data. The profile-based measures are expected to reveal the variation of speed along the spatial dimension of the roadway, which might indicate safety issues.

Some of the proposed measures are designed to capture speed consistency. This requires the speed data to be under free-flow conditions. However, the original GPS-based trajectory data did not include a direct information of the flow regime. As a result, a series of speed data processing filters were developed to identify likely free-flow speed data. The sensitivity of the data filters to the speed-related measures was analyzed and the findings are as follows:

- The speed measures derived from the 85th percentile speed profile including $SD85$, $M85$, $CV85$, and $IQ85$ are not sensitive to the data filters. This is because the filters are designed to remove potentially non-free-flow trips, which usually contain low speed data points, and thereby the filter effect on the high percentile speeds is minimal. As a result, the speed measures derived from the high percentile speed data might not require the free-flow filtering process.

- The speed measures utilizing the low speed data, such as, mean and variation of the speed band and the acceleration noise are affected by the free-flow filters on many corridors. Therefore, the free-flow condition is an important factor to determine the values of these measures when they are desired for operating speed analysis.

- It is observed that the off-peak period cannot be effectively used as a surrogate of the free-flow filters because different urban streets have different peak period. Therefore, the speed profile pattern filters provide more reliable information about the traffic condition than a time-of-day filter.

In the model development effort, the study corridors were divided into two classes, namely, higher and lower functional classifications. The crash prediction model was constructed using the regression tree and generalized linear modeling approaches. Findings from the models are summarized as follows:

- The safety characteristics of roadways are likely a function of the roadway classification. Separation of corridors by functional classification allows for the identification of high crash LC corridors (relative to other LC corridors) that might not be otherwise identified.

- For the higher classification roadways, the most important explanatory variable used to construct the regression model is the acceleration noise under all traffic condition (AN_A) which includes all continuous trips made during day light and no rain conditions. The model explains 27 percent of the total variation in the crash data. The AN_A has a positive relationship with the crash frequency on the HC corridors, i.e., the higher the acceleration noise, the higher the expected crash frequency.

- For the lower classification roadways, the variables stop frequency ($STOPS$) and acceleration noise (AN_A) were used to construct the crash prediction model. Both variables have positive relationships with the crash frequency on the LC corridors, as expected.

- Several measures derived from likely free-flow speed data such as $SD85$, $M85$, SD_BND, M_BND, and AN_F do not have significant relationship with the crash frequency. This might be because accidents on the urban streets are mainly a function of a combination of traffic congestion, and the roadway design, and roadside characteristics while the mentioned variables are designed to measure the speed consistency along the corridor without considering the level of traffic congestion. The traffic condition of the roadway can be partially explained by the variables $STOPS$ and AN_A.

- While most of the crash prediction models include traffic volume as their traffic exposure measure, the model in this study does not require the comprehensive traffic volume information because this can be represented by the road functional classification, the acceleration noise, and the stop frequency variables.

8.2 CONTRIBUTIONS

Existing crash prediction models require data such as physical road geometries, traffic volume, and speed characteristics. These data can be expensive and time consuming to collect. Most of the speed-safety models also use point-specific speed measures, which do not capture speed consistency along the corridor. This study explores the use of the speed profile data from GPS instrumented vehicles. The major contributions of this research are summarized as follows:

- The research developed a methodology to obtain the speed profile data under various conditions.

- The research provides an understanding of the impact of data filtering processes on the speed measures.

- The crash prediction model provides a foundation for a speed profile based screening tool in a road safety improvement program. In particular, traffic engineers can use this model to identify potential problem corridors using only the speed data collected from the instrumented vehicles.

8.3 LIMITATIONS AND FUTURE WORK

Additional research work that can be conducted includes the followings:

- Since the GDOT crash database includes information regarding the manner of collision and level of severity (e.g., number of injuries and fatalities) , the model should be extended to examine the relationship of speed profile characteristics with different crash types and different degrees of severity.

- The geometric elements and roadside features that can be readily obtained could be incorporated in the crash prediction model to examine the improvement in predictive power.

- The design consistency indicator, $SVLIM$, measures the variation of operating speed from the speed limit. However, this variable was not found to be significant in the final model. It might be interesting to see how this measure is improved when only the over speed data, i.e., speed data above the speed limit, are included.

Additionally, the following is suggested as the future research when the required data are available:

- The study corridors were selected mainly based on availability of GPS data and the need of FHWA study, Effects of Urban Street Environment on Operating Speeds. Therefore, the selected sites do not necessarily portray an unbiased distribution of crash data. A future study should include a large sample of randomly selected corridors.

- It is also possible that additional variables potentially influencing road safety have not been included in this study. For example, a bias in driver demographics across the corridors may introduce a bias in the incident statistics. Future research should investigate the impact of driver and vehicle characteristics on the surrogate measures.

- The speed measures, stop frequency, and acceleration noise are partially influenced by the traffic congestion on the corridor. The trip data distribution used to calculate these measures may potentially be biased by time of day drivers of the instrumented vehicles tended to traverse the corridor. For example, if most of the instrumented vehicles travel on a particular corridor during the peak time and only a few trips are made during off-peak, the stop frequency and the acceleration noise determined are likely to be overestimated for the corridor. Future research into the impact of the trip sampling on measures should be conducted.

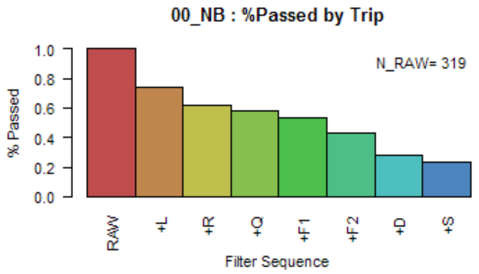
- When more speed data are available during the inclement weather and nighttime period, speed and safety characteristics during inclement weather and low visibility condition should be inspected as these might expose additional geometric design or other roadway characteristic issues.

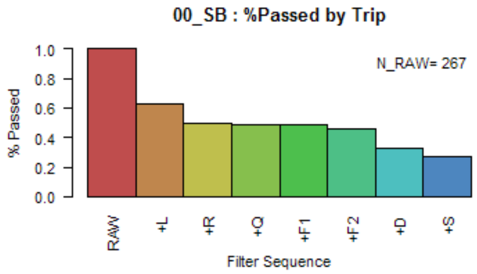
- Regarding the modeling approach, each direction of the road should be model separately as each direction potentially has different speed and safety characteristics. In this study, accidents were combined from the two directions of travel as the impact direction of travel could not be accurately identified from the given crash database. Additional explanatory power in the model is expected when the model incorporates direction of travel.

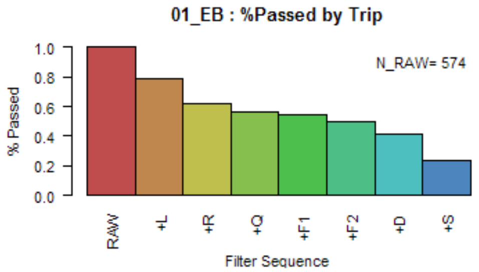
APPENDIX A: SUMMARY OF DATA PROCESSING RESULTS

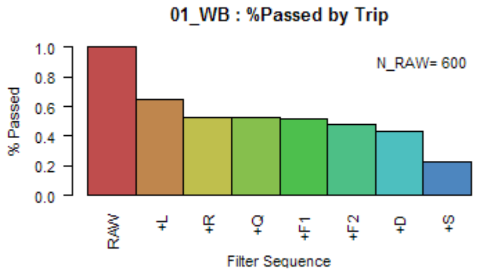
The following bar charts are the summary of data reduction from data processing steps performed in Chapter 4. Each figure represents percent trips that passed the filtering criteria of one directional corridor. Number of trips of the initial data set is placed at the top right corner of the chart. The notations of the filters are described below:

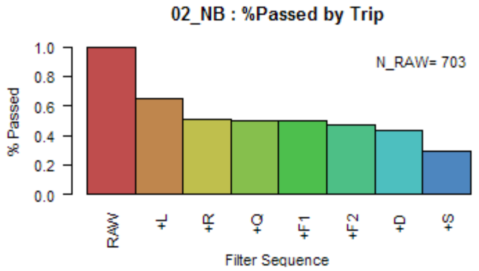
- L=Light
- W=Weekday
- R=No Rain
- F=Free-Flow
- O=Off-Peak
- D=Deviated Trips
- S=GPS Signal

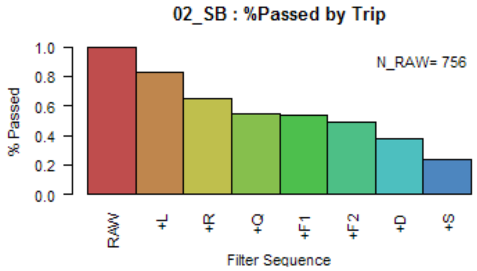

00_NB : %Passed by Trip
% Passed
1.0
0.8
0.6
0.4
0.2
0.0
N_RAW= 319
RAW
+L
+R
+Q
+F1
+F2
+D
+S
Filter Sequence

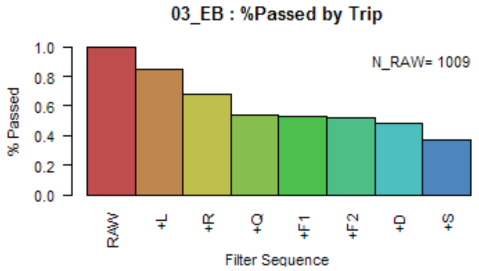

00_SB : %Passed by Trip
% Passed
1.0
0.8
0.6
0.4
0.2
0.0
N_RAW= 267
RAW
+L
+R
+Q
+F1
+F2
+D
+S
Filter Sequence

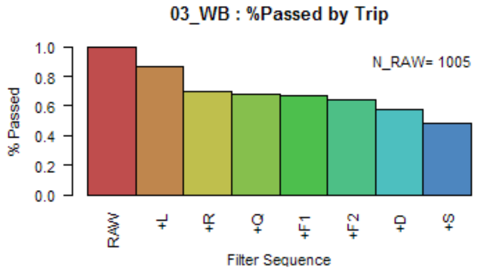

01_EB : %Passed by Trip
% Passed
1.0
0.8
0.6
0.4
0.2
0.0
N_RAW= 574
RAW
+L
+R
+Q
+F1
+F2
+D
+S
Filter Sequence

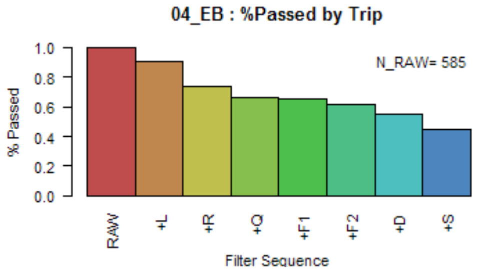

01_WB : %Passed by Trip
% Passed
1.0
0.8
0.6
0.4
0.2
0.0
N_RAW= 600
RAW
+L
+R
+Q
+F1
+F2
+D
+S
Filter Sequence

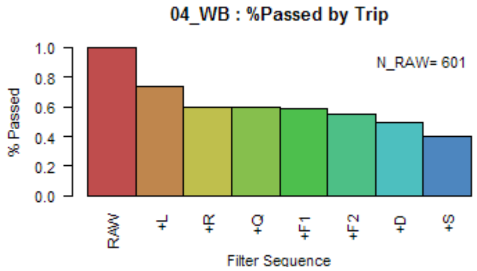

02_NB : %Passed by Trip
% Passed
1.0
0.8
0.6
0.4
0.2
0.0
N_RAW= 703
RAW
+L
+R
+Q
+F1
+F2
+D
+S
Filter Sequence


02_SB : %Passed by Trip
% Passed
1.0
0.8
0.6
0.4
0.2
0.0
N_RAW= 756
RAW
+L
+R
+Q
+F1
+F2
+D
+S
Filter Sequence

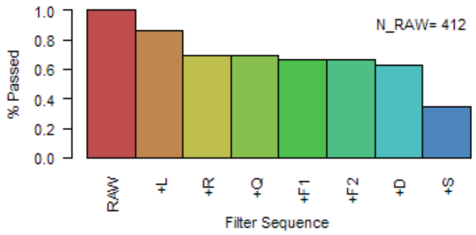

03_EB : %Passed by Trip
% Passed
1.0
0.8
0.6
0.4
0.2
0.0
N_RAW= 1009
RAW
+L
+R
+Q
+F1
+F2
+D
+S
Filter Sequence


03_WB : %Passed by Trip
% Passed
1.0
0.8
0.6
0.4
0.2
0.0
N_RAW= 1005
RAW
+L
+R
+Q
+F1
+F2
+D
+S
Filter Sequence

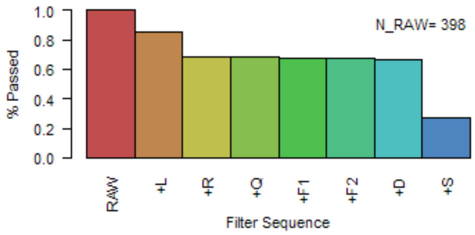

04_EB : %Passed by Trip
% Passed
1.0
0.8
0.6
0.4
0.2
0.0
N_RAW= 585
RAW
+L
+R
+Q
+F1
+F2
+D
+S
Filter Sequence


04_WB : %Passed by Trip
% Passed
1.0
0.8
0.6
0.4
0.2
0.0
N_RAW= 601
RAW
+L
+R
+Q
+F1
+F2
+D
+S
Filter Sequence

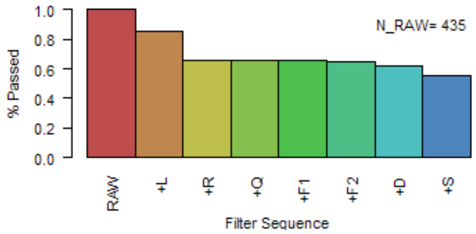
05_NB : %Passed by Trip



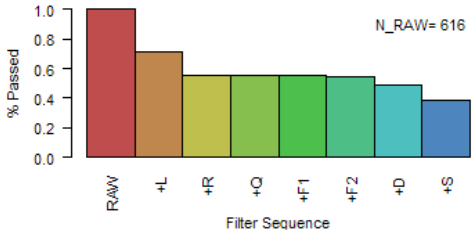
05_SB : %Passed by Trip



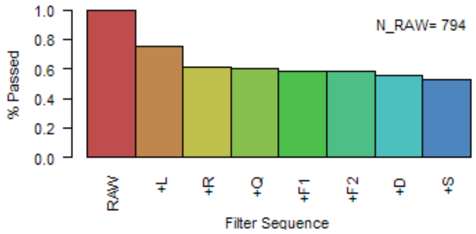
07_NB : %Passed by Trip



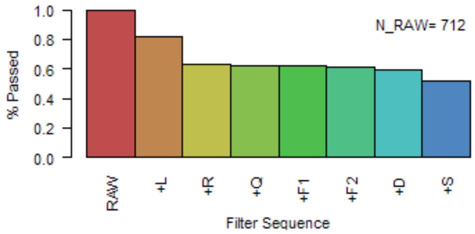
07_SB : %Passed by Trip



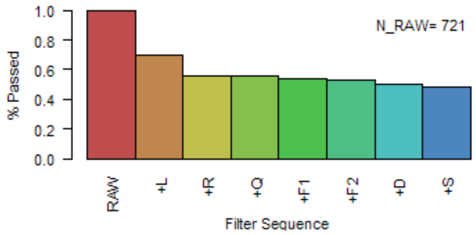
08_EB : %Passed by Trip



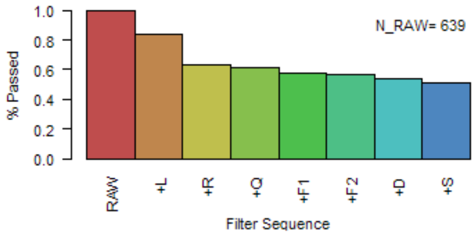
08_WB : %Passed by Trip



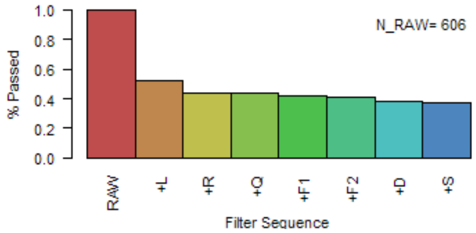
09_EB : %Passed by Trip



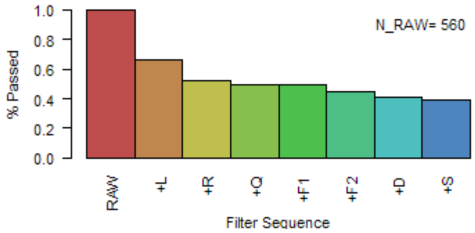
09_WB : %Passed by Trip

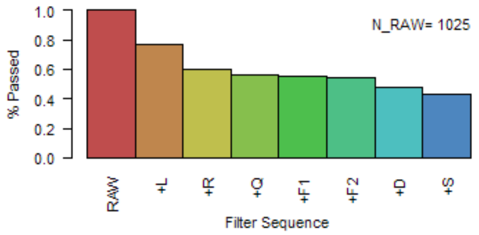


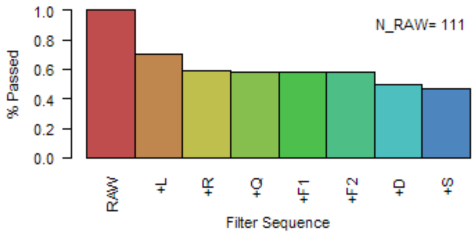
10_NB : %Passed by Trip

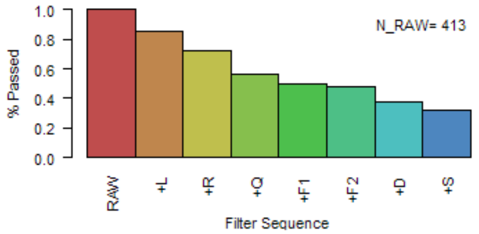


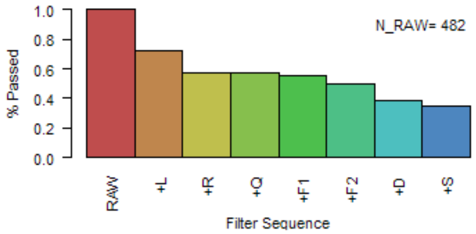
10_SB : %Passed by Trip

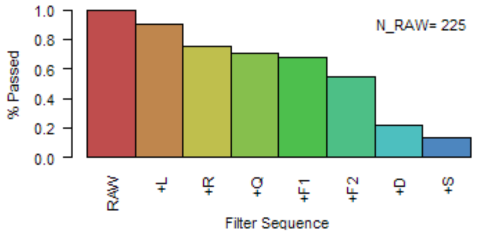


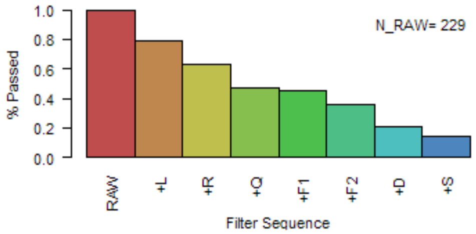

12_NB : %Passed by Trip

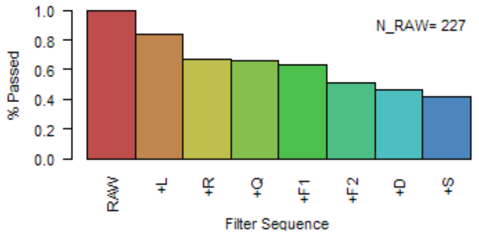

12_SB : %Passed by Trip

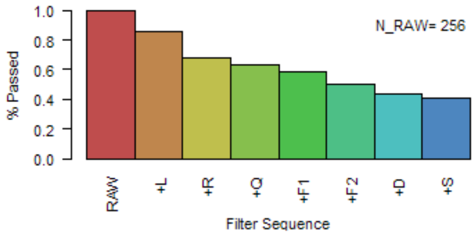

14_EB : %Passed by Trip

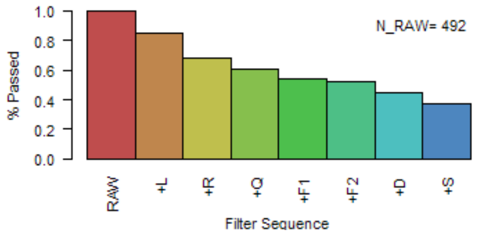

14_WB : %Passed by Trip

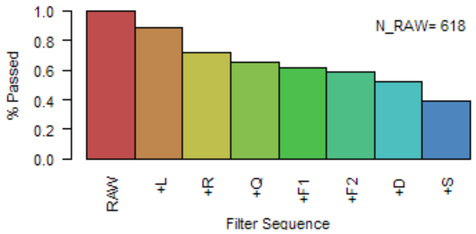

15_EB : %Passed by Trip

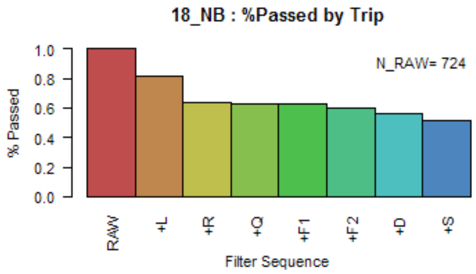

15_WB : %Passed by Trip

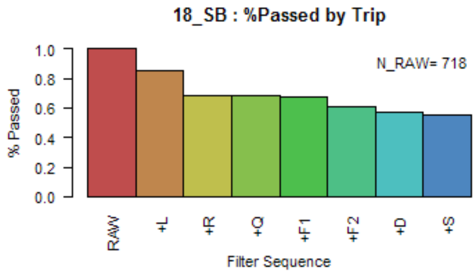

16_NB : %Passed by Trip

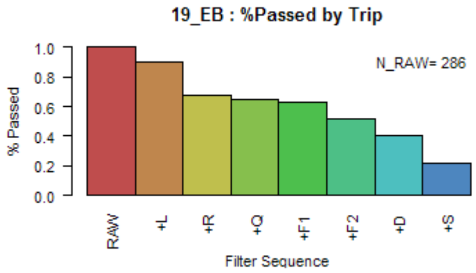

16_SB : %Passed by Trip

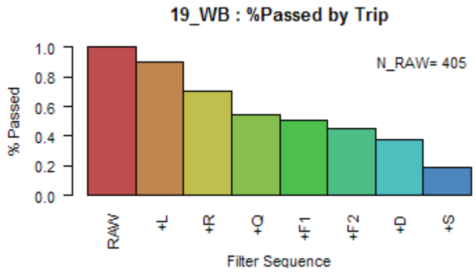

17_NB : %Passed by Trip

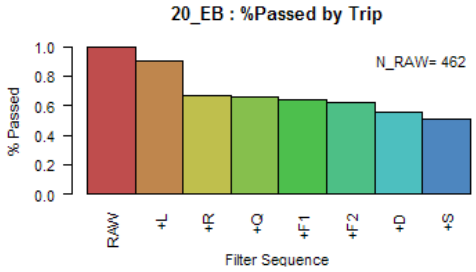

17_SB : %Passed by Trip

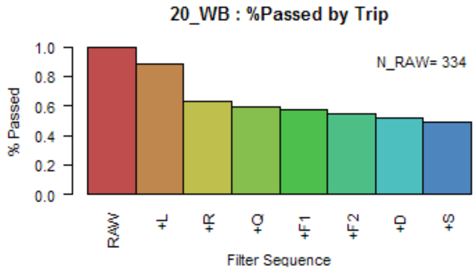

18_NB : %Passed by Trip
N_RAW= 724
% Passed
0.0
0.2
0.4
0.6
0.8
1.0
RAW
+L
+R
+Q
+F1
+F2
+D
+S
Filter Sequence

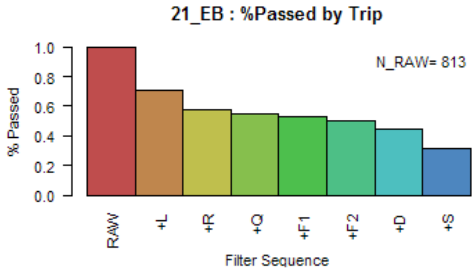

18_SB : %Passed by Trip
N_RAW= 718
% Passed
0.0
0.2
0.4
0.6
0.8
1.0
RAW
+L
+R
+Q
+F1
+F2
+D
+S
Filter Sequence

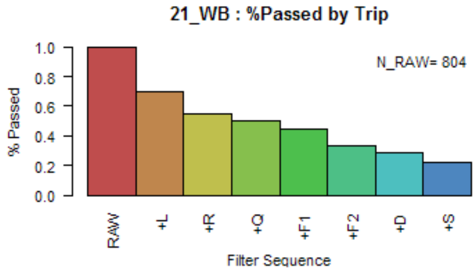

19_EB : %Passed by Trip
N_RAW= 286
% Passed
0.0
0.2
0.4
0.6
0.8
1.0
RAW
+L
+R
+Q
+F1
+F2
+D
+S
Filter Sequence

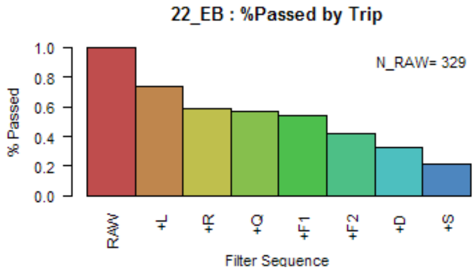

19_WB : %Passed by Trip
N_RAW= 405
% Passed
0.0
0.2
0.4
0.6
0.8
1.0
RAW
+L
+R
+Q
+F1
+F2
+D
+S
Filter Sequence

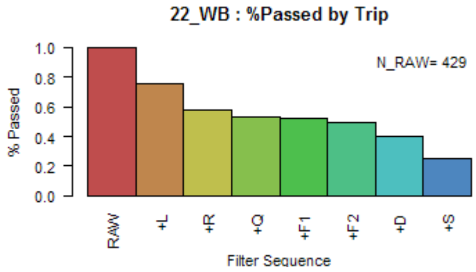

20_EB : %Passed by Trip
N_RAW= 462
% Passed
0.0
0.2
0.4
0.6
0.8
1.0
RAW
+L
+R
+Q
+F1
+F2
+D
+S
Filter Sequence

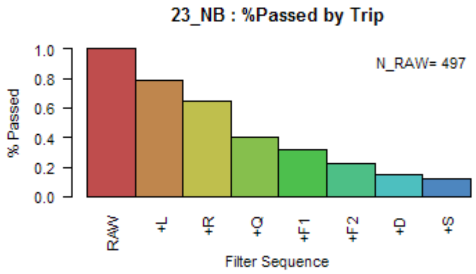

20_WB : %Passed by Trip
N_RAW= 334
% Passed
0.0
0.2
0.4
0.6
0.8
1.0
RAW
+L
+R
+Q
+F1
+F2
+D
+S
Filter Sequence

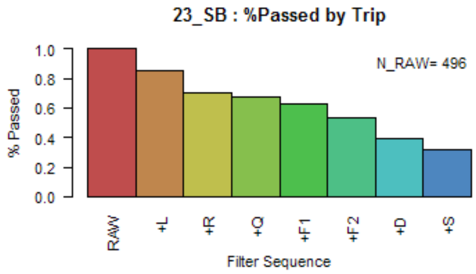

21_EB : %Passed by Trip
N_RAW= 813
% Passed
0.0
0.2
0.4
0.6
0.8
1.0
RAW
+L
+R
+Q
+F1
+F2
+D
+S
Filter Sequence

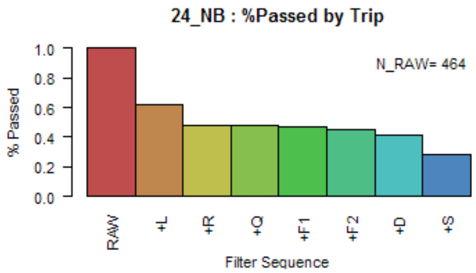

21_WB : %Passed by Trip
N_RAW= 804
% Passed
0.0
0.2
0.4
0.6
0.8
1.0
RAW
+L
+R
+Q
+F1
+F2
+D
+S
Filter Sequence

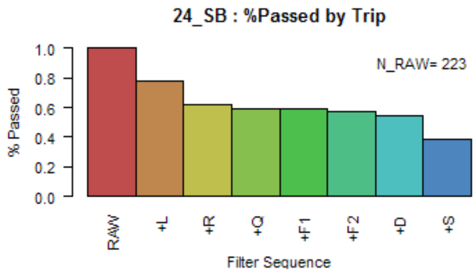

22_EB : %Passed by Trip
N_RAW= 329
% Passed
0.0
0.2
0.4
0.6
0.8
1.0
RAW
+L
+R
+Q
+F1
+F2
+D
+S
Filter Sequence

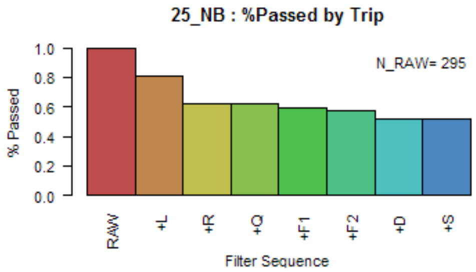

22_WB : %Passed by Trip
N_RAW= 429
% Passed
0.0
0.2
0.4
0.6
0.8
1.0
RAW
+L
+R
+Q
+F1
+F2
+D
+S
Filter Sequence

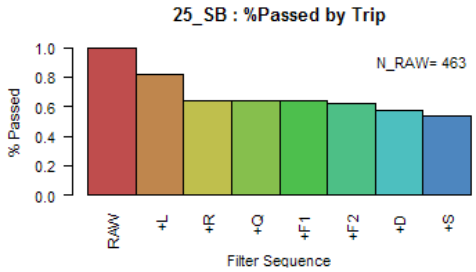

23_NB : %Passed by Trip

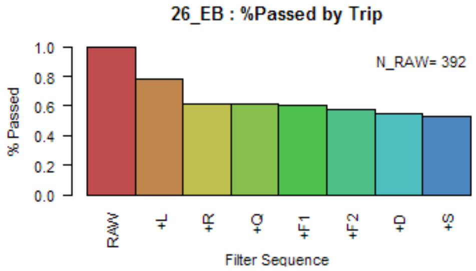

23_SB : %Passed by Trip

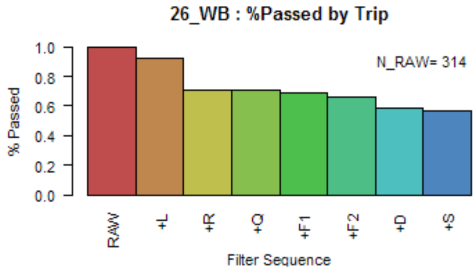

24_NB : %Passed by Trip

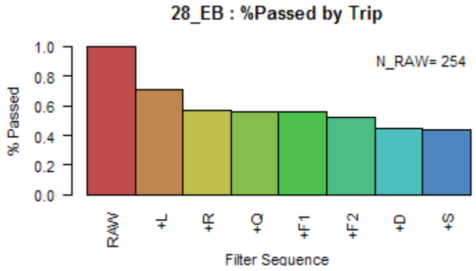

24_SB : %Passed by Trip

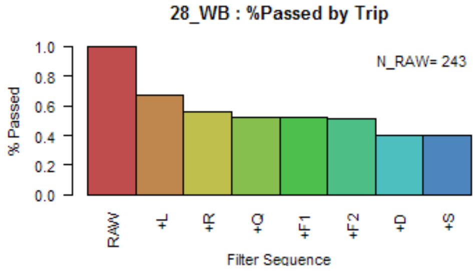

25_NB : %Passed by Trip


25_SB : %Passed by Trip

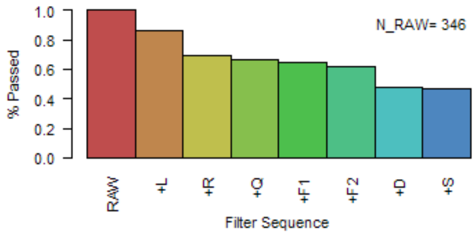

26_EB : %Passed by Trip


26_WB : %Passed by Trip

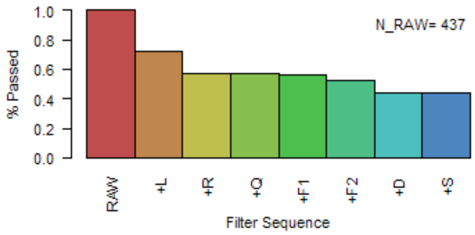

28_EB : %Passed by Trip


28_WB : %Passed by Trip

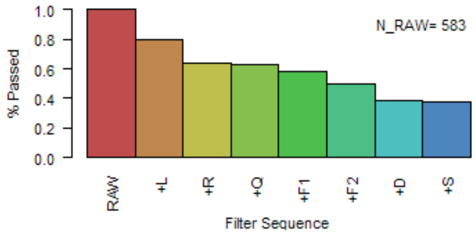
29_EB : %Passed by Trip



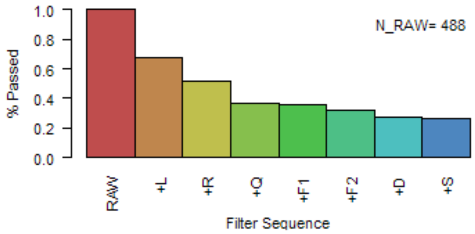
29_WB : %Passed by Trip



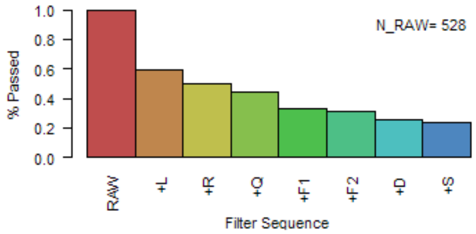
30_NB : %Passed by Trip



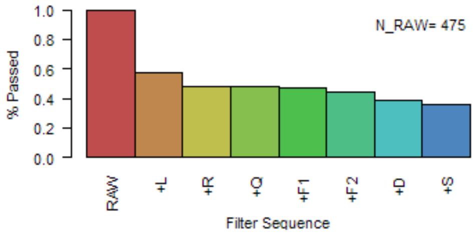
30_SB : %Passed by Trip



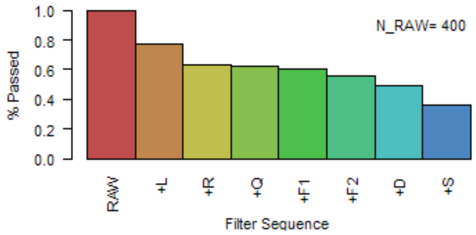
31_EB : %Passed by Trip



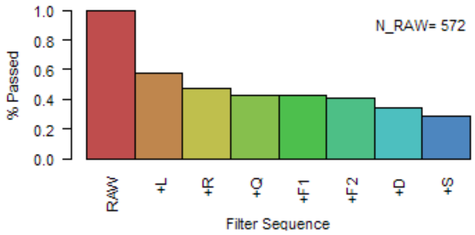
31_WB : %Passed by Trip



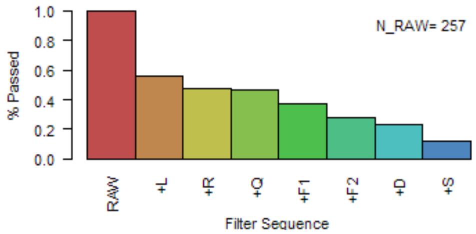
32_NB : %Passed by Trip



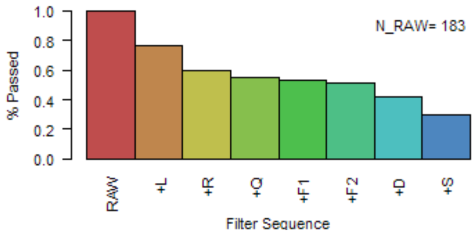
32_SB : %Passed by Trip

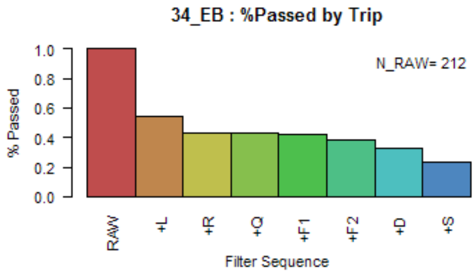


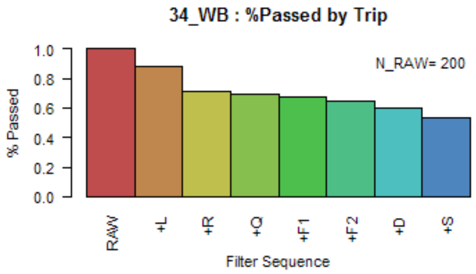
33_NB : %Passed by Trip

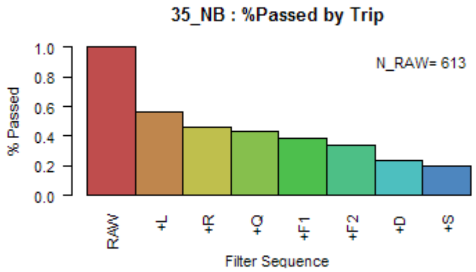


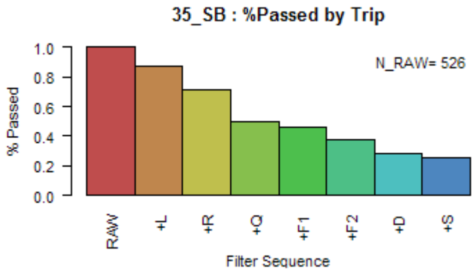
33_SB : %Passed by Trip

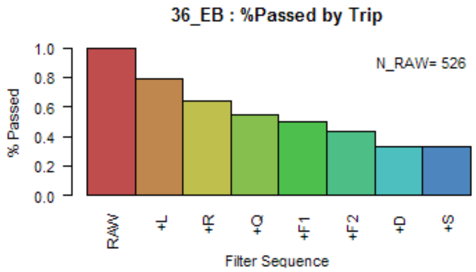


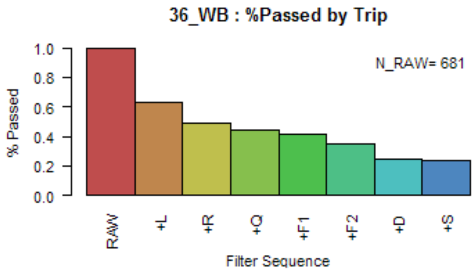

34_EB : %Passed by Trip
N_RAW= 212
% Passed
1.0
0.8
0.6
0.4
0.2
0.0
RAW
+L
+R
+Q
+F1
+F2
+D
+S
Filter Sequence

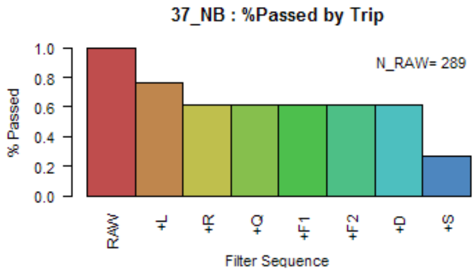

34_WB : %Passed by Trip
N_RAW= 200
% Passed
1.0
0.8
0.6
0.4
0.2
0.0
RAW
+L
+R
+Q
+F1
+F2
+D
+S
Filter Sequence

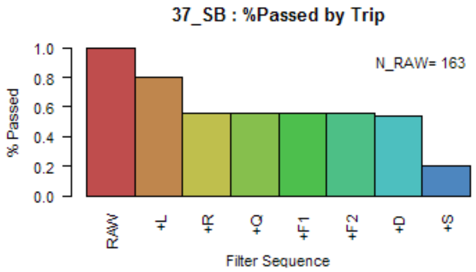

35_NB : %Passed by Trip
N_RAW= 613
% Passed
1.0
0.8
0.6
0.4
0.2
0.0
RAW
+L
+R
+Q
+F1
+F2
+D
+S
Filter Sequence

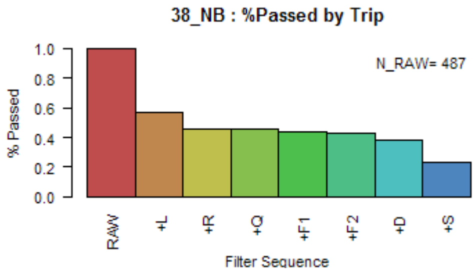

35_SB : %Passed by Trip
N_RAW= 526
% Passed
1.0
0.8
0.6
0.4
0.2
0.0
RAW
+L
+R
+Q
+F1
+F2
+D
+S
Filter Sequence

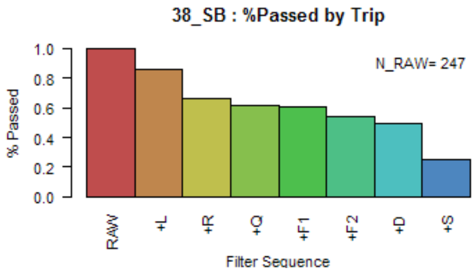

36_EB : %Passed by Trip
N_RAW= 526
% Passed
1.0
0.8
0.6
0.4
0.2
0.0
RAW
+L
+R
+Q
+F1
+F2
+D
+S
Filter Sequence

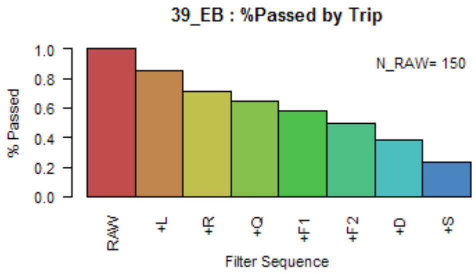

36_WB : %Passed by Trip
N_RAW= 681
% Passed
1.0
0.8
0.6
0.4
0.2
0.0
RAW
+L
+R
+Q
+F1
+F2
+D
+S
Filter Sequence

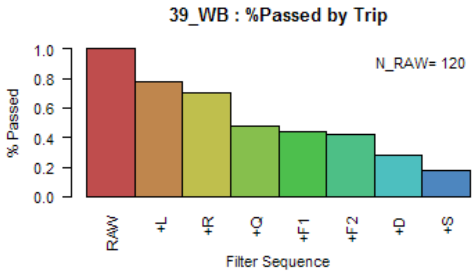

37_NB : %Passed by Trip
N_RAW= 289
% Passed
1.0
0.8
0.6
0.4
0.2
0.0
RAW
+L
+R
+Q
+F1
+F2
+D
+S
Filter Sequence

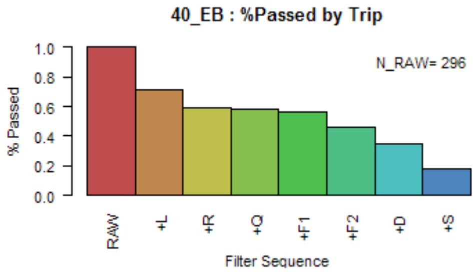

37_SB : %Passed by Trip
N_RAW= 163
% Passed
1.0
0.8
0.6
0.4
0.2
0.0
RAW
+L
+R
+Q
+F1
+F2
+D
+S
Filter Sequence

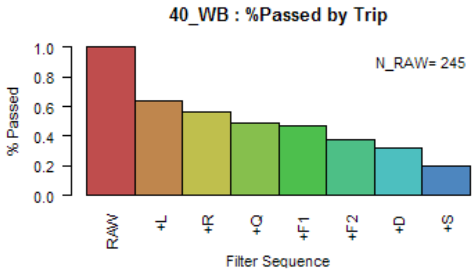

38_NB : %Passed by Trip
N_RAW= 487
% Passed
1.0
0.8
0.6
0.4
0.2
0.0
RAW
+L
+R
+Q
+F1
+F2
+D
+S
Filter Sequence

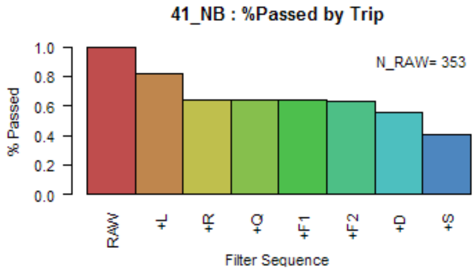

38_SB : %Passed by Trip
N_RAW= 247
% Passed
1.0
0.8
0.6
0.4
0.2
0.0
RAW
+L
+R
+Q
+F1
+F2
+D
+S
Filter Sequence

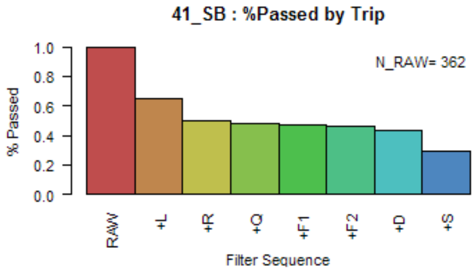

39_EB : %Passed by Trip
N_RAW= 150
% Passed
0.0
0.2
0.4
0.6
0.8
1.0
RAW
+L
+R
+Q
+F1
+F2
+D
+S
Filter Sequence

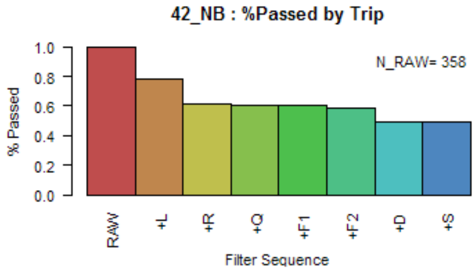

39_WB : %Passed by Trip
N_RAW= 120
% Passed
0.0
0.2
0.4
0.6
0.8
1.0
RAW
+L
+R
+Q
+F1
+F2
+D
+S
Filter Sequence

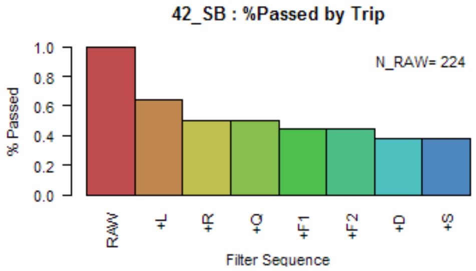

40_EB : %Passed by Trip
N_RAW= 296
% Passed
0.0
0.2
0.4
0.6
0.8
1.0
RAW
+L
+R
+Q
+F1
+F2
+D
+S
Filter Sequence

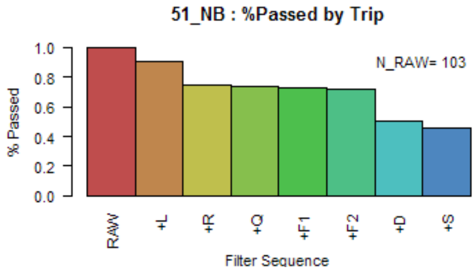

40_WB : %Passed by Trip
N_RAW= 245
% Passed
0.0
0.2
0.4
0.6
0.8
1.0
RAW
+L
+R
+Q
+F1
+F2
+D
+S
Filter Sequence

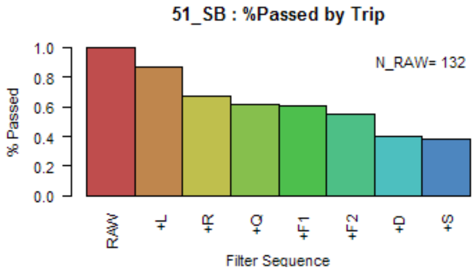

41_NB : %Passed by Trip
N_RAW= 353
% Passed
0.0
0.2
0.4
0.6
0.8
1.0
RAW
+L
+R
+Q
+F1
+F2
+D
+S
Filter Sequence

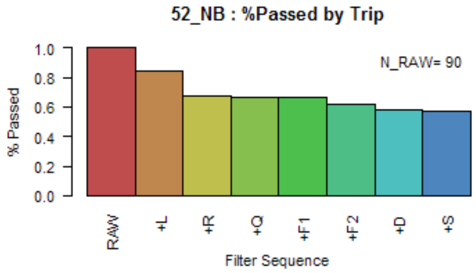

41_SB : %Passed by Trip
N_RAW= 362
% Passed
0.0
0.2
0.4
0.6
0.8
1.0
RAW
+L
+R
+Q
+F1
+F2
+D
+S
Filter Sequence

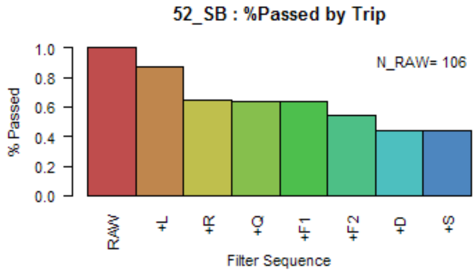

42_NB : %Passed by Trip
N_RAW= 358
% Passed
0.0
0.2
0.4
0.6
0.8
1.0
RAW
+L
+R
+Q
+F1
+F2
+D
+S
Filter Sequence

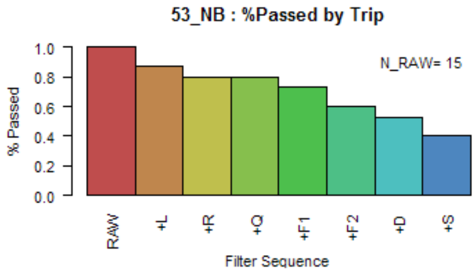

42_SB : %Passed by Trip
N_RAW= 224
% Passed
0.0
0.2
0.4
0.6
0.8
1.0
RAW
+L
+R
+Q
+F1
+F2
+D
+S
Filter Sequence

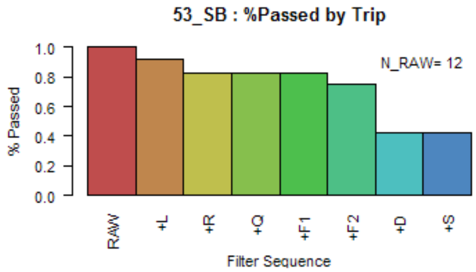

51_NB : %Passed by Trip
N_RAW= 103
% Passed
0.0
0.2
0.4
0.6
0.8
1.0
RAW
+L
+R
+Q
+F1
+F2
+D
+S
Filter Sequence

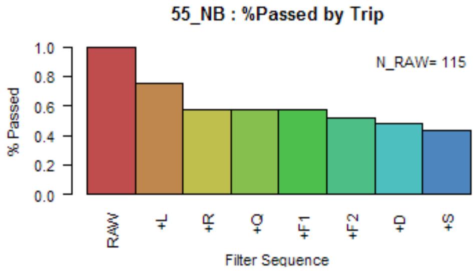

51_SB : %Passed by Trip
N_RAW= 132
% Passed
0.0
0.2
0.4
0.6
0.8
1.0
RAW
+L
+R
+Q
+F1
+F2
+D
+S
Filter Sequence

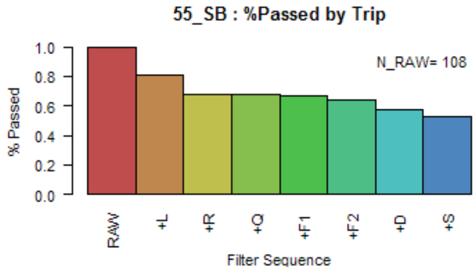

52_NB : %Passed by Trip
N_RAW= 90
% Passed
1.0
0.8
0.6
0.4
0.2
0.0
RAW
+L
+R
+Q
+F1
+F2
+D
+S
Filter Sequence

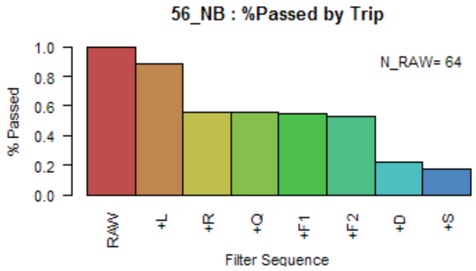

52_SB : %Passed by Trip
N_RAW= 106
% Passed
1.0
0.8
0.6
0.4
0.2
0.0
RAW
+L
+R
+Q
+F1
+F2
+D
+S
Filter Sequence

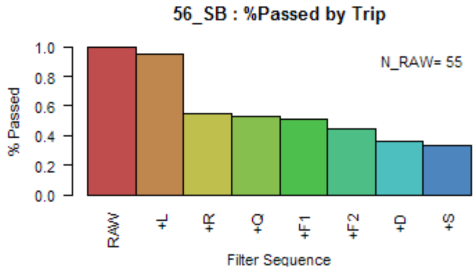

53_NB : %Passed by Trip
N_RAW= 15
% Passed
1.0
0.8
0.6
0.4
0.2
0.0
RAW
+L
+R
+Q
+F1
+F2
+D
+S
Filter Sequence

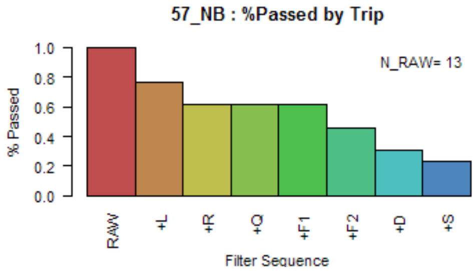

53_SB : %Passed by Trip
N_RAW= 12
% Passed
1.0
0.8
0.6
0.4
0.2
0.0
RAW
+L
+R
+Q
+F1
+F2
+D
+S
Filter Sequence

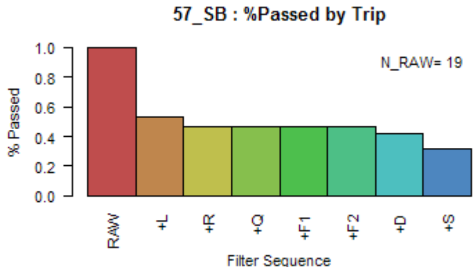

55_NB : %Passed by Trip
N_RAW= 115
% Passed
1.0
0.8
0.6
0.4
0.2
0.0
RAW
+L
+R
+Q
+F1
+F2
+D
+S
Filter Sequence

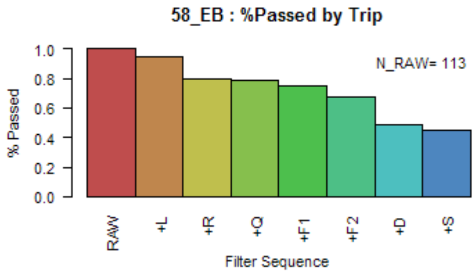

55_SB : %Passed by Trip
N_RAW= 108
% Passed
1.0
0.8
0.6
0.4
0.2
0.0
RAW
+L
+R
+Q
+F1
+F2
+D
+S
Filter Sequence

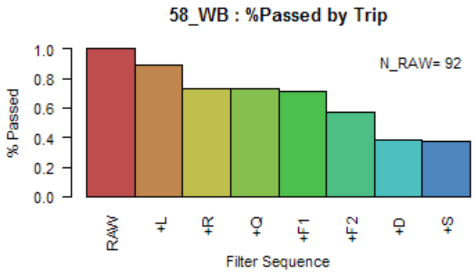

56_NB : %Passed by Trip
N_RAW= 64
% Passed
1.0
0.8
0.6
0.4
0.2
0.0
RAW
+L
+R
+Q
+F1
+F2
+D
+S
Filter Sequence

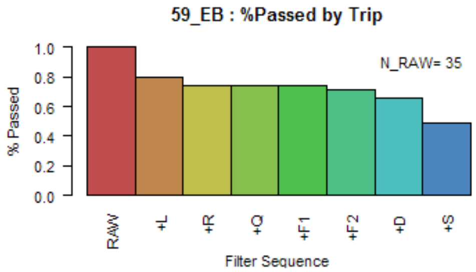

56_SB : %Passed by Trip
N_RAW= 55
% Passed
1.0
0.8
0.6
0.4
0.2
0.0
RAW
+L
+R
+Q
+F1
+F2
+D
+S
Filter Sequence

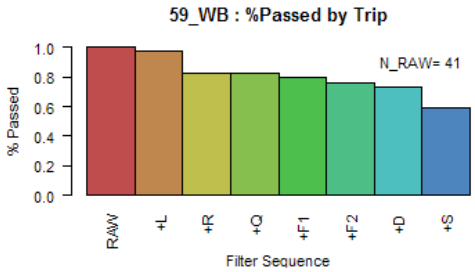

57_NB : %Passed by Trip
N_RAW= 13
% Passed
1.0
0.8
0.6
0.4
0.2
0.0
RAW
+L
+R
+Q
+F1
+F2
+D
+S
Filter Sequence

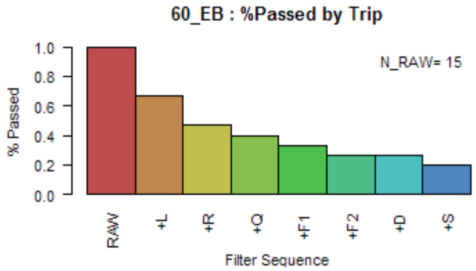

57_SB : %Passed by Trip
N_RAW= 19
% Passed
1.0
0.8
0.6
0.4
0.2
0.0
RAW
+L
+R
+Q
+F1
+F2
+D
+S
Filter Sequence

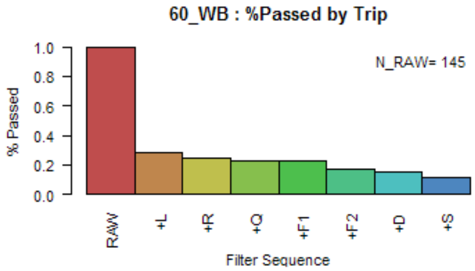

58_EB : %Passed by Trip
N_RAW= 113
% Passed
1.0 0.8 0.6 0.4 0.2 0.0
RAW +L +R +Q +F1 +F2 +D +S
Filter Sequence

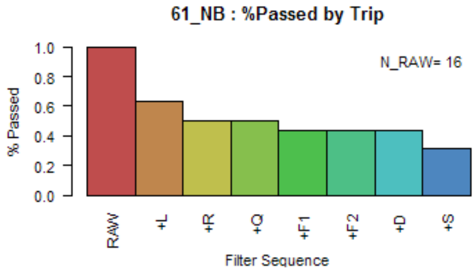

58_WB : %Passed by Trip
N_RAW= 92
% Passed
1.0 0.8 0.6 0.4 0.2 0.0
RAW +L +R +Q +F1 +F2 +D +S
Filter Sequence

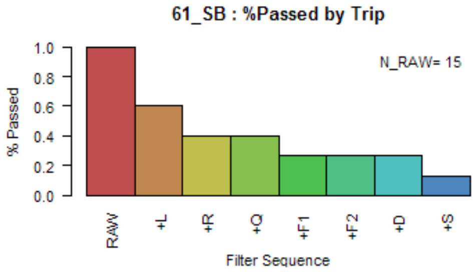

59_EB : %Passed by Trip
N_RAW= 35
% Passed
1.0 0.8 0.6 0.4 0.2 0.0
RAW +L +R +Q +F1 +F2 +D +S
Filter Sequence

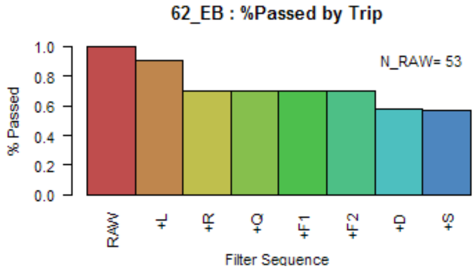

59_WB : %Passed by Trip
N_RAW= 41
% Passed
1.0 0.8 0.6 0.4 0.2 0.0
RAW +L +R +Q +F1 +F2 +D +S
Filter Sequence

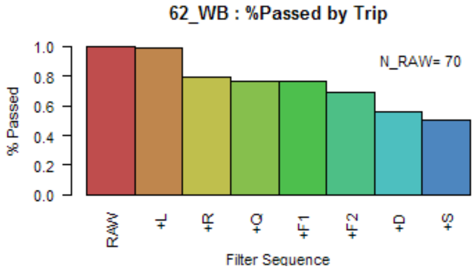

60_EB : %Passed by Trip
N_RAW= 15
% Passed
1.0 0.8 0.6 0.4 0.2 0.0
RAW +L +R +Q +F1 +F2 +D +S
Filter Sequence

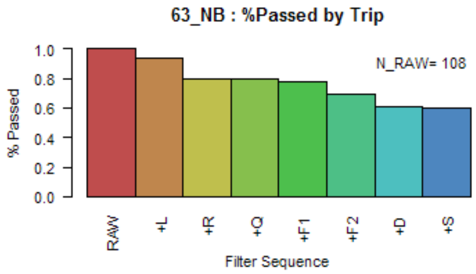

60_WB : %Passed by Trip
N_RAW= 145
% Passed
1.0 0.8 0.6 0.4 0.2 0.0
RAW +L +R +Q +F1 +F2 +D +S
Filter Sequence

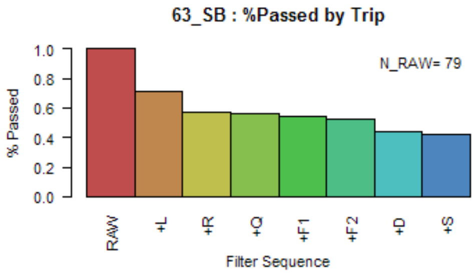

61_NB : %Passed by Trip
N_RAW= 16
% Passed
1.0 0.8 0.6 0.4 0.2 0.0
RAW +L +R +Q +F1 +F2 +D +S
Filter Sequence

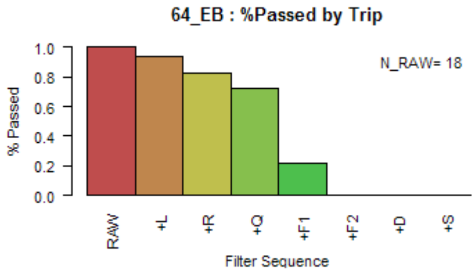

61_SB : %Passed by Trip
N_RAW= 15
% Passed
1.0 0.8 0.6 0.4 0.2 0.0
RAW +L +R +Q +F1 +F2 +D +S
Filter Sequence

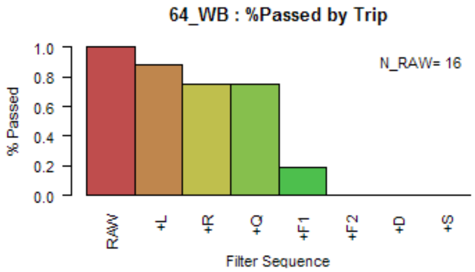

62_EB : %Passed by Trip
N_RAW= 53
% Passed
1.0 0.8 0.6 0.4 0.2 0.0
RAW +L +R +Q +F1 +F2 +D +S
Filter Sequence

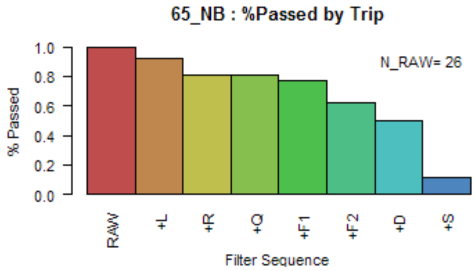

62_WB : %Passed by Trip
N_RAW= 70
% Passed
1.0 0.8 0.6 0.4 0.2 0.0
RAW +L +R +Q +F1 +F2 +D +S
Filter Sequence

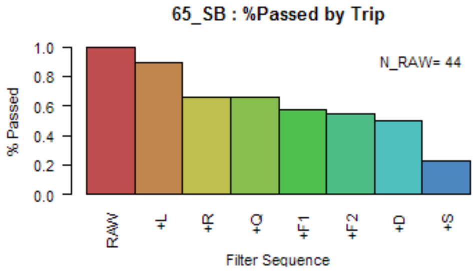

63_NB : %Passed by Trip
N_RAW= 108
% Passed
0.0
0.2
0.4
0.6
0.8
1.0
RAW
+L
+R
+Q
+F1
+F2
+D
+S
Filter Sequence

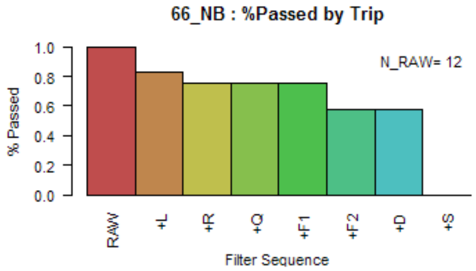

63_SB : %Passed by Trip
N_RAW= 79
% Passed
0.0
0.2
0.4
0.6
0.8
1.0
RAW
+L
+R
+Q
+F1
+F2
+D
+S
Filter Sequence

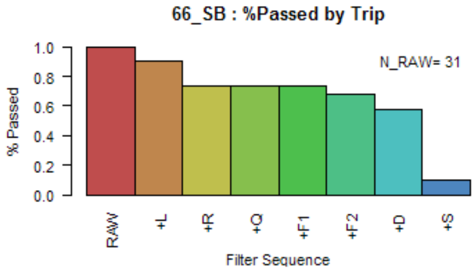

64_EB : %Passed by Trip
N_RAW= 18
% Passed
0.0
0.2
0.4
0.6
0.8
1.0
RAW
+L
+R
+Q
+F1
+F2
+D
+S
Filter Sequence

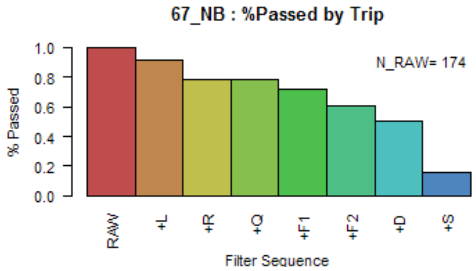

64_WB : %Passed by Trip
N_RAW= 16
% Passed
0.0
0.2
0.4
0.6
0.8
1.0
RAW
+L
+R
+Q
+F1
+F2
+D
+S
Filter Sequence

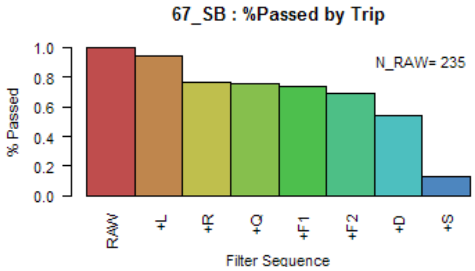

65_NB : %Passed by Trip
N_RAW= 26
% Passed
0.0
0.2
0.4
0.6
0.8
1.0
RAW
+L
+R
+Q
+F1
+F2
+D
+S
Filter Sequence

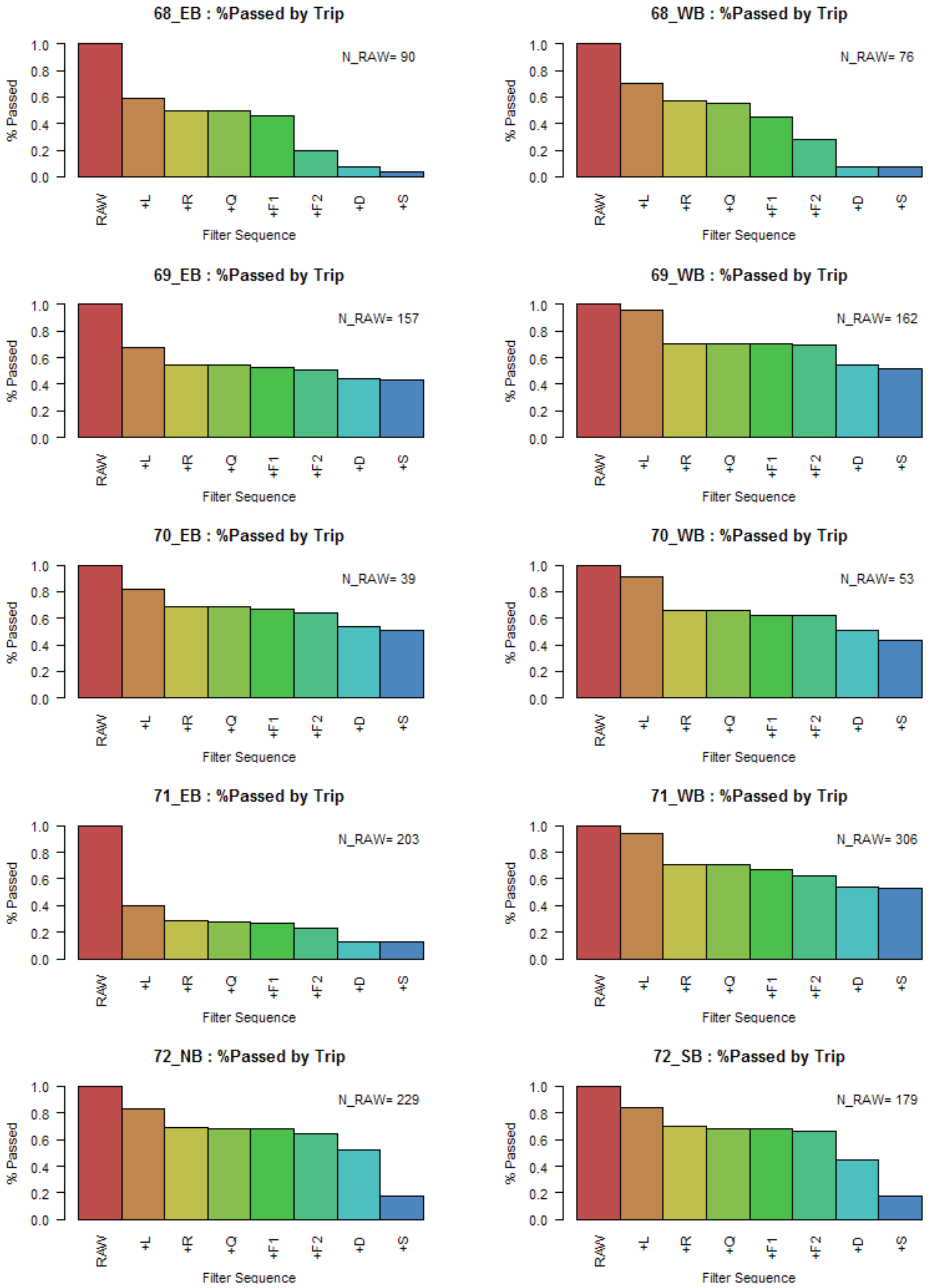

65_SB : %Passed by Trip
N_RAW= 44
% Passed
0.0
0.2
0.4
0.6
0.8
1.0
RAW
+L
+R
+Q
+F1
+F2
+D
+S
Filter Sequence

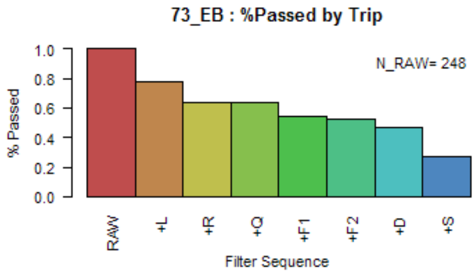

66_NB : %Passed by Trip
N_RAW= 12
% Passed
0.0
0.2
0.4
0.6
0.8
1.0
RAW
+L
+R
+Q
+F1
+F2
+D
+S
Filter Sequence

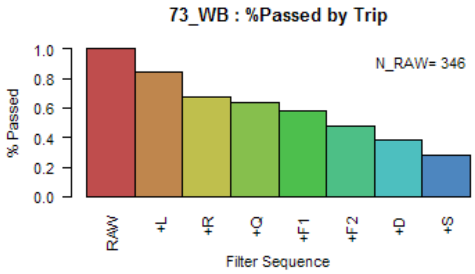

66_SB : %Passed by Trip
N_RAW= 31
% Passed
0.0
0.2
0.4
0.6
0.8
1.0
RAW
+L
+R
+Q
+F1
+F2
+D
+S
Filter Sequence

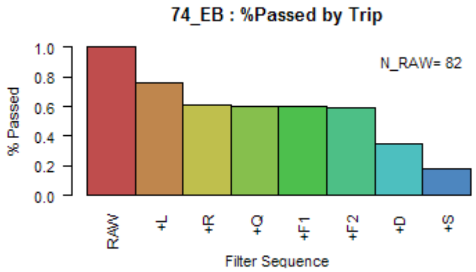

67_NB : %Passed by Trip
N_RAW= 174
% Passed
0.0
0.2
0.4
0.6
0.8
1.0
RAW
+L
+R
+Q
+F1
+F2
+D
+S
Filter Sequence

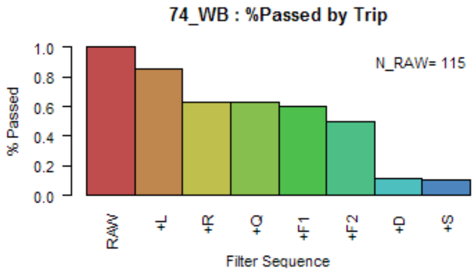

67_SB : %Passed by Trip
N_RAW= 235
% Passed
0.0
0.2
0.4
0.6
0.8
1.0
RAW
+L
+R
+Q
+F1
+F2
+D
+S
Filter Sequence

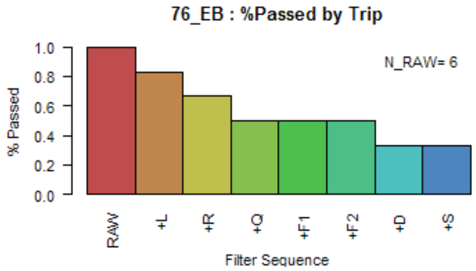

68_EB : %Passed by Trip
N_RAW= 90
68_WB : %Passed by Trip
N_RAW= 76
69_EB : %Passed by Trip
N_RAW= 157
69_WB : %Passed by Trip
N_RAW= 162
70_EB : %Passed by Trip
N_RAW= 39
70_WB : %Passed by Trip
N_RAW= 53
71_EB : %Passed by Trip
N_RAW= 203
71_WB : %Passed by Trip
N_RAW= 306
72_NB : %Passed by Trip
N_RAW= 229
72_SB : %Passed by Trip
N_RAW= 179
% Passed
0.0
0.2
0.4
0.6
0.8
1.0
RAW
+L
+R
+Q
+F1
+F2
+D
+S
Filter Sequence

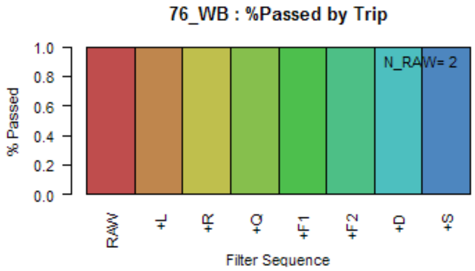

73_EB : %Passed by Trip

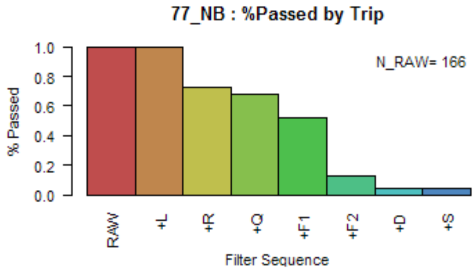

73_WB : %Passed by Trip

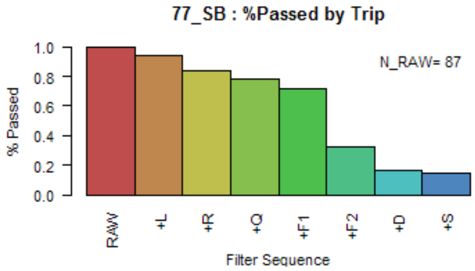

74_EB : %Passed by Trip

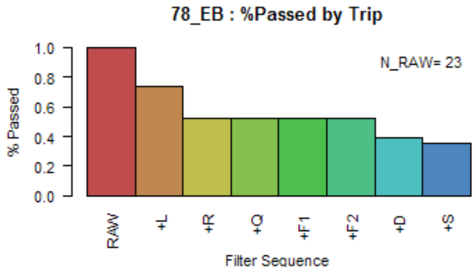

74_WB : %Passed by Trip

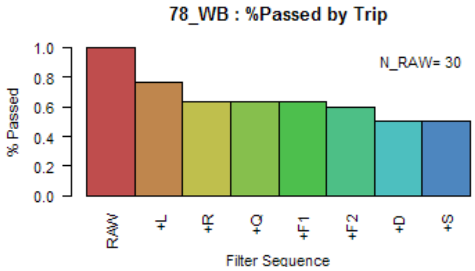

76_EB : %Passed by Trip

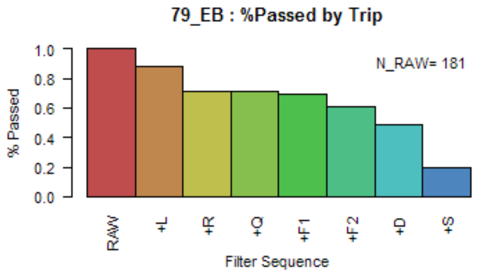

76_WB : %Passed by Trip

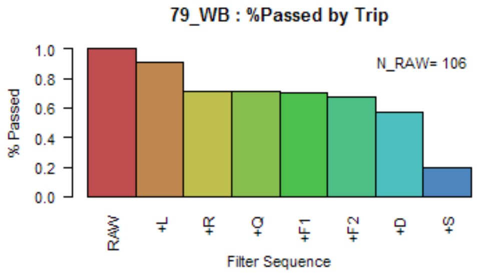

77_NB : %Passed by Trip

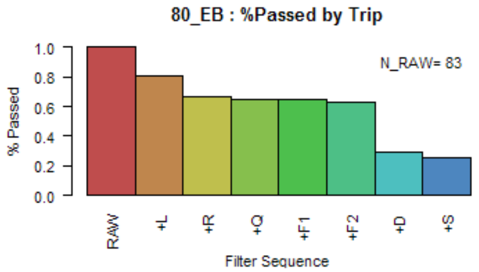

77_SB : %Passed by Trip

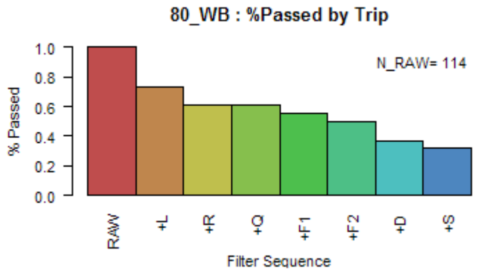

78_EB : %Passed by Trip

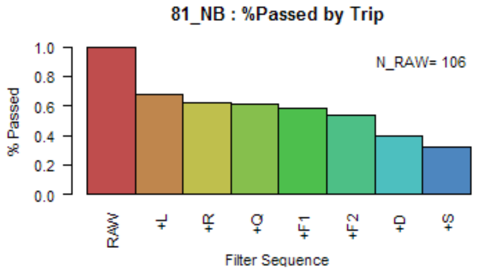

78_WB : %Passed by Trip

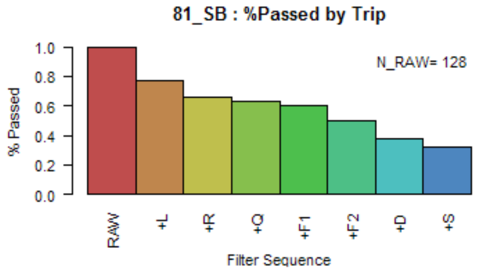

79_EB : %Passed by Trip
N_RAW= 181
% Passed
1.0
0.8
0.6
0.4
0.2
0.0
RAW
+L
+R
+Q
+F1
+F2
+D
+S
Filter Sequence

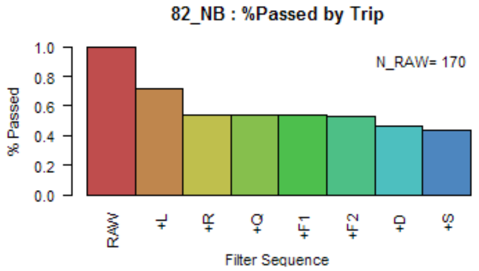

79_WB : %Passed by Trip
N_RAW= 106
% Passed
1.0
0.8
0.6
0.4
0.2
0.0
RAW
+L
+R
+Q
+F1
+F2
+D
+S
Filter Sequence

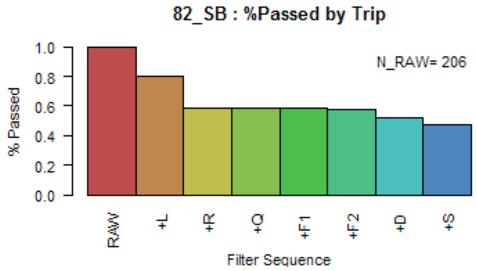

80_EB : %Passed by Trip
N_RAW= 83
% Passed
1.0
0.8
0.6
0.4
0.2
0.0
RAW
+L
+R
+Q
+F1
+F2
+D
+S
Filter Sequence

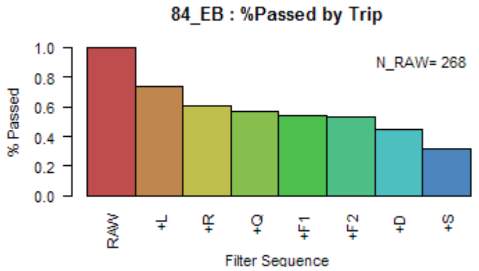

80_WB : %Passed by Trip
N_RAW= 114
% Passed
1.0
0.8
0.6
0.4
0.2
0.0
RAW
+L
+R
+Q
+F1
+F2
+D
+S
Filter Sequence

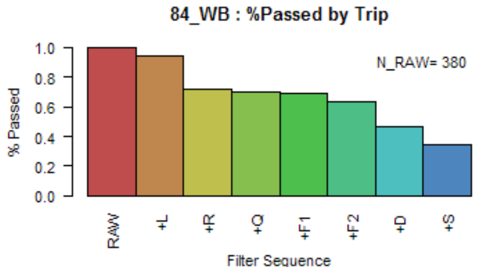

81_NB : %Passed by Trip
N_RAW= 106
% Passed
1.0
0.8
0.6
0.4
0.2
0.0
RAW
+L
+R
+Q
+F1
+F2
+D
+S
Filter Sequence

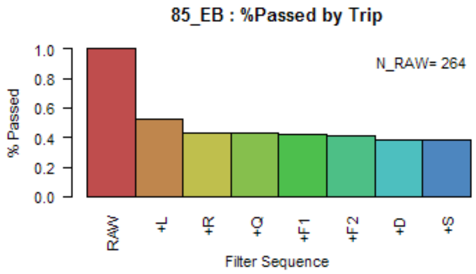

81_SB : %Passed by Trip
N_RAW= 128
% Passed
1.0
0.8
0.6
0.4
0.2
0.0
RAW
+L
+R
+Q
+F1
+F2
+D
+S
Filter Sequence

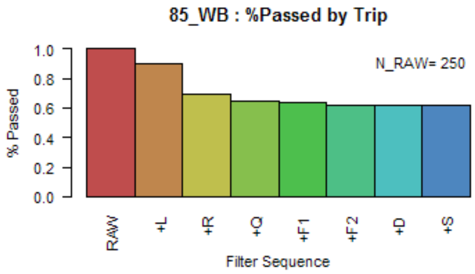

82_NB : %Passed by Trip
N_RAW= 170
% Passed
1.0
0.8
0.6
0.4
0.2
0.0
RAW
+L
+R
+Q
+F1
+F2
+D
+S
Filter Sequence

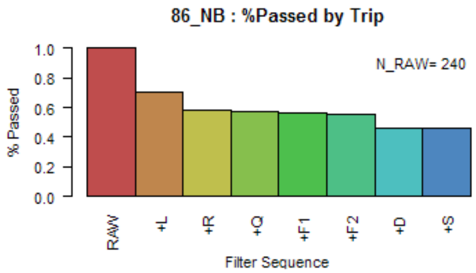

82_SB : %Passed by Trip
N_RAW= 206
% Passed
1.0
0.8
0.6
0.4
0.2
0.0
RAW
+L
+R
+Q
+F1
+F2
+D
+S
Filter Sequence

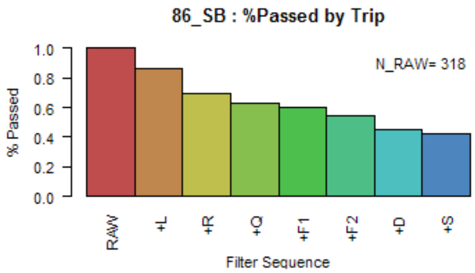

84_EB : %Passed by Trip
N_RAW= 268
% Passed
1.0
0.8
0.6
0.4
0.2
0.0
RAW
+L
+R
+Q
+F1
+F2
+D
+S
Filter Sequence

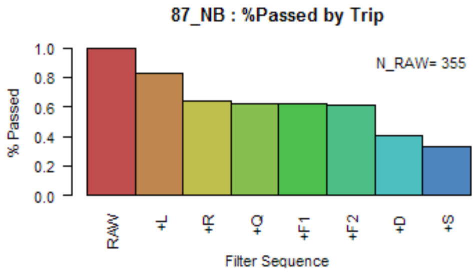

84_WB : %Passed by Trip
N_RAW= 380
% Passed
1.0
0.8
0.6
0.4
0.2
0.0
RAW
+L
+R
+Q
+F1
+F2
+D
+S
Filter Sequence

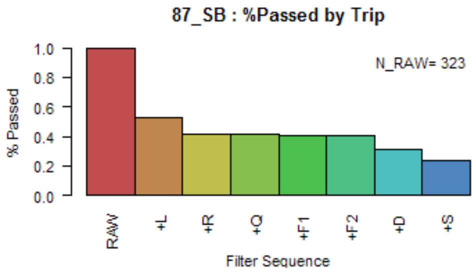

85_EB : %Passed by Trip
N_RAW= 264
% Passed
0.0
0.2
0.4
0.6
0.8
1.0
RAW
+L
+R
+Q
+F1
+F2
+D
+S
Filter Sequence

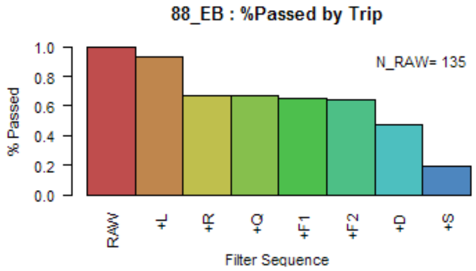

85_WB : %Passed by Trip
N_RAW= 250
% Passed
0.0
0.2
0.4
0.6
0.8
1.0
RAW
+L
+R
+Q
+F1
+F2
+D
+S
Filter Sequence

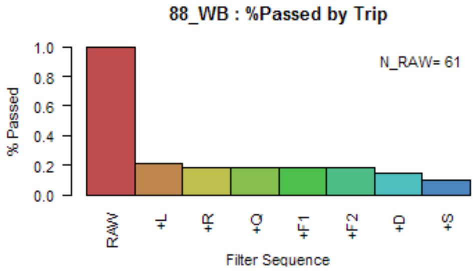

86_NB : %Passed by Trip
N_RAW= 240
% Passed
0.0
0.2
0.4
0.6
0.8
1.0
RAW
+L
+R
+Q
+F1
+F2
+D
+S
Filter Sequence

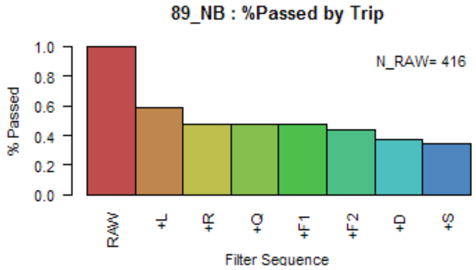

86_SB : %Passed by Trip
N_RAW= 318
% Passed
0.0
0.2
0.4
0.6
0.8
1.0
RAW
+L
+R
+Q
+F1
+F2
+D
+S
Filter Sequence

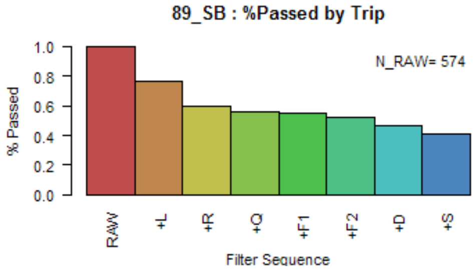

87_NB : %Passed by Trip
N_RAW= 355
% Passed
0.0
0.2
0.4
0.6
0.8
1.0
RAW
+L
+R
+Q
+F1
+F2
+D
+S
Filter Sequence


87_SB : %Passed by Trip
N_RAW= 323
% Passed
0.0
0.2
0.4
0.6
0.8
1.0
RAW
+L
+R
+Q
+F1
+F2
+D
+S
Filter Sequence

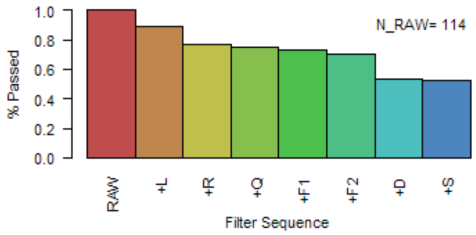

88_EB : %Passed by Trip
N_RAW= 135
% Passed
0.0
0.2
0.4
0.6
0.8
1.0
RAW
+L
+R
+Q
+F1
+F2
+D
+S
Filter Sequence


88_WB : %Passed by Trip
N_RAW= 61
% Passed
0.0
0.2
0.4
0.6
0.8
1.0
RAW
+L
+R
+Q
+F1
+F2
+D
+S
Filter Sequence

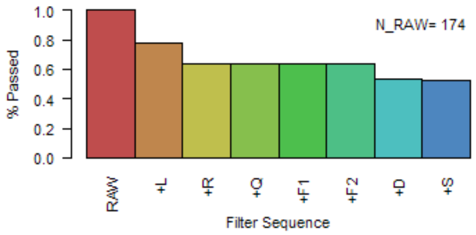

89_NB : %Passed by Trip
N_RAW= 416
% Passed
0.0
0.2
0.4
0.6
0.8
1.0
RAW
+L
+R
+Q
+F1
+F2
+D
+S
Filter Sequence


89_SB : %Passed by Trip
N_RAW= 574
% Passed
0.0
0.2
0.4
0.6
0.8
1.0
RAW
+L
+R
+Q
+F1
+F2
+D
+S
Filter Sequence

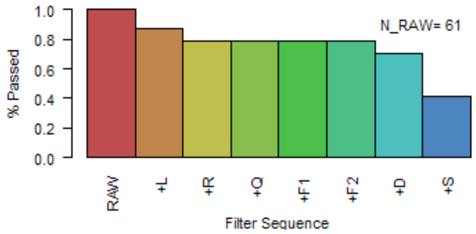
90_NB : %Passed by Trip



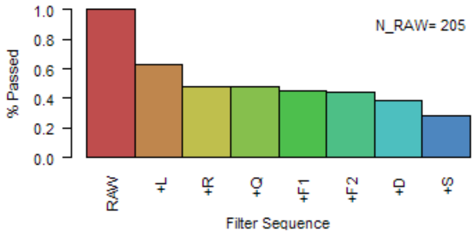
90_SB : %Passed by Trip



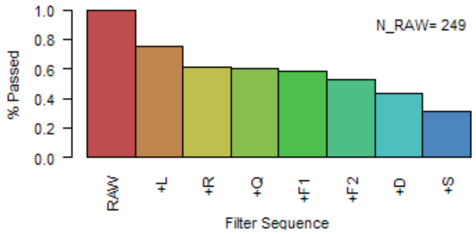
91_NB : %Passed by Trip



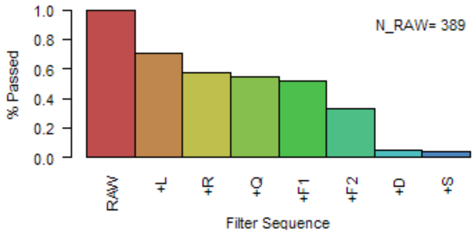
91_SB : %Passed by Trip



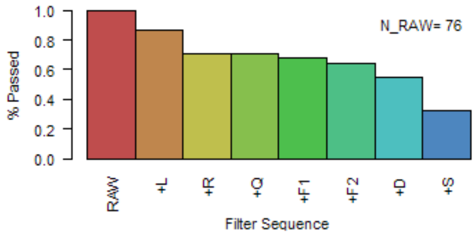
92_NB : %Passed by Trip



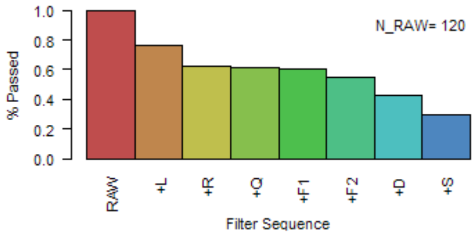
92_SB : %Passed by Trip



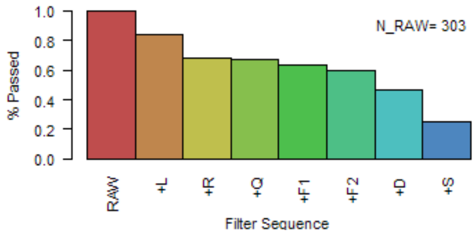
93_NB : %Passed by Trip



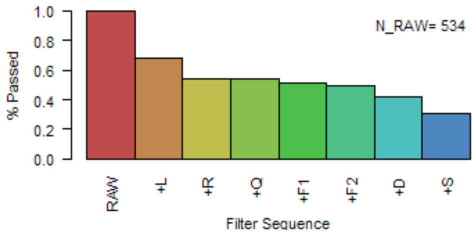
93_SB : %Passed by Trip

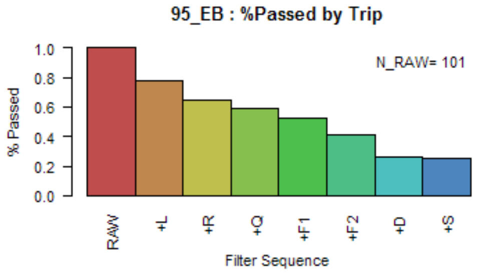


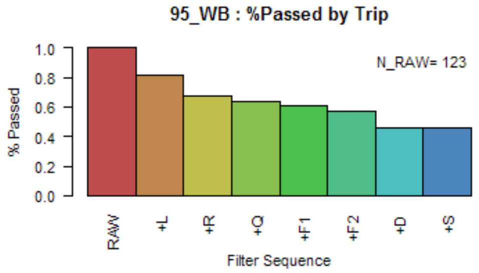
94_NB : %Passed by Trip

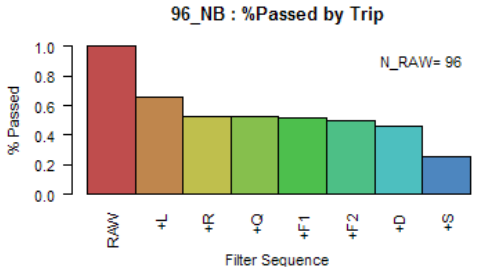


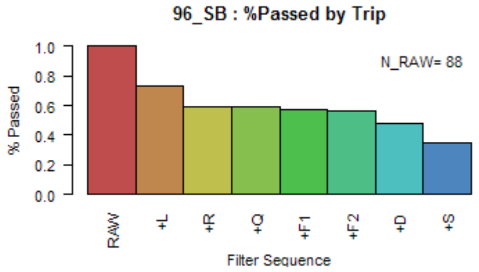
94_SB : %Passed by Trip

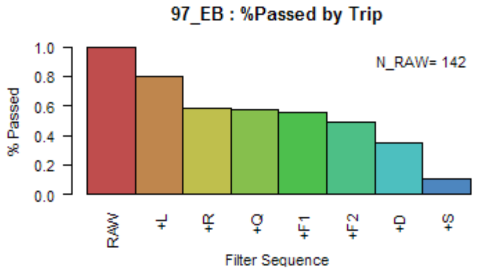


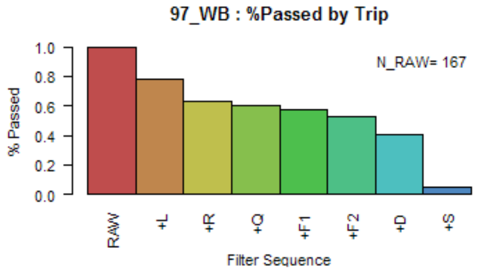

95_EB : %Passed by Trip
N_RAW= 101
% Passed
0.0
0.2
0.4
0.6
0.8
1.0
RAW
+L
+R
+Q
+F1
+F2
+D
+S
Filter Sequence

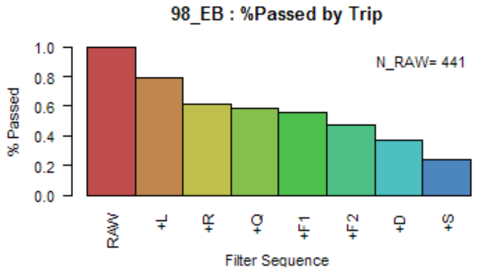

95_WB : %Passed by Trip
N_RAW= 123
% Passed
0.0
0.2
0.4
0.6
0.8
1.0
RAW
+L
+R
+Q
+F1
+F2
+D
+S
Filter Sequence

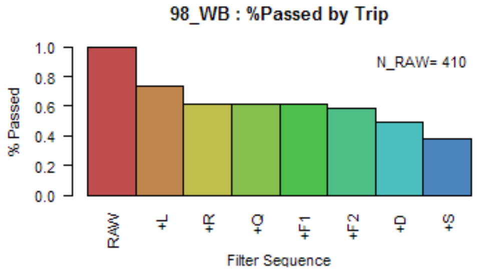

96_NB : %Passed by Trip
N_RAW= 96
% Passed
0.0
0.2
0.4
0.6
0.8
1.0
RAW
+L
+R
+Q
+F1
+F2
+D
+S
Filter Sequence

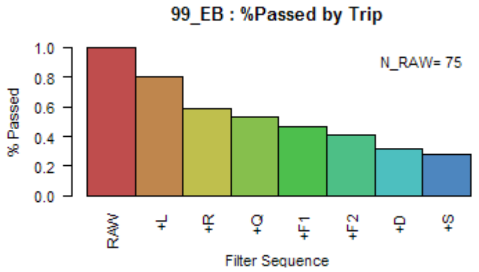

96_SB : %Passed by Trip
N_RAW= 88
% Passed
0.0
0.2
0.4
0.6
0.8
1.0
RAW
+L
+R
+Q
+F1
+F2
+D
+S
Filter Sequence

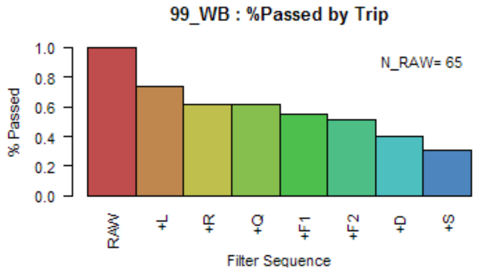

97_EB : %Passed by Trip
N_RAW= 142
% Passed
0.0
0.2
0.4
0.6
0.8
1.0
RAW
+L
+R
+Q
+F1
+F2
+D
+S
Filter Sequence


97_WB : %Passed by Trip
N_RAW= 167
% Passed
0.0
0.2
0.4
0.6
0.8
1.0
RAW
+L
+R
+Q
+F1
+F2
+D
+S
Filter Sequence


98_EB : %Passed by Trip
N_RAW= 441
% Passed
0.0
0.2
0.4
0.6
0.8
1.0
RAW
+L
+R
+Q
+F1
+F2
+D
+S
Filter Sequence


98_WB : %Passed by Trip
N_RAW= 410
% Passed
0.0
0.2
0.4
0.6
0.8
1.0
RAW
+L
+R
+Q
+F1
+F2
+D
+S
Filter Sequence


99_EB : %Passed by Trip
N_RAW= 75
% Passed
0.0
0.2
0.4
0.6
0.8
1.0
RAW
+L
+R
+Q
+F1
+F2
+D
+S
Filter Sequence


99_WB : %Passed by Trip
N_RAW= 65
% Passed
0.0
0.2
0.4
0.6
0.8
1.0
RAW
+L
+R
+Q
+F1
+F2
+D
+S
Filter Sequence

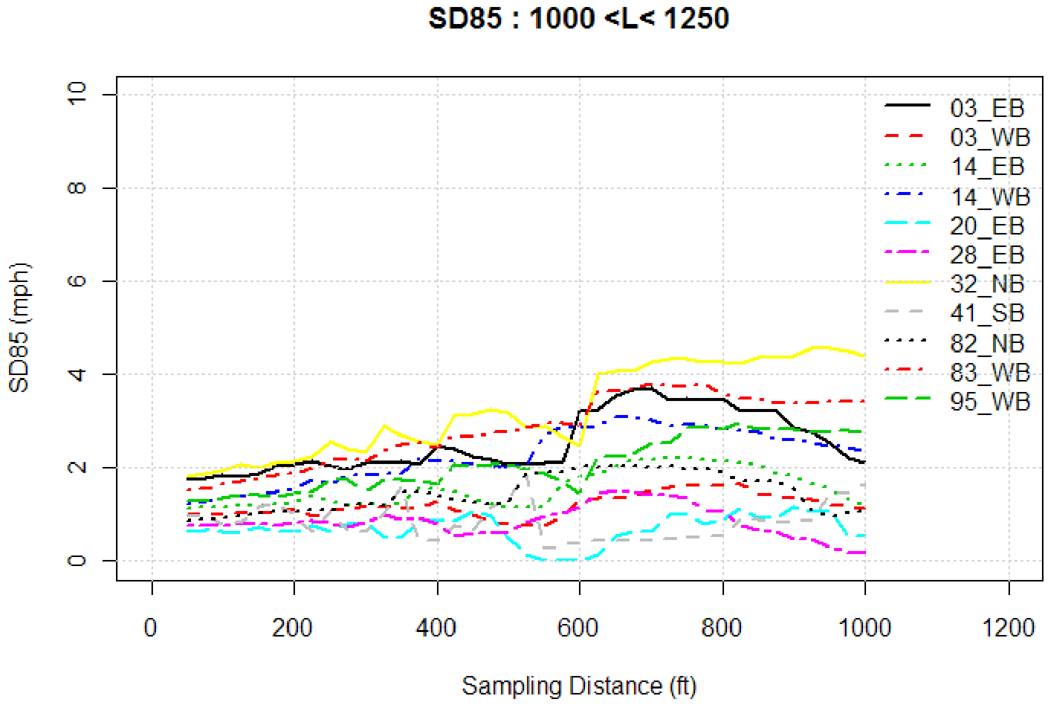
APPENDIX B: SENSITIVITY ANALYSIS RESULTS

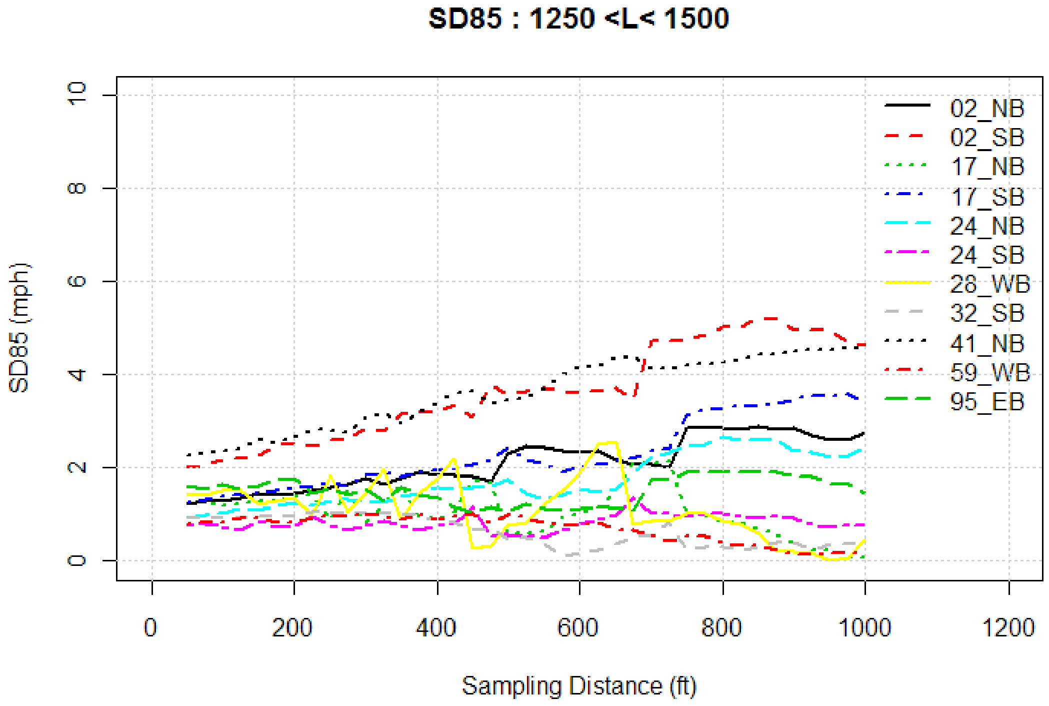
The charts in this appendix are the results of the sensitivity analyses of the speed measures to the spacing distance discussed in Chapter 5. The sensitivity plots are grouped by the effective corridor length, i.e., the corridor length subtracted by the traffic control influential zones.

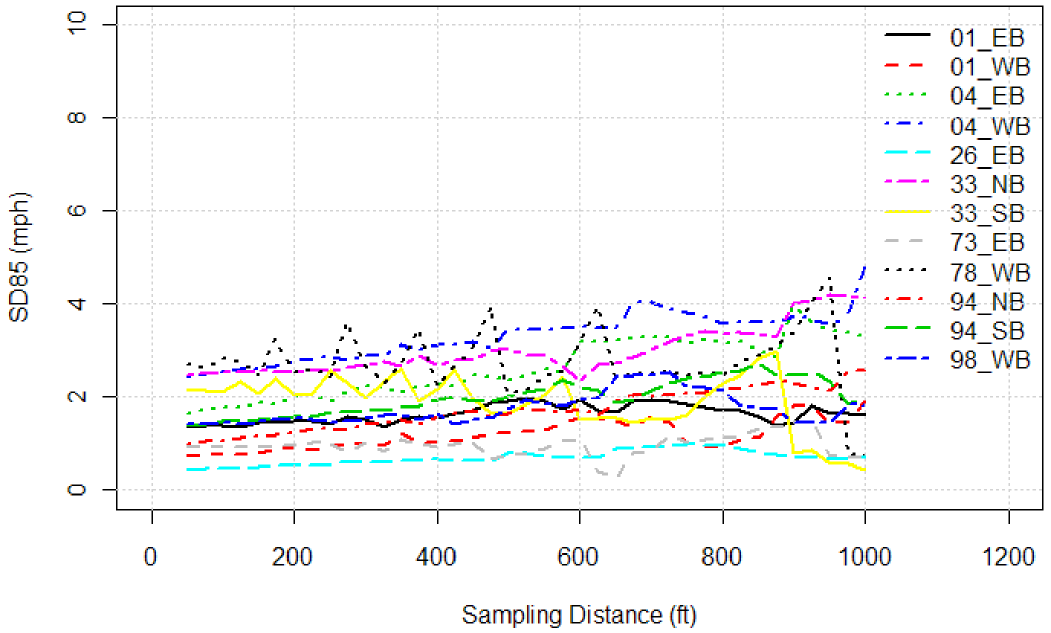
The speed measures exhibited in this section include:

- Speed variation (SD85)
- Mean of 85th percentile speed (M85)
- Coefficient of Variation of 85th percentile speed (CV85)
- Interquartile Range of 85th percentile speed (IQ85)
- Variation of 85th percentile speed from the speed limit (SVLIM)
- Mean of speed bands (M_BND)
- Variation of speed bands (SD_BND)
- Acceleration noise under free-flow condition (AN_FF)
- Acceleration noise under all-flow condition (AN_AF)

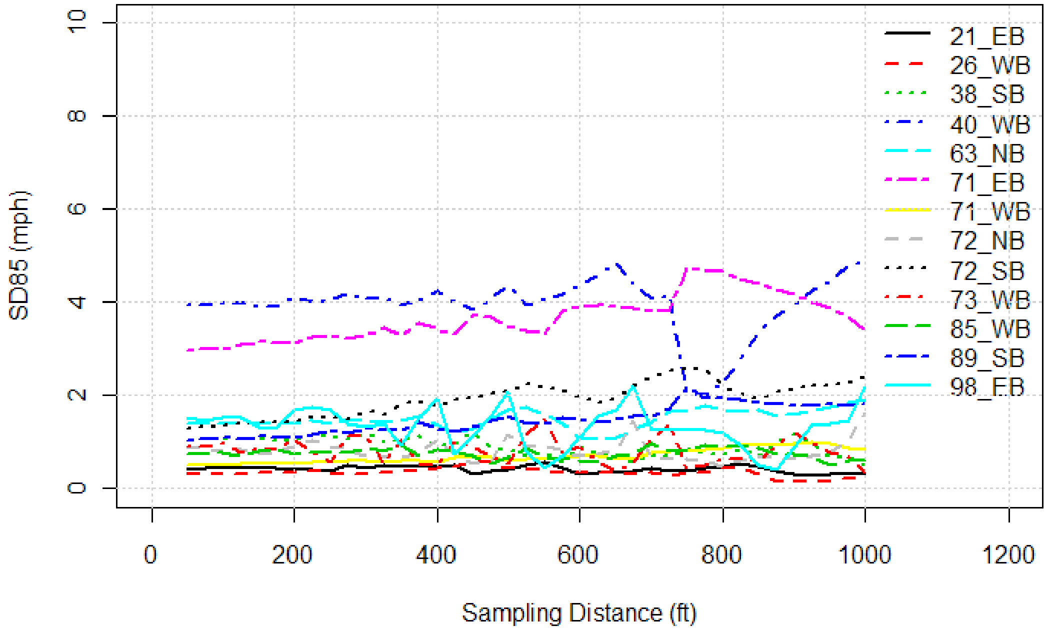
Sensitivity of Speed Variation (SD85) Plots



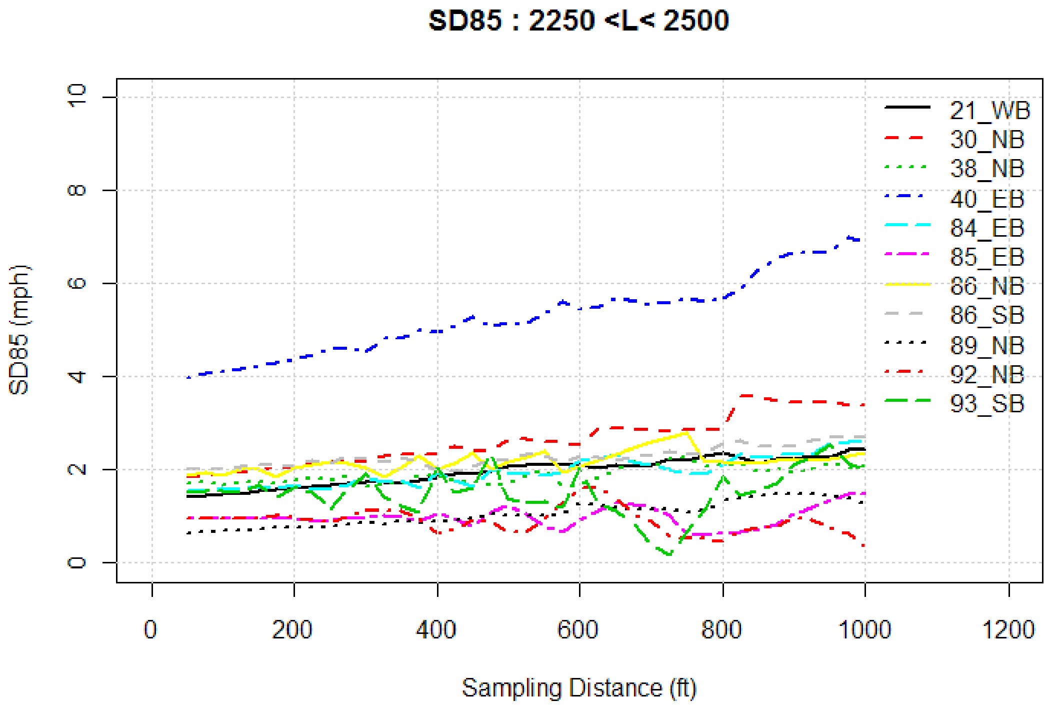


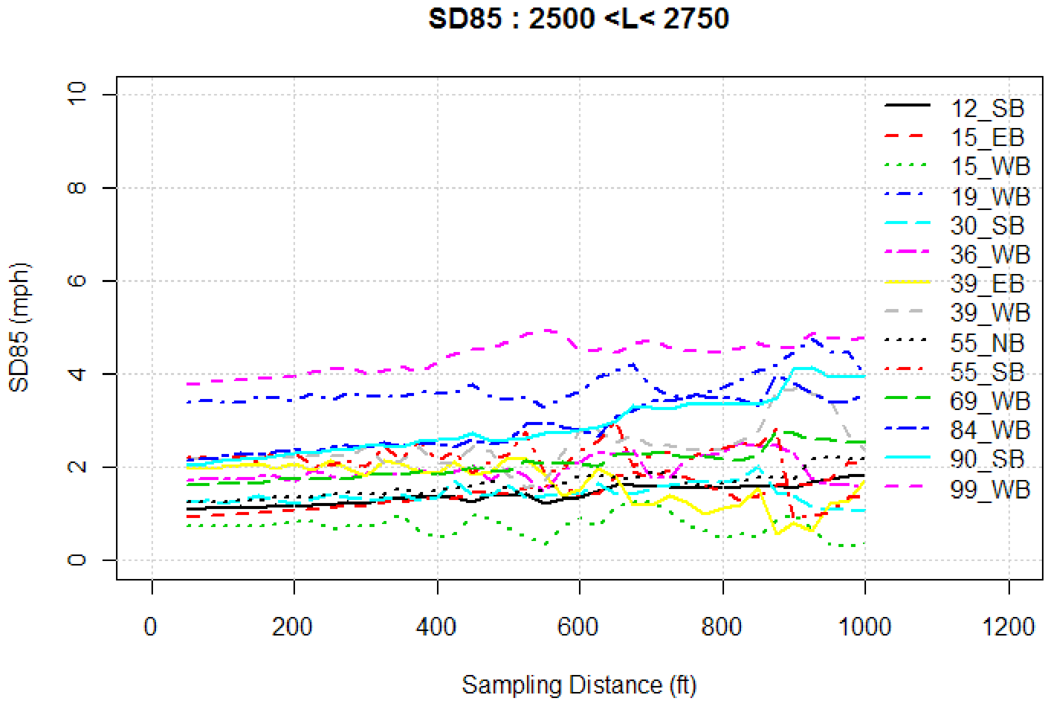


SD85 : 1500 <L< 2000

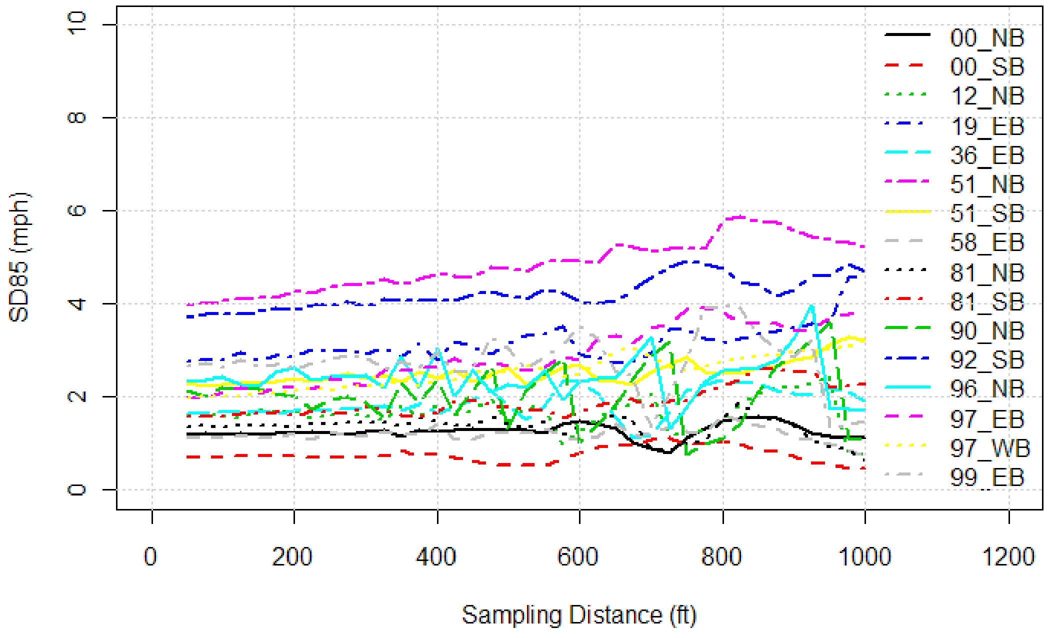


SD85 : 2000 <L< 2250

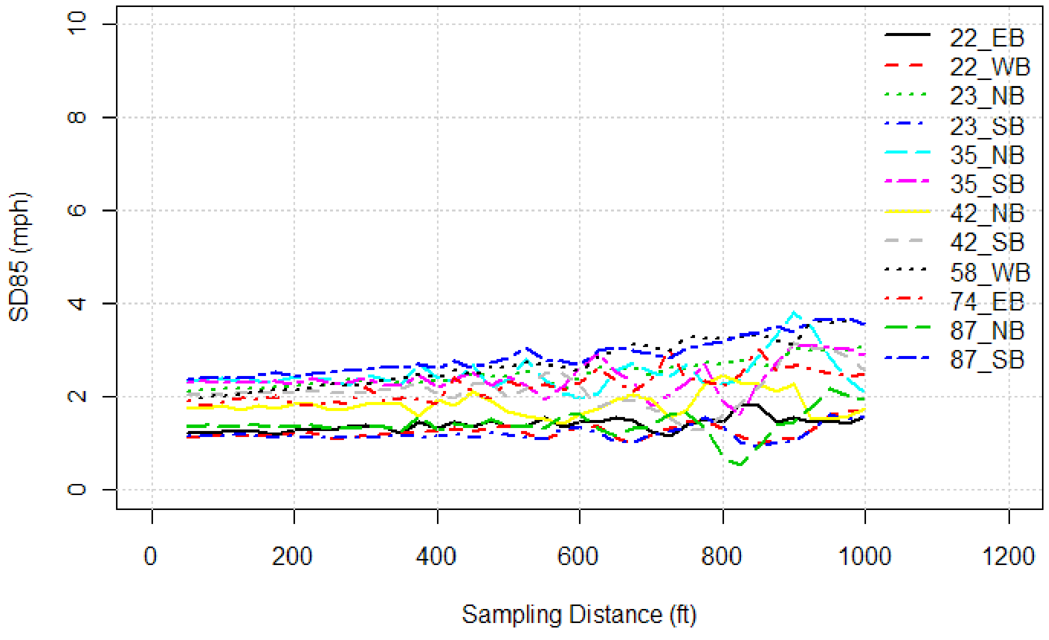

SD85 : 2250 <L< 2500
SD85 (mph)
0
2
4
6
8
10
0
200
400
600
800
1000
1200
Sampling Distance (ft)
21_WB
30_NB
38_NB
40_EB
84_EB
85_EB
86_NB
86_SB
89_NB
92_NB
93_SB


SD85 : 2500 <L< 2750
SD85 (mph)
0
2
4
6
8
10
0
200
400
600
800
1000
1200
Sampling Distance (ft)
12_SB
15_EB
15_WB
19_WB
30_SB
36_WB
39_EB
39_WB
55_NB
55_SB
69_WB
84_WB
90_SB
99_WB

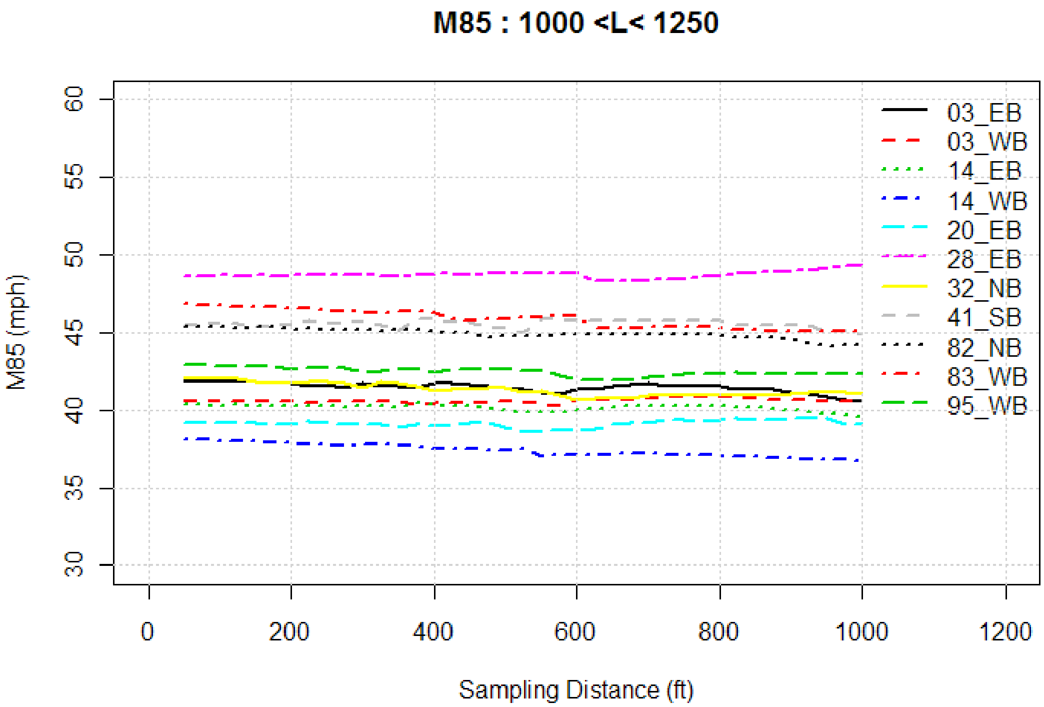
SD85 : 2750 <L< 3500

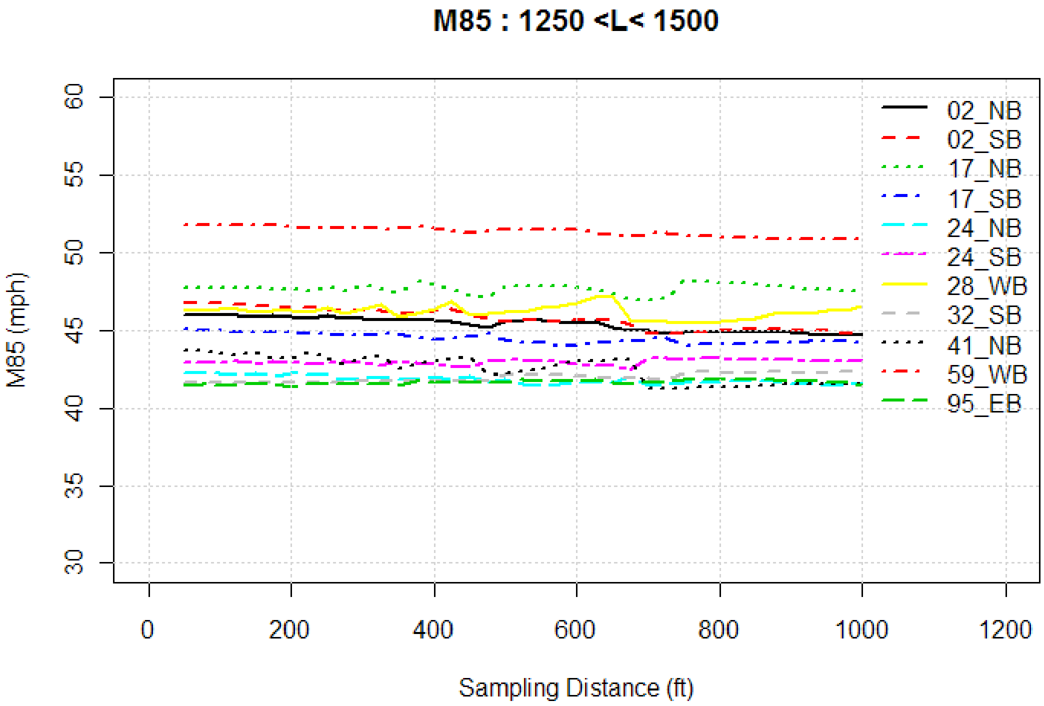


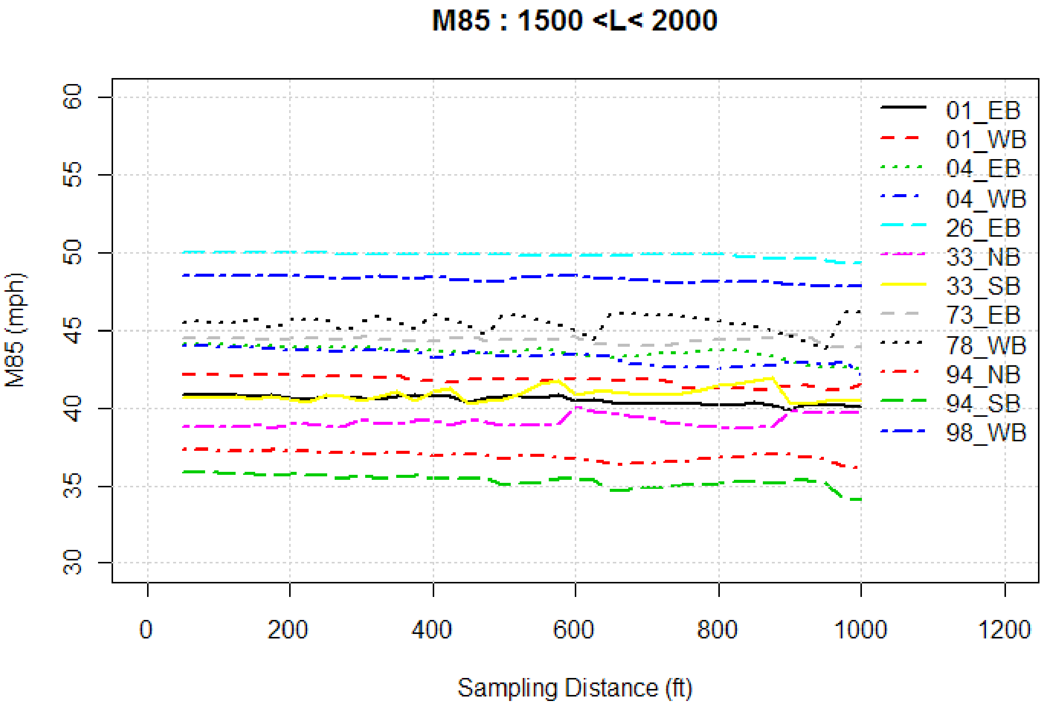
SD85 : 3500 <L< 7000

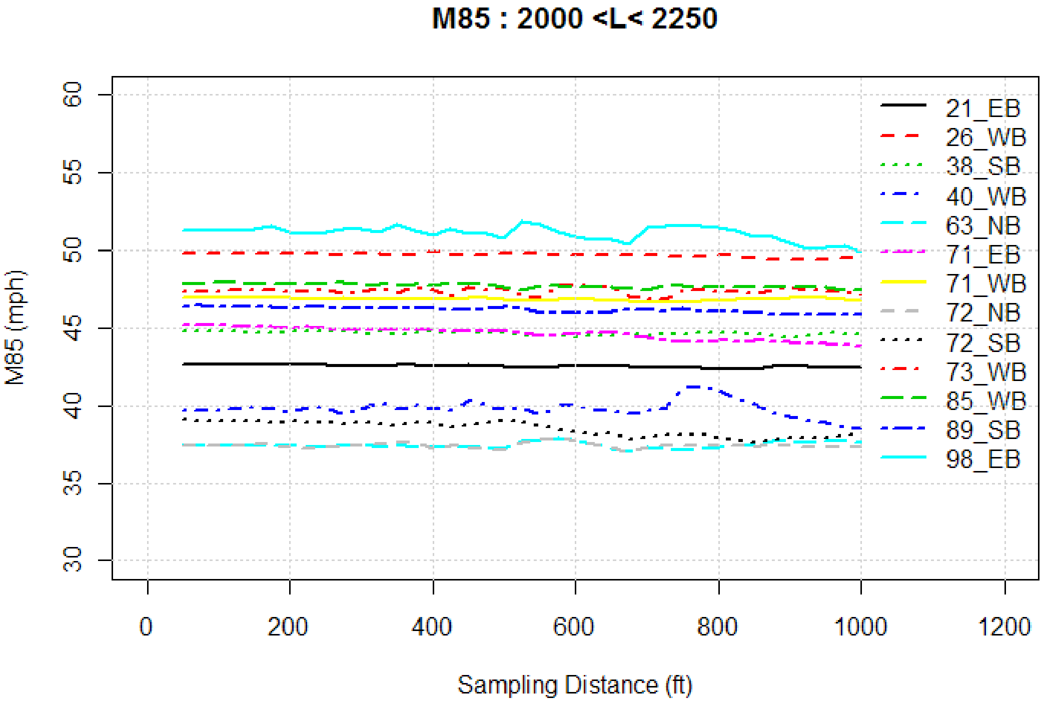


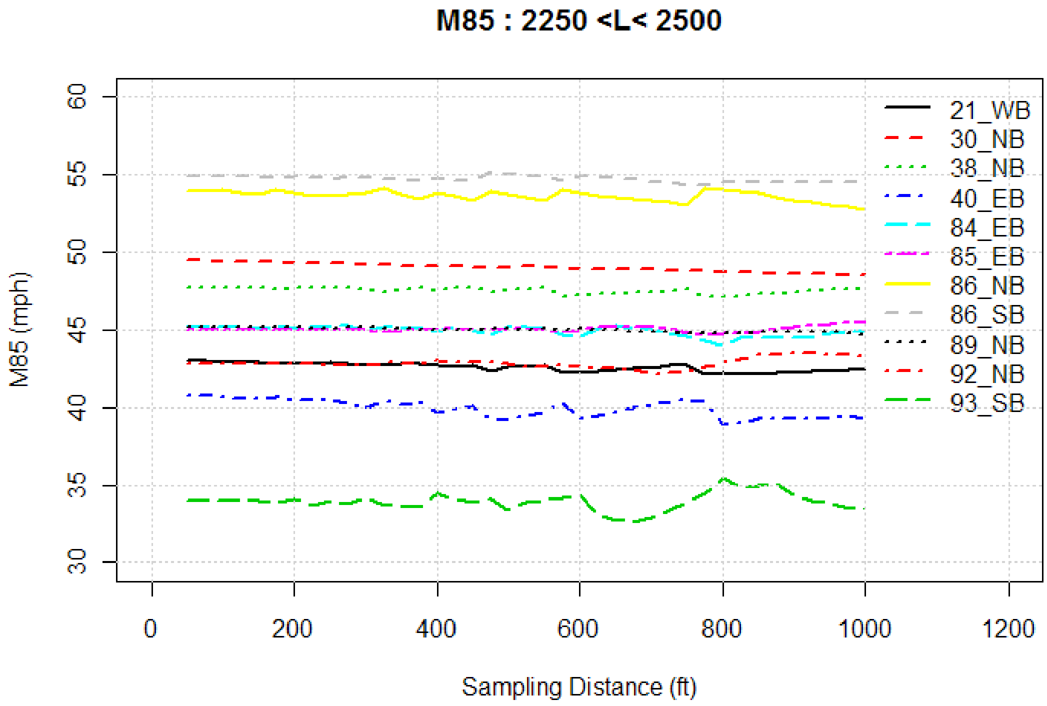
Sensitivity of Mean of 85th Percentile Speed (M85) Plots

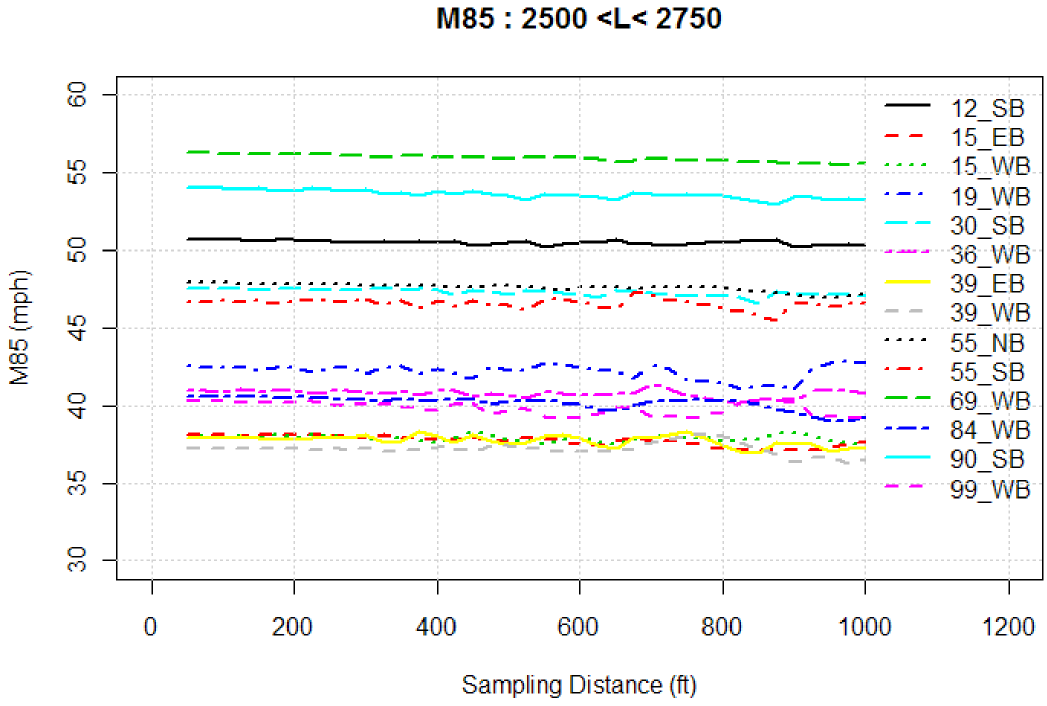




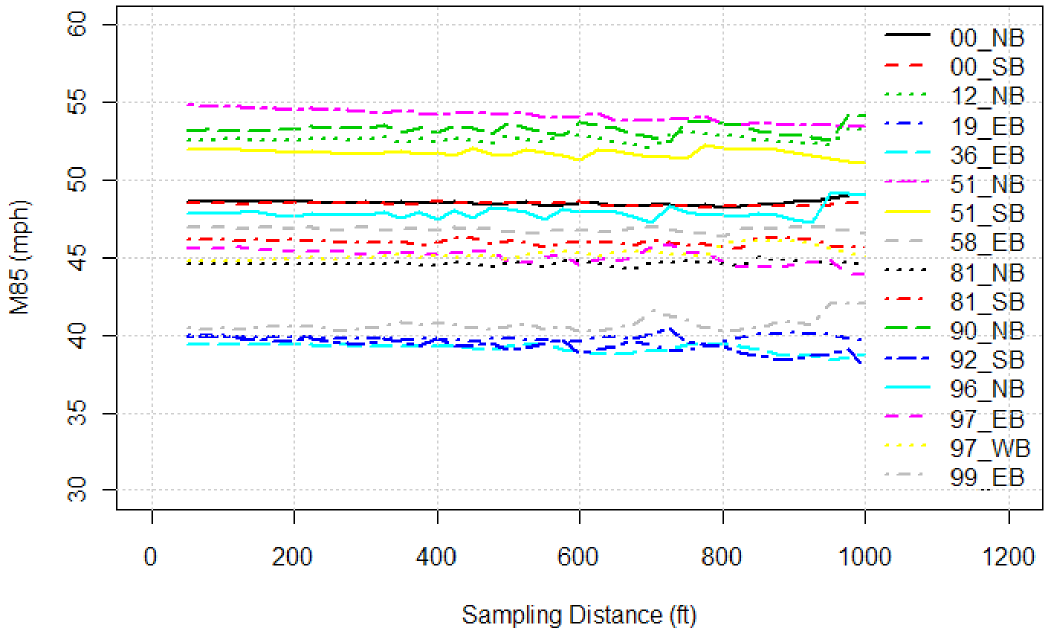

M85 : 1500 <L< 2000
M85 (mph)
30
35
40
45
50
55
60
0
200
400
600
800
1000
1200
Sampling Distance (ft)
01_EB
01_WB
04_EB
04_WB
26_EB
33_NB
33_SB
73_EB
78_WB
94_NB
94_SB
98_WB


M85 : 2000 <L< 2250
M85 (mph)
30
35
40
45
50
55
60
0
200
400
600
800
1000
1200
Sampling Distance (ft)
21_EB
26_WB
38_SB
40_WB
63_NB
71_EB
71_WB
72_NB
72_SB
73_WB
85_WB
89_SB
98_EB

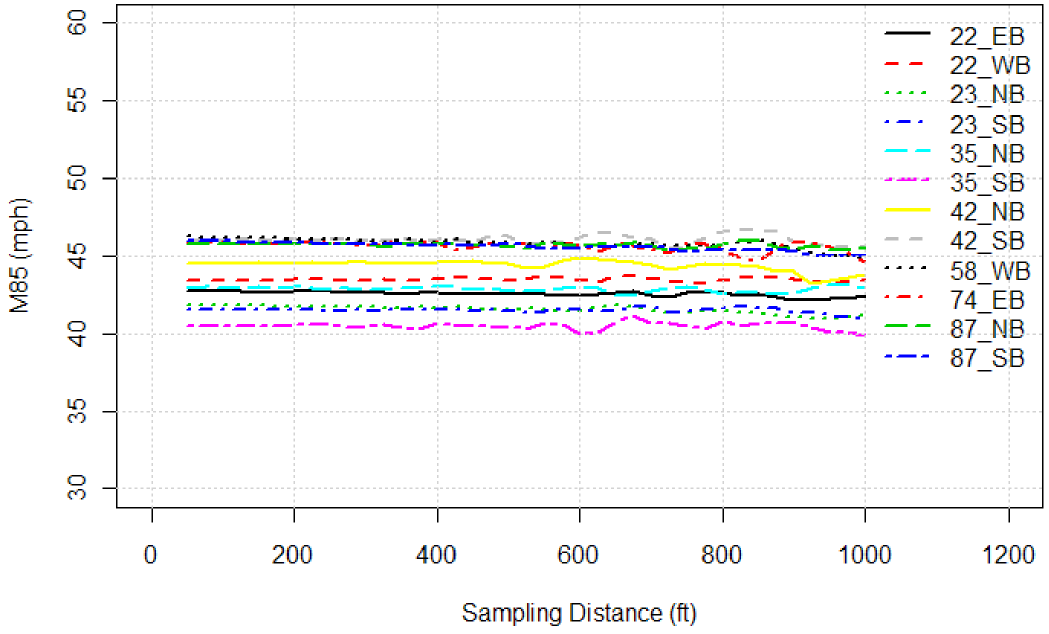

M85 : 2250 <L< 2500
M85 (mph)
60
55
50
45
40
35
30
0
200
400
600
800
1000
1200
Sampling Distance (ft)
21_WB
30_NB
38_NB
40_EB
84_EB
85_EB
86_NB
86_SB
89_NB
92_NB
93_SB


M85 : 2500 <L< 2750
M85 (mph)
60
55
50
45
40
35
30
0
200
400
600
800
1000
1200
Sampling Distance (ft)
12_SB
15_EB
15_WB
19_WB
30_SB
36_WB
39_EB
39_WB
55_NB
55_SB
69_WB
84_WB
90_SB
99_WB

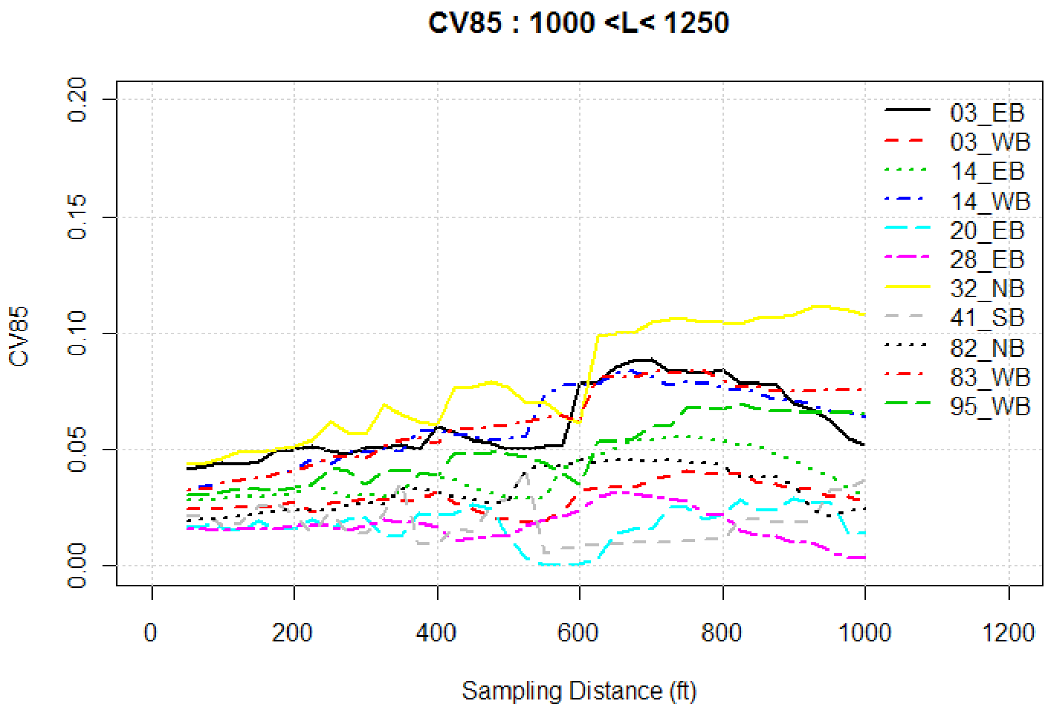
M85 : 2750 <L< 3500

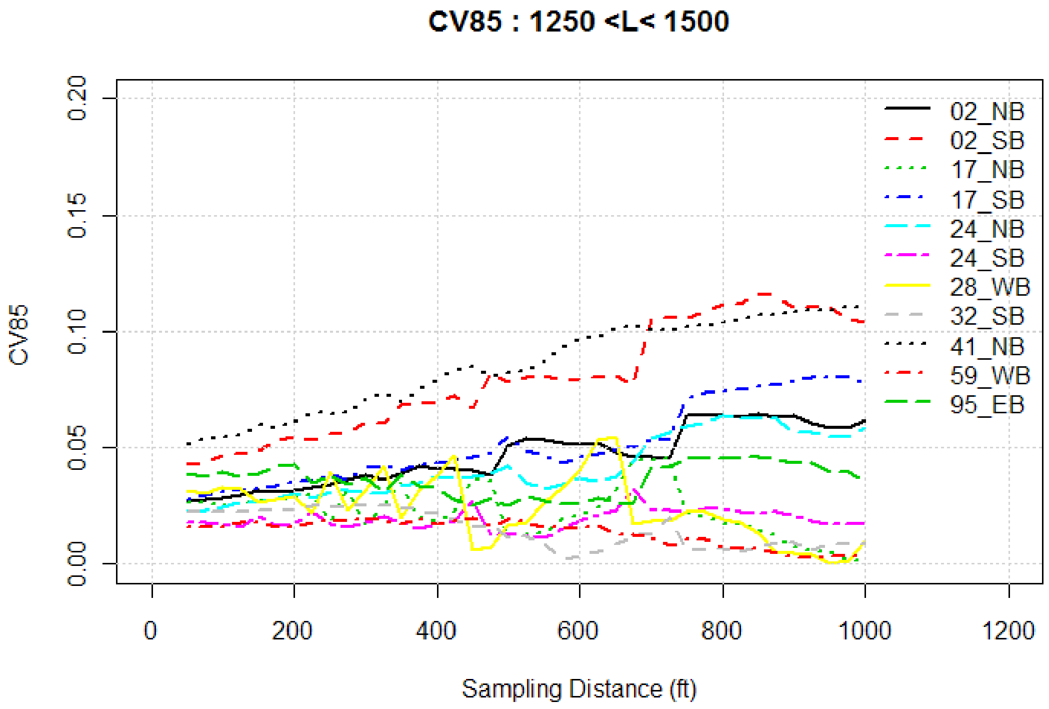


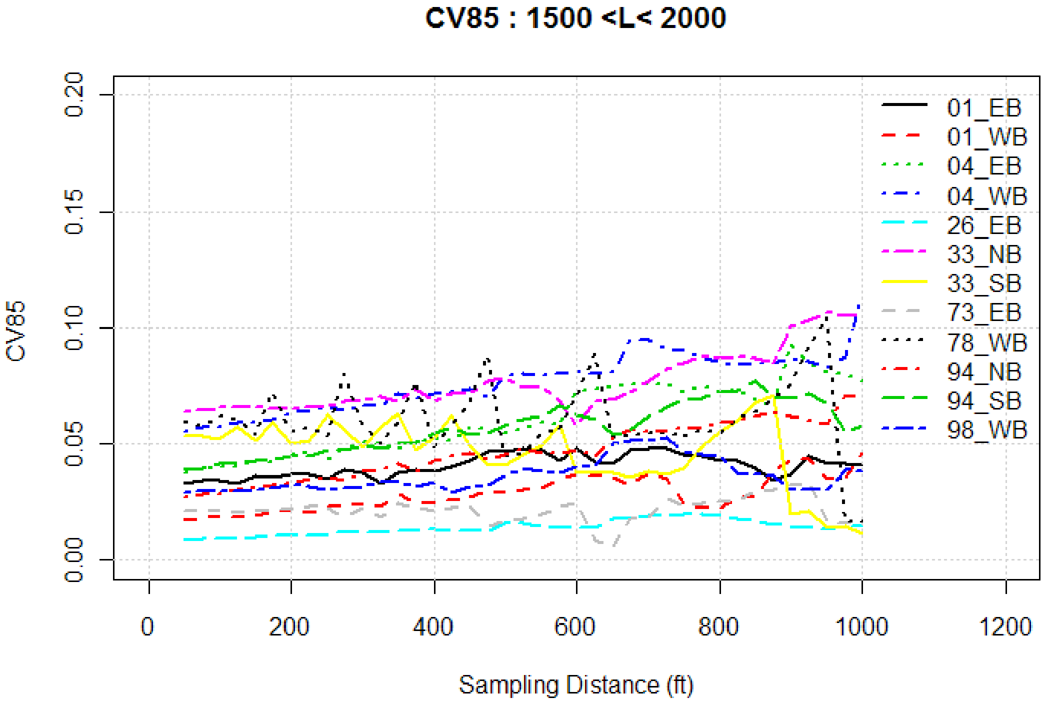
M85 : 3500 <L< 7000

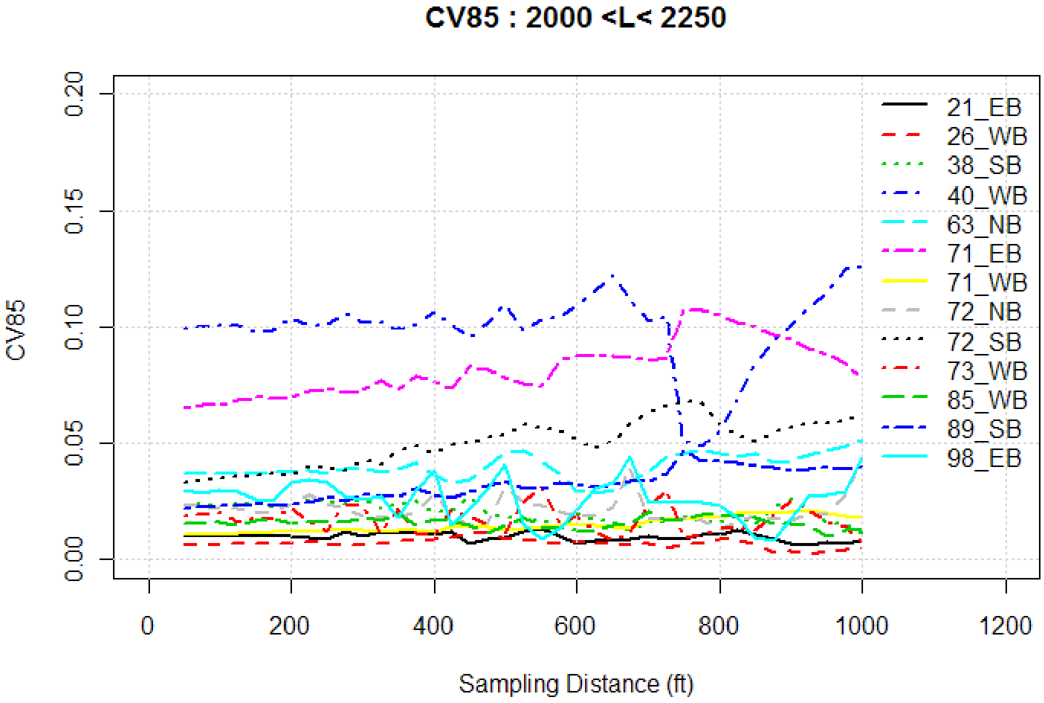


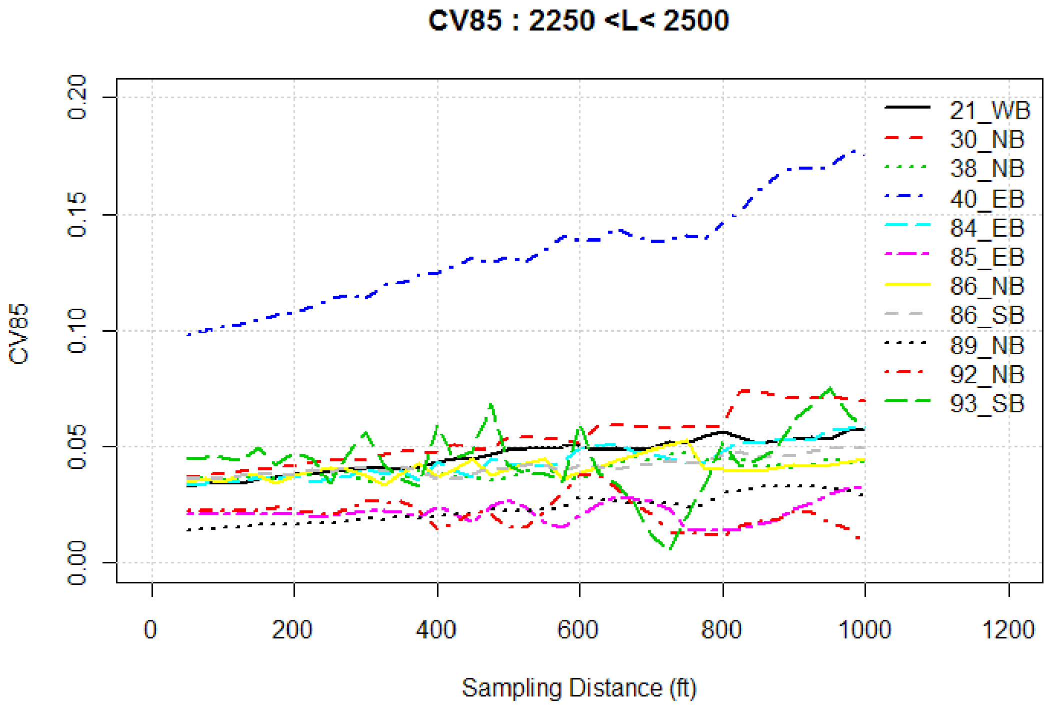
Sensitivity of Coefficient of Variation of 85th Percentile Speed (CV85) Plots

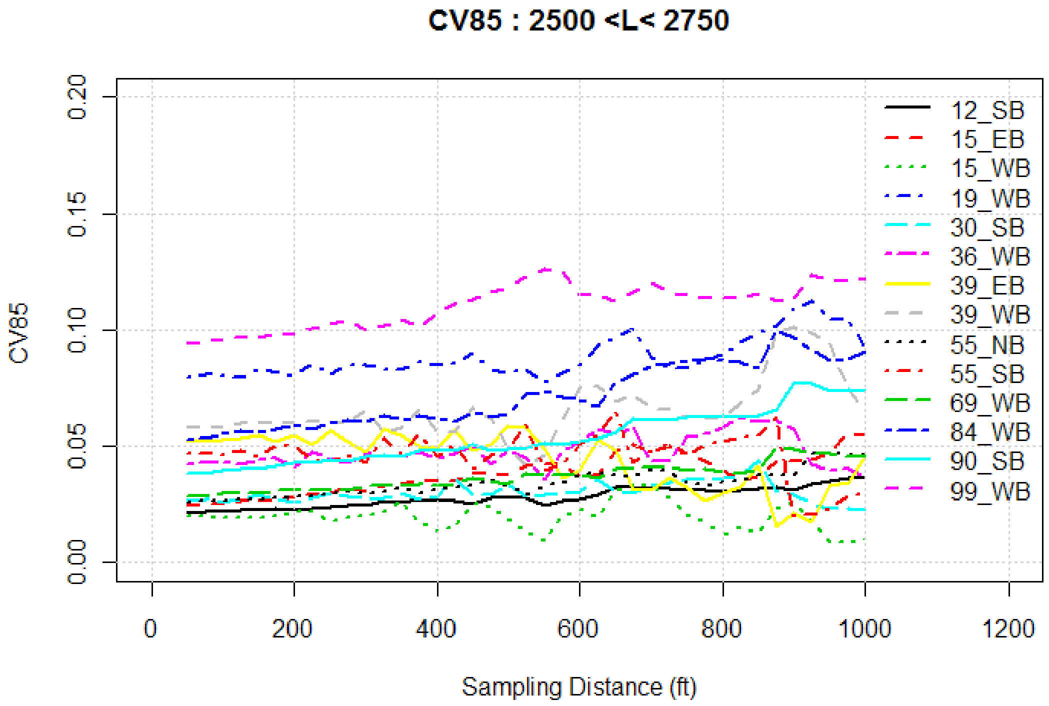


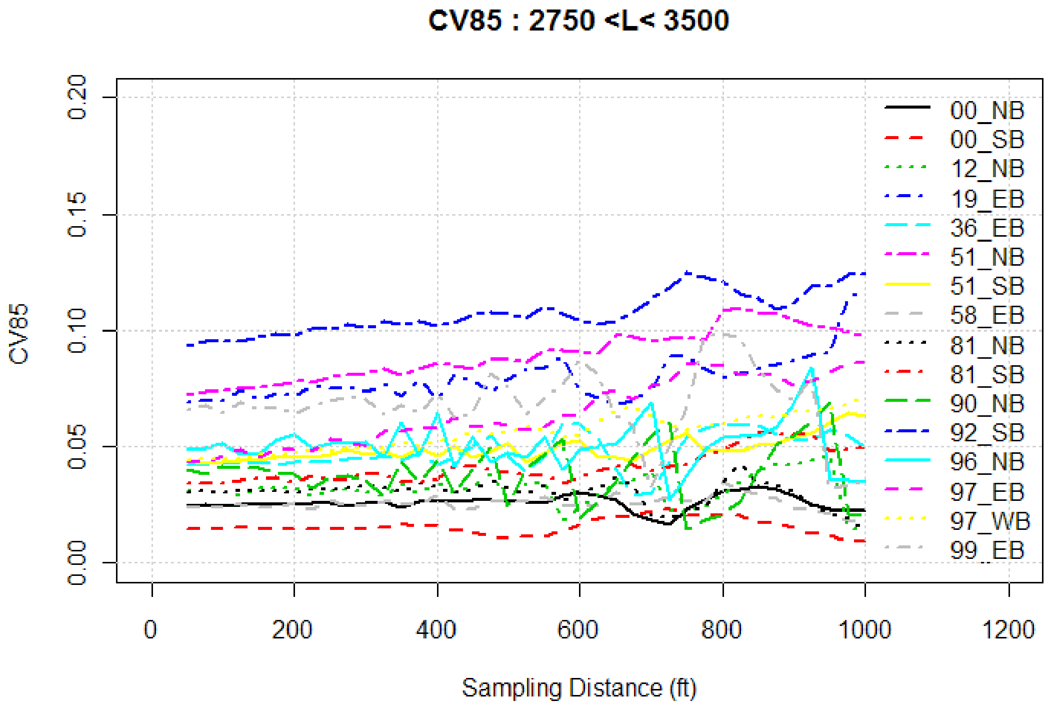


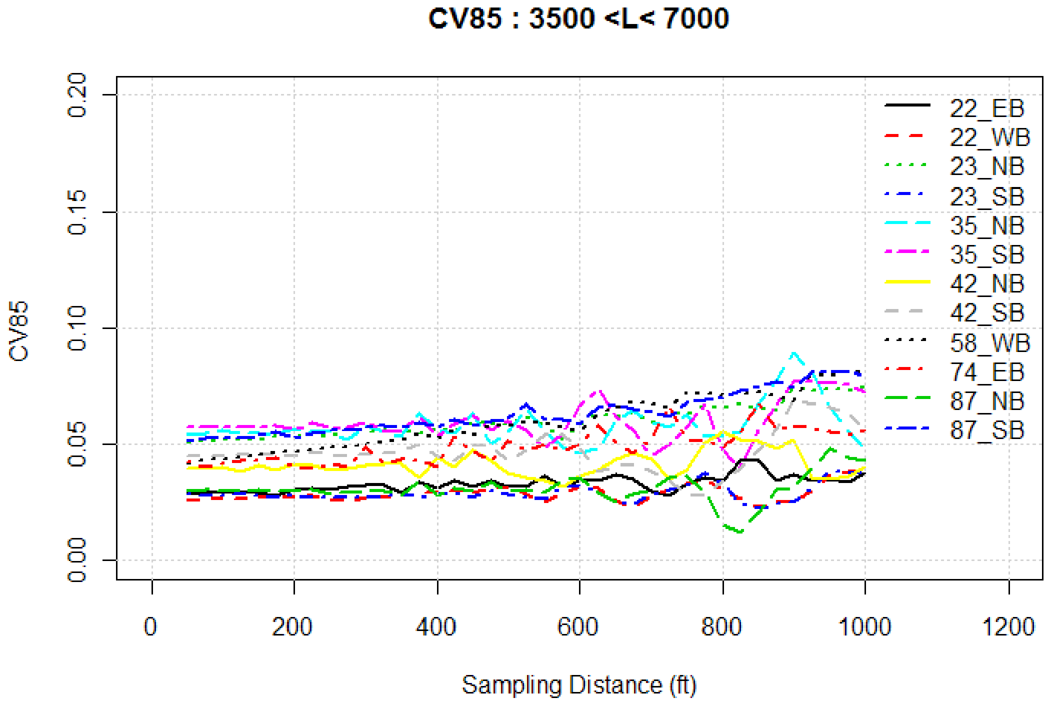

CV85 : 1500 <L< 2000
CV85
0.00
0.05
0.10
0.15
0.20
0
200
400
600
800
1000
1200
Sampling Distance (ft)
01_EB
01_WB
04_EB
04_WB
26_EB
33_NB
33_SB
73_EB
78_WB
94_NB
94_SB
98_WB


CV85 : 2000 <L< 2250
CV85
0.00
0.05
0.10
0.15
0.20
0
200
400
600
800
1000
1200
Sampling Distance (ft)
21_EB
26_WB
38_SB
40_WB
63_NB
71_EB
71_WB
72_NB
72_SB
73_WB
85_WB
89_SB
98_EB

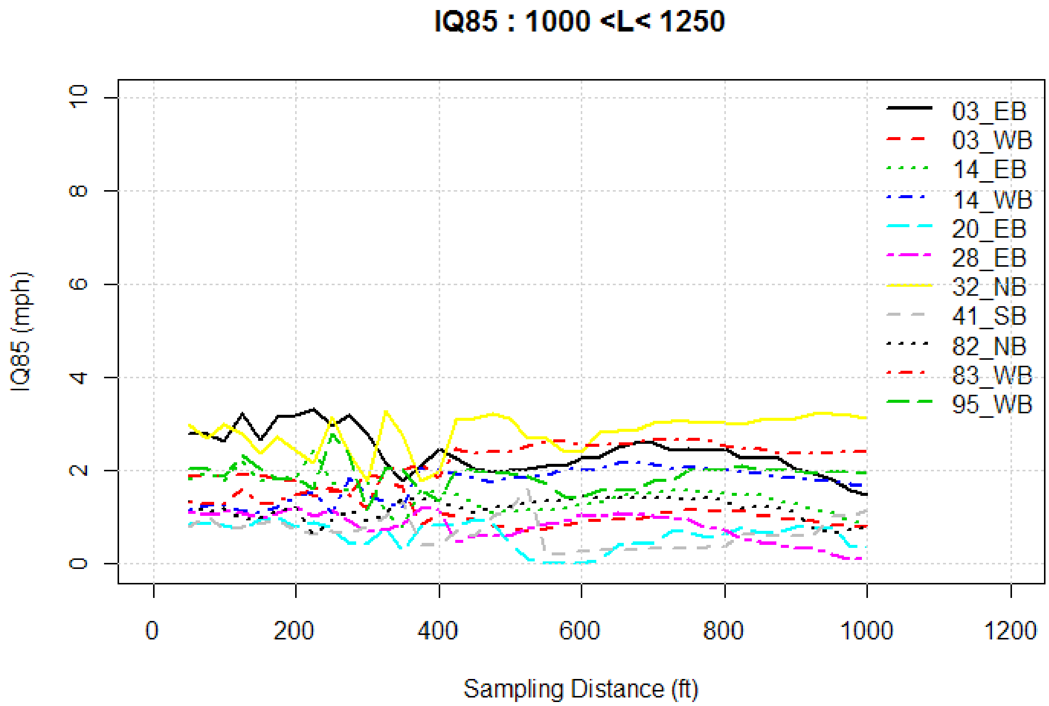

CV85 : 2250 <L< 2500
CV85
0.00
0.05
0.10
0.15
0.20
0
200
400
600
800
1000
1200
Sampling Distance (ft)
21_WB
30_NB
38_NB
40_EB
84_EB
85_EB
86_NB
86_SB
89_NB
92_NB
93_SB

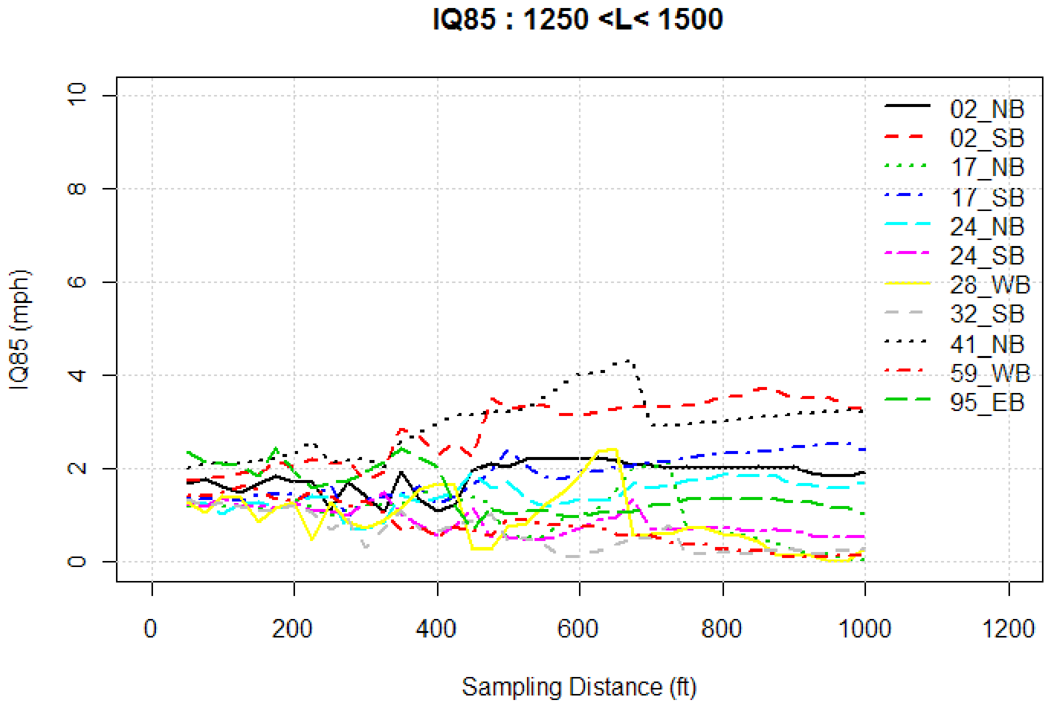

CV85 : 2500 <L< 2750
CV85
0.00
0.05
0.10
0.15
0.20
0
200
400
600
800
1000
1200
Sampling Distance (ft)
12_SB
15_EB
15_WB
19_WB
30_SB
36_WB
39_EB
39_WB
55_NB
55_SB
69_WB
84_WB
90_SB
99_WB


CV85 : 2750 <L< 3500
CV85
0.00
0.05
0.10
0.15
0.20
0
200
400
600
800
1000
1200
Sampling Distance (ft)
00_NB
00_SB
12_NB
19_EB
36_EB
51_NB
51_SB
58_EB
81_NB
81_SB
90_NB
92_SB
96_NB
97_EB
97_WB
99_EB

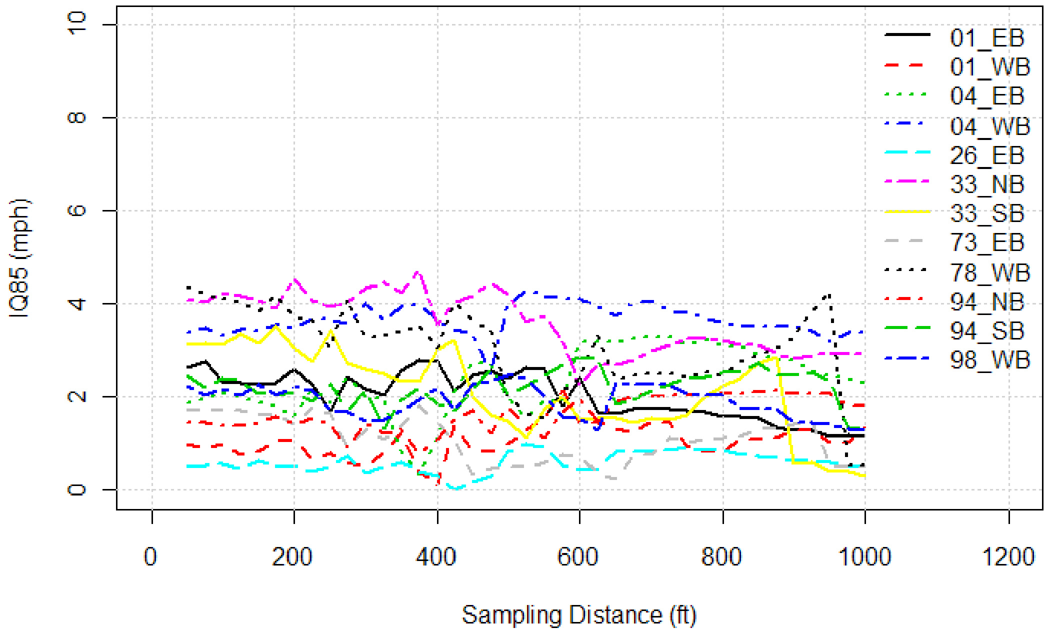

CV85 : 3500 <L< 7000
CV85
0.00
0.05
0.10
0.15
0.20
0
200
400
600
800
1000
1200
Sampling Distance (ft)
22_EB
22_WB
23_NB
23_SB
35_NB
35_SB
42_NB
42_SB
58_WB
74_EB
87_NB
87_SB

Sensitivity of Interquartile Range of 85th Percentile Speed (IQ85) Plots

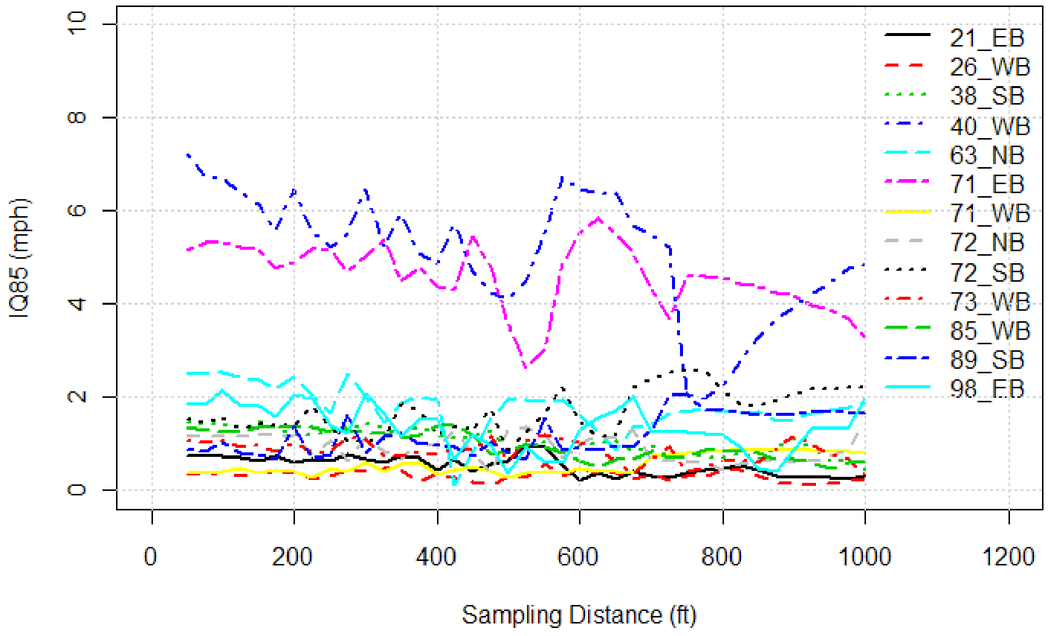


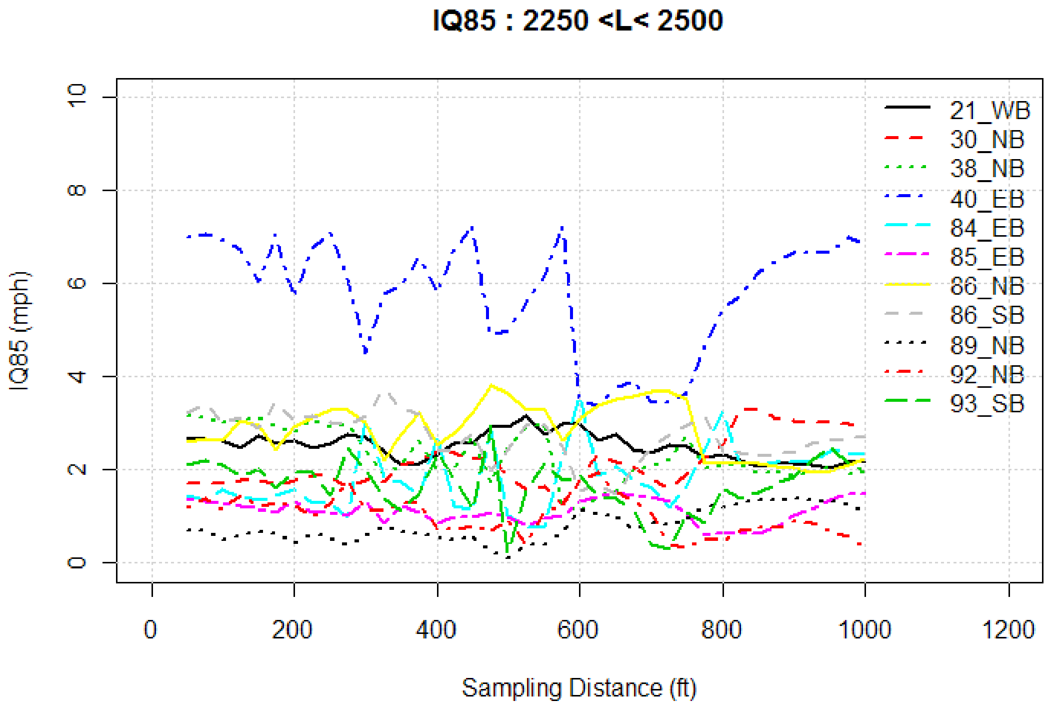


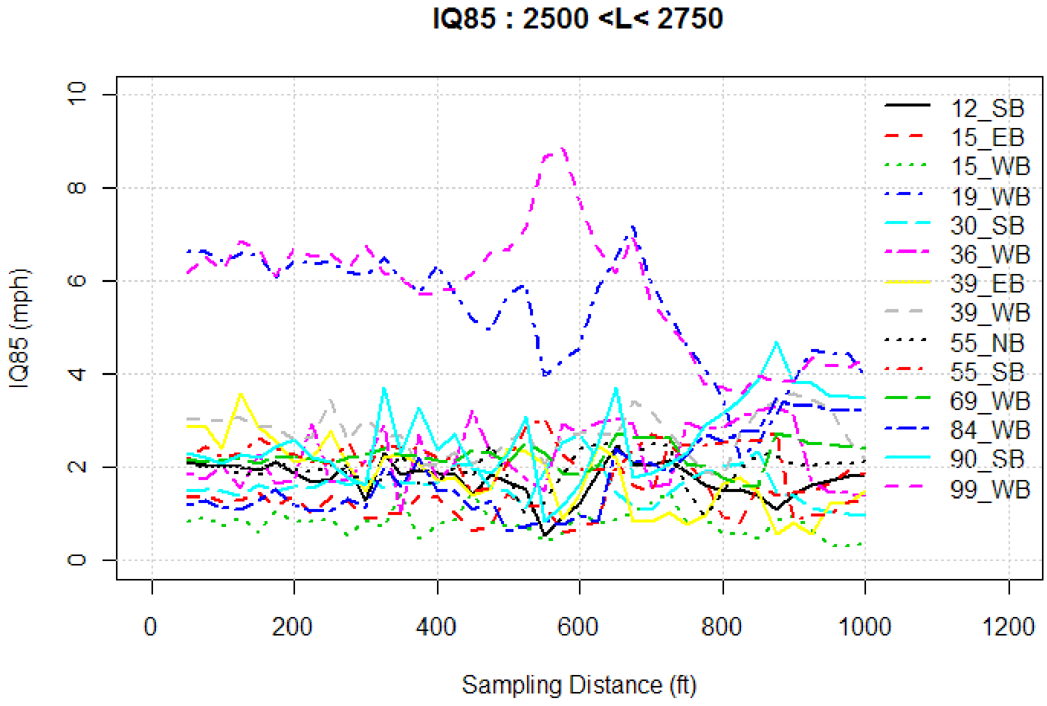
IQ85 : 1500 <L< 2000

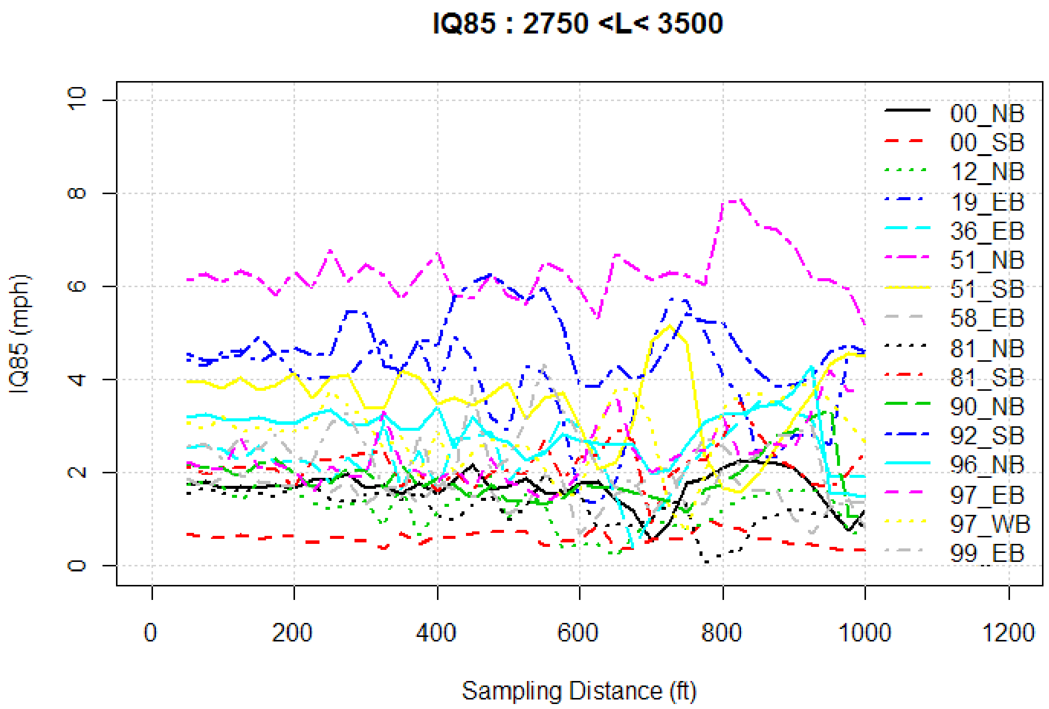


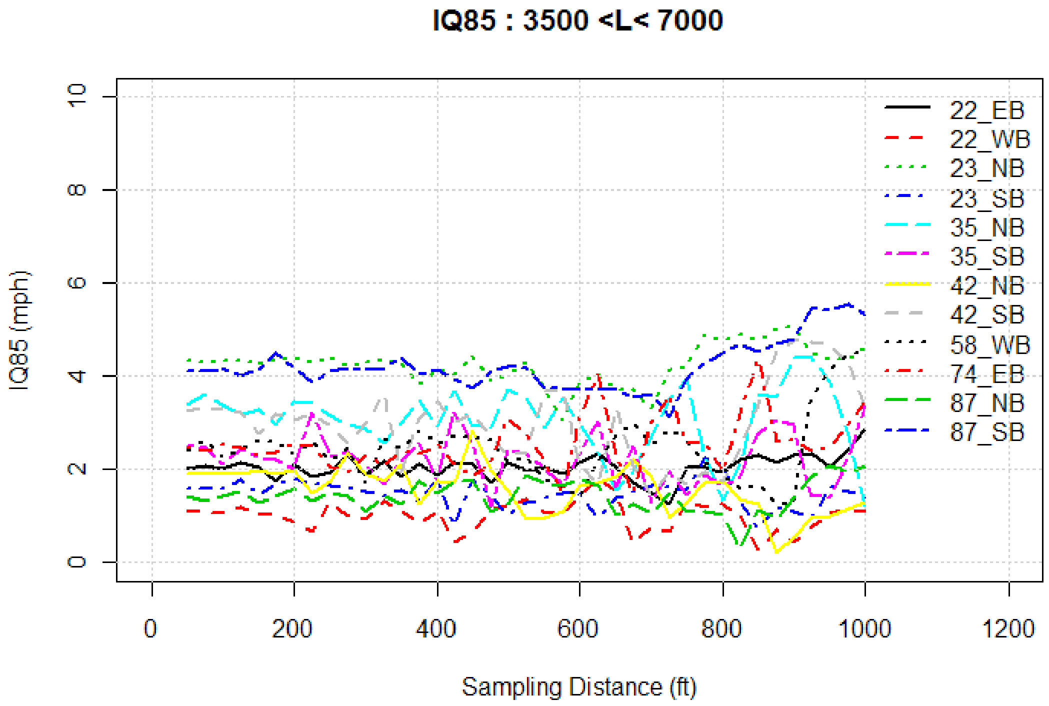
IQ85 : 2000 <L< 2250



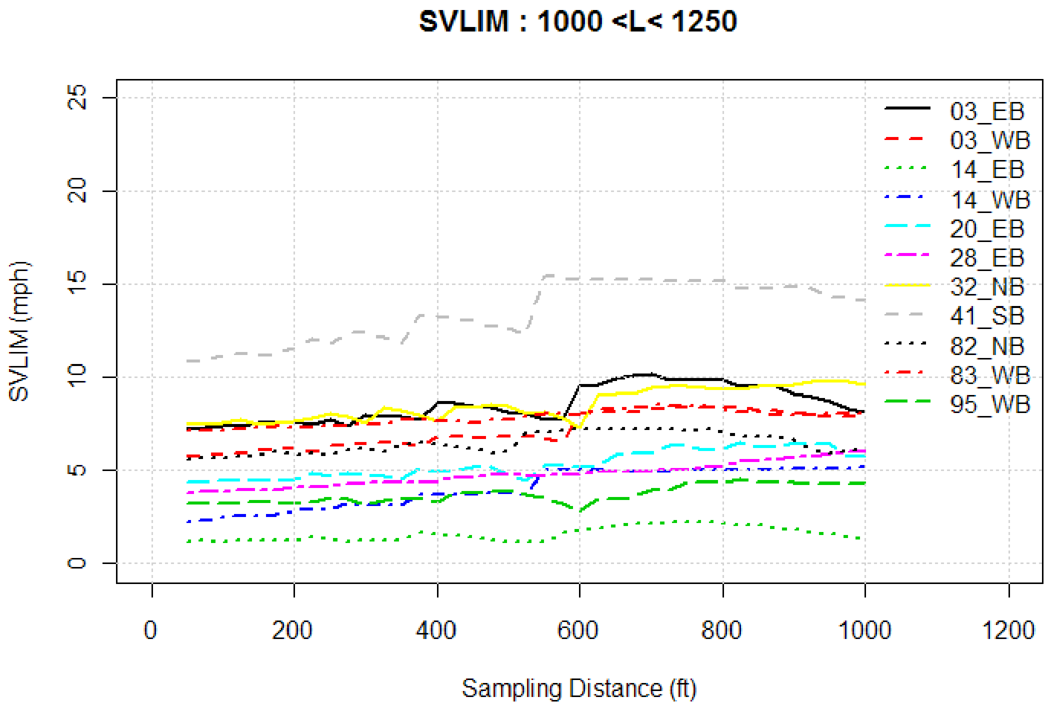

IQ85 : 2250 <L< 2500
IQ85 (mph)
0
2
4
6
8
10
0
200
400
600
800
1000
1200
Sampling Distance (ft)
21_WB
30_NB
38_NB
40_EB
84_EB
85_EB
86_NB
86_SB
89_NB
92_NB
93_SB

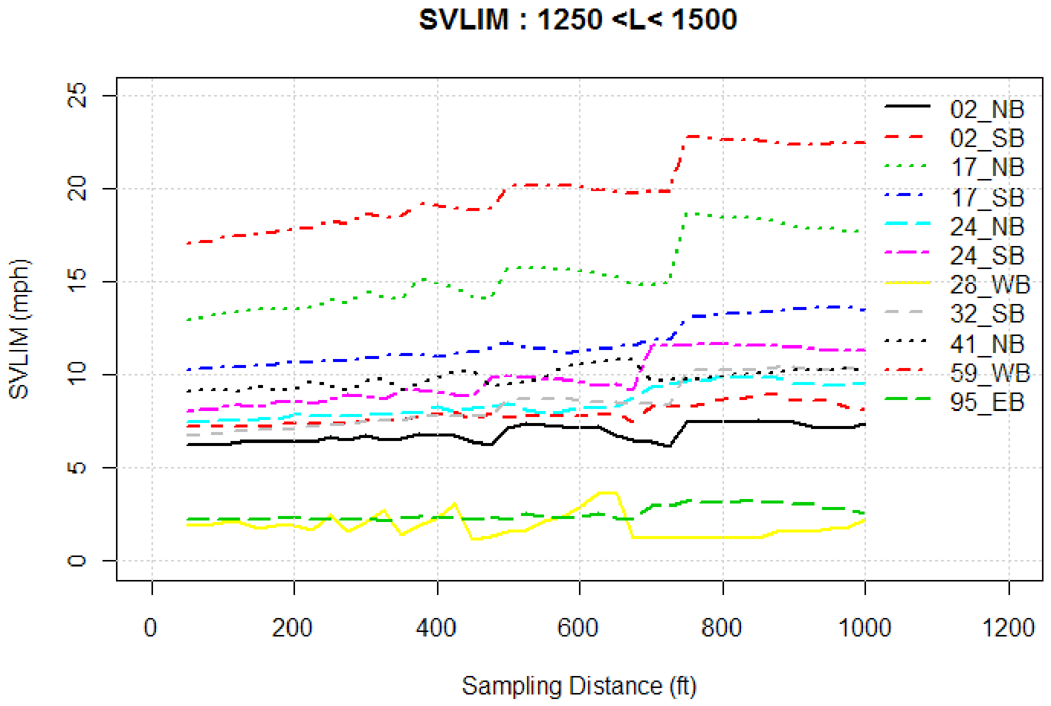

IQ85 : 2500 <L< 2750
IQ85 (mph)
0
2
4
6
8
10
0
200
400
600
800
1000
1200
Sampling Distance (ft)
12_SB
15_EB
15_WB
19_WB
30_SB
36_WB
39_EB
39_WB
55_NB
55_SB
69_WB
84_WB
90_SB
99_WB


IQ85 : 2750 <L< 3500
IQ85 (mph)
0
2
4
6
8
10
0
200
400
600
800
1000
1200
Sampling Distance (ft)
00_NB
00_SB
12_NB
19_EB
36_EB
51_NB
51_SB
58_EB
81_NB
81_SB
90_NB
92_SB
96_NB
97_EB
97_WB
99_EB

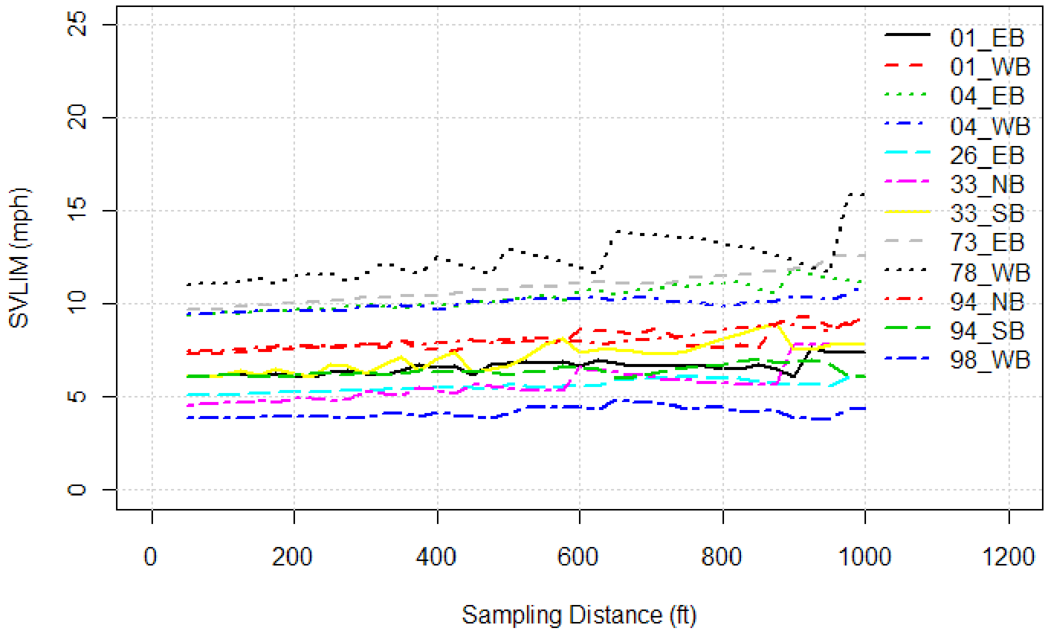

IQ85 : 3500 <L< 7000
IQ85 (mph)
0
2
4
6
8
10
0
200
400
600
800
1000
1200
Sampling Distance (ft)
22_EB
22_WB
23_NB
23_SB
35_NB
35_SB
42_NB
42_SB
58_WB
74_EB
87_NB
87_SB

Sensitivity of Variation of 85th Percentile Speed from the Speed Limit (SVLIM) Plots

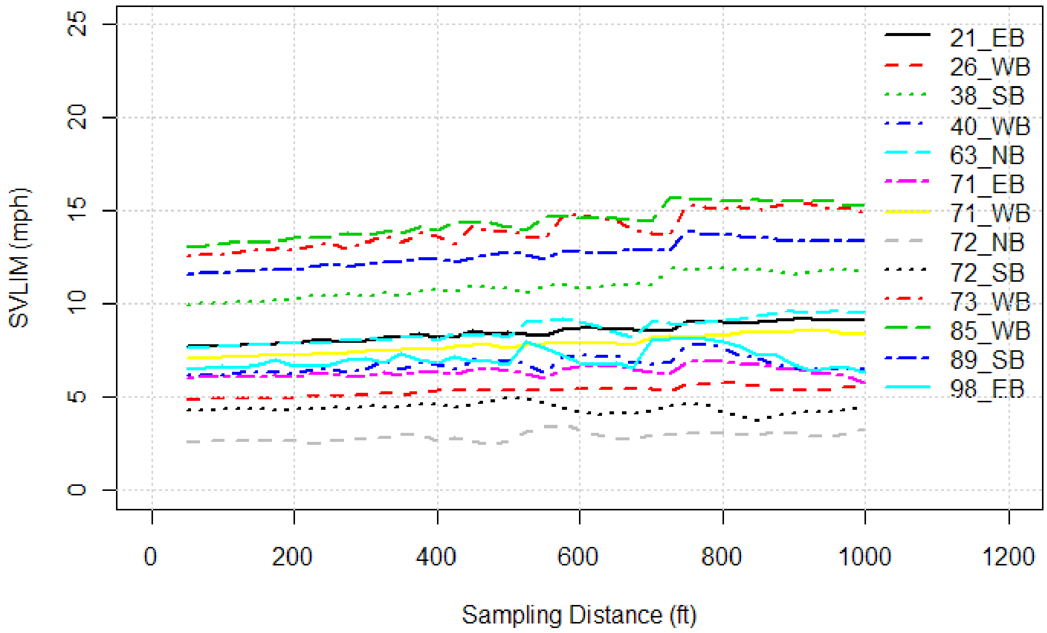




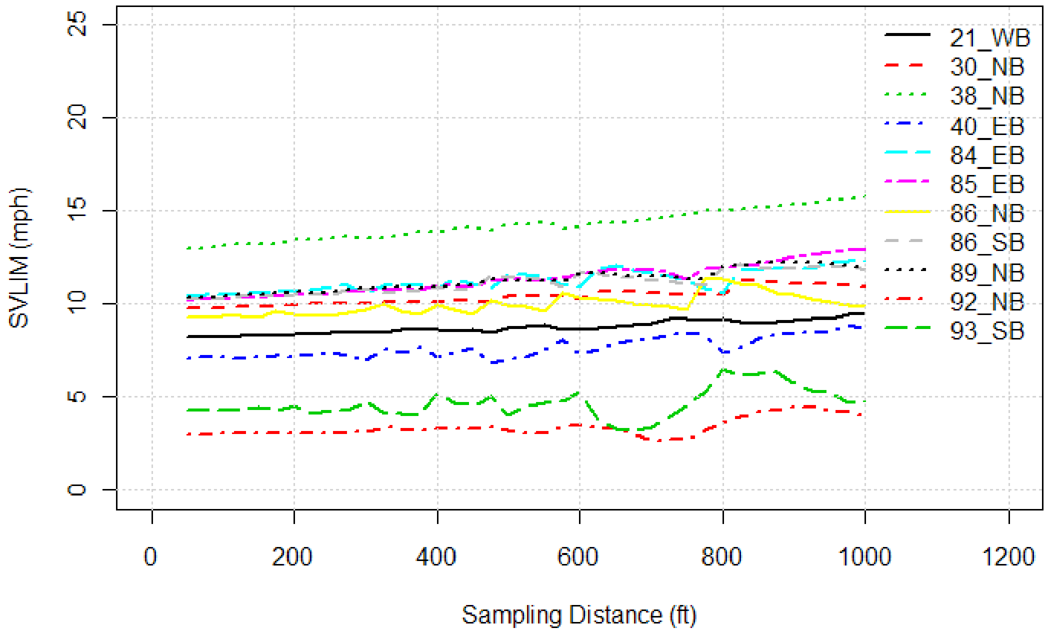
SVLIM : 1500 <L< 2000



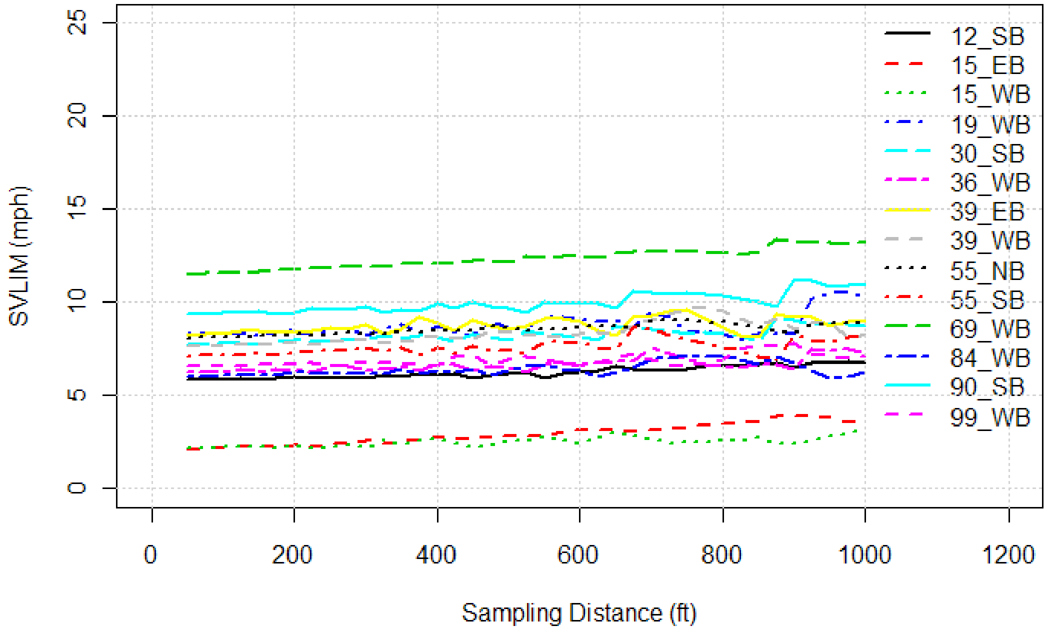
SVLIM : 2000 <L< 2250

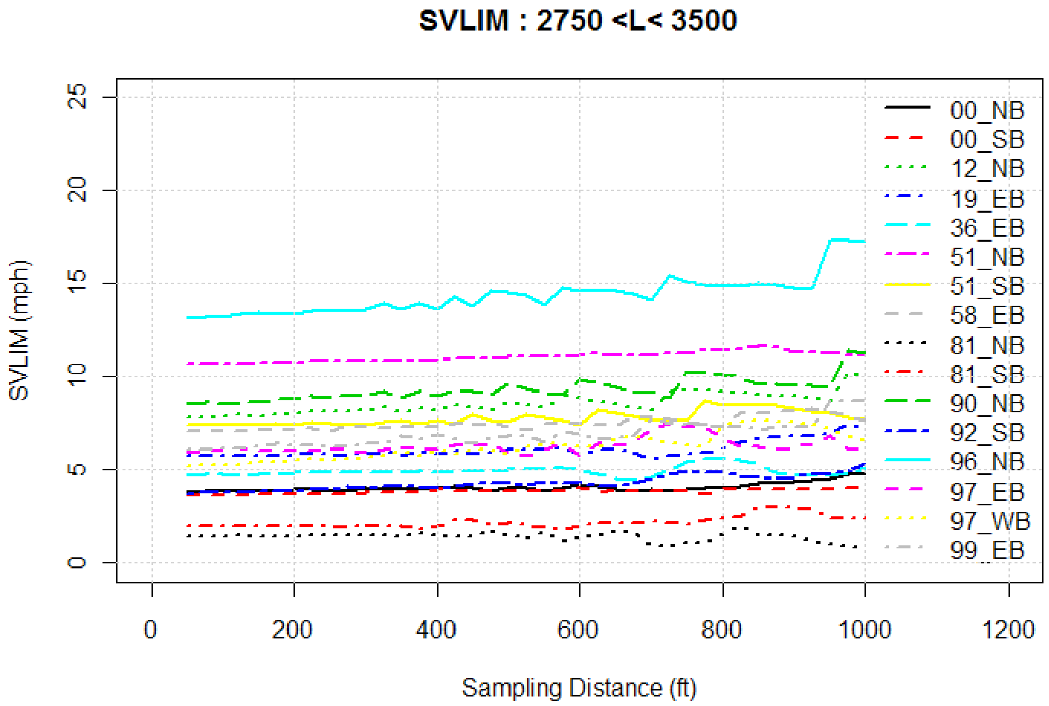


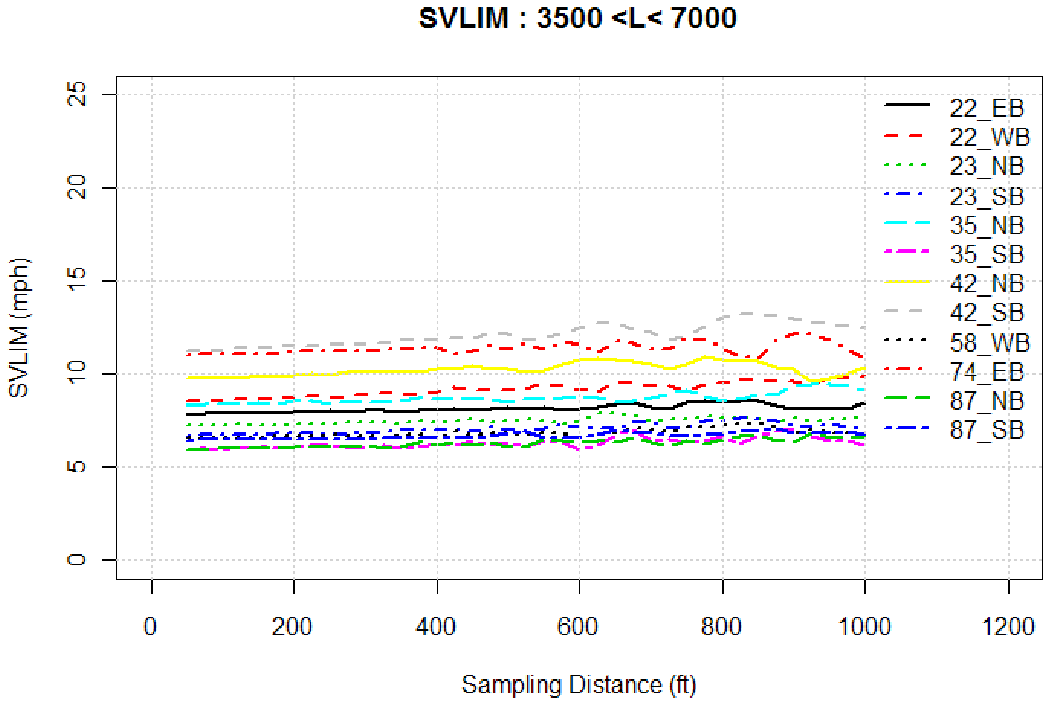
SVLIM : 2250 <L< 2500



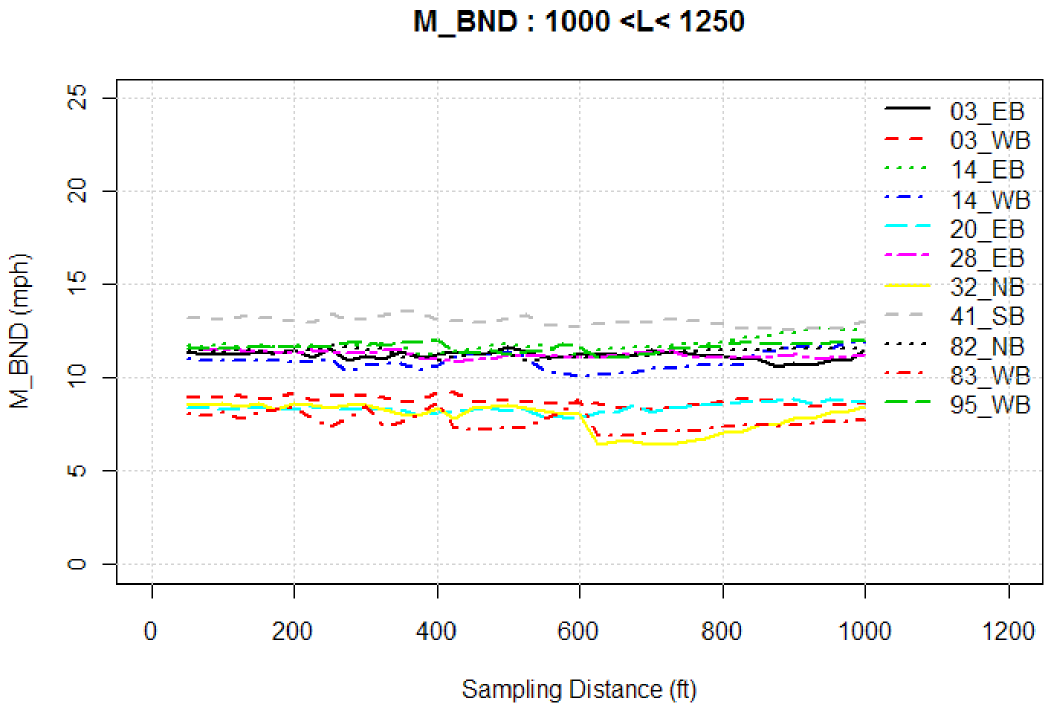
SVLIM : 2500 <L< 2750

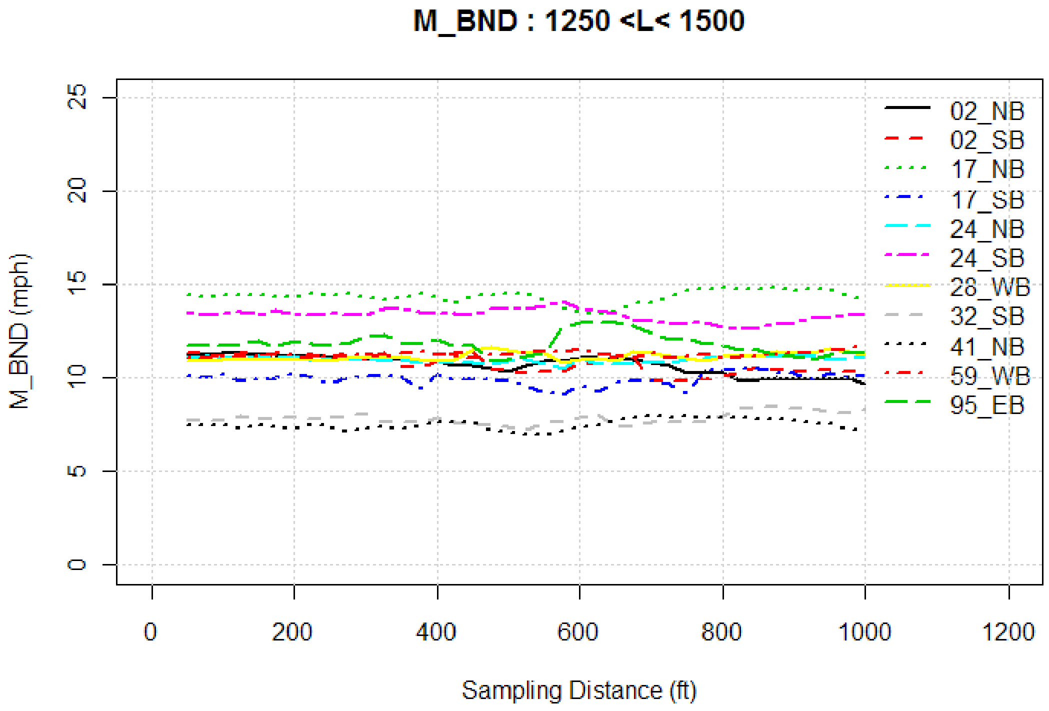



SVLIM : 2750 <L< 3500
SVLIM (mph)
0
5
10
15
20
25
0
200
400
600
800
1000
1200
Sampling Distance (ft)
00_NB
00_SB
12_NB
19_EB
36_EB
51_NB
51_SB
58_EB
81_NB
81_SB
90_NB
92_SB
96_NB
97_EB
97_WB
99_EB

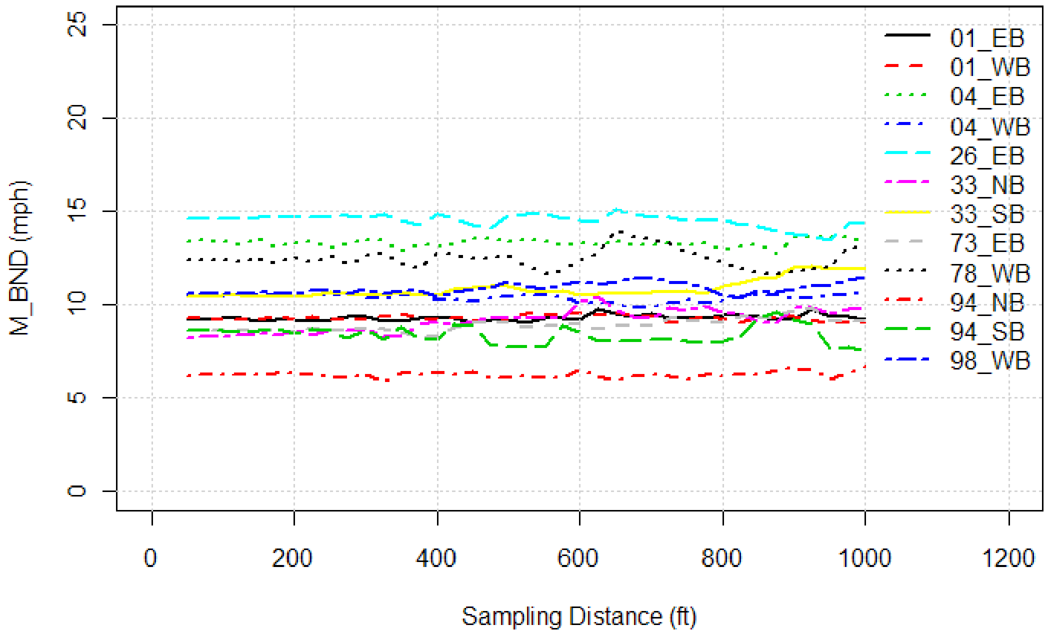

SVLIM : 3500 <L< 7000
SVLIM (mph)
0
5
10
15
20
25
0
200
400
600
800
1000
1200
Sampling Distance (ft)
22_EB
22_WB
23_NB
23_SB
35_NB
35_SB
42_NB
42_SB
58_WB
74_EB
87_NB
87_SB

Sensitivity of Mean of Speed Bands (M_BND) Plots

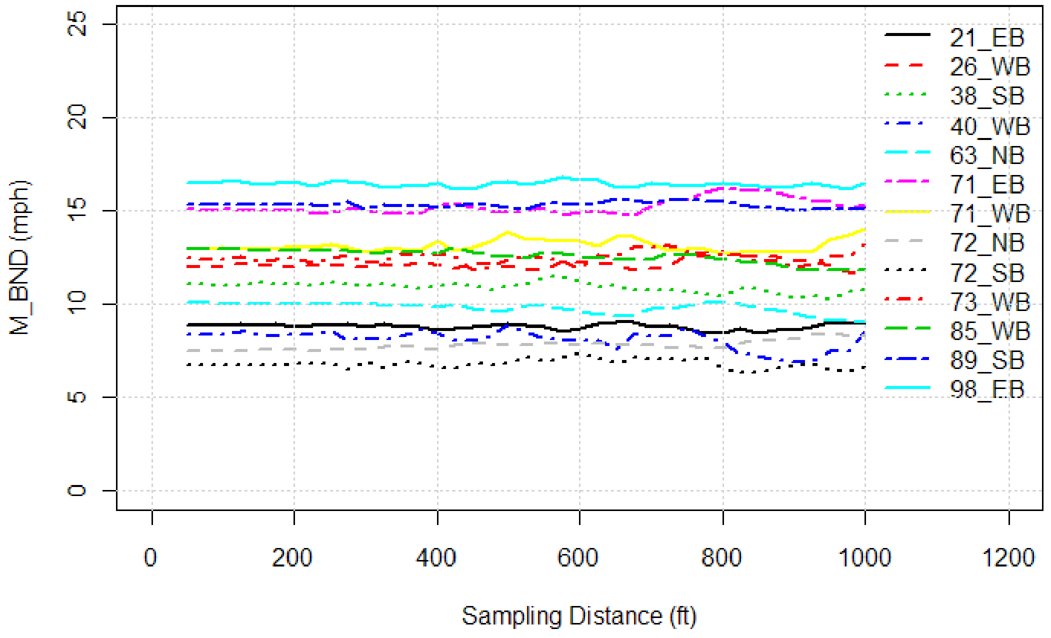


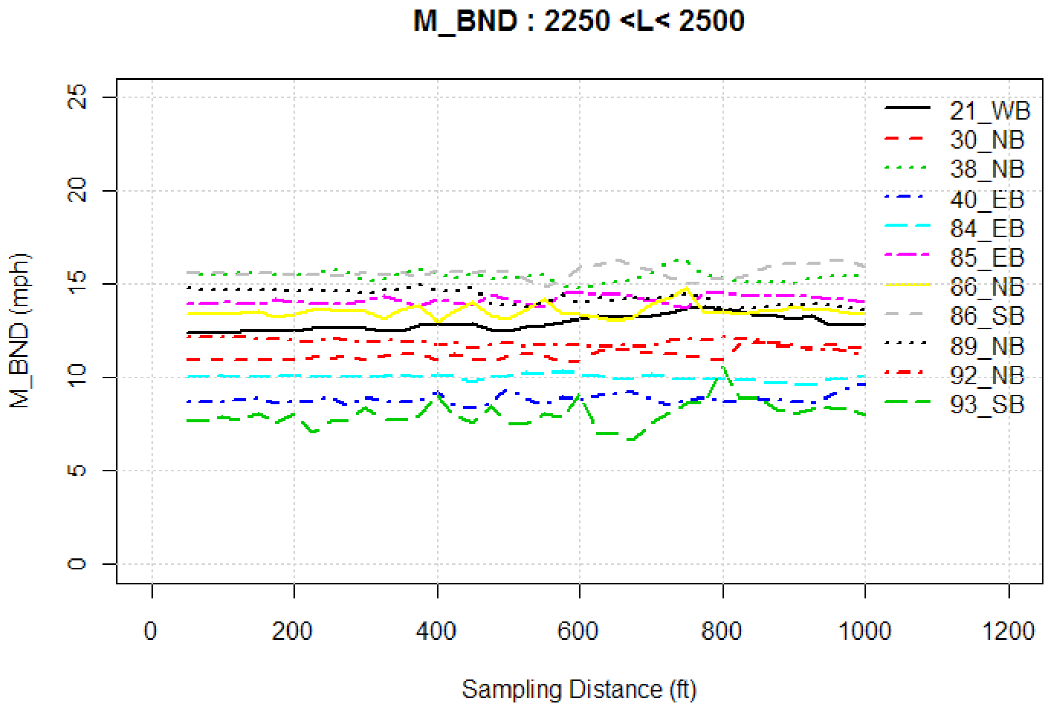


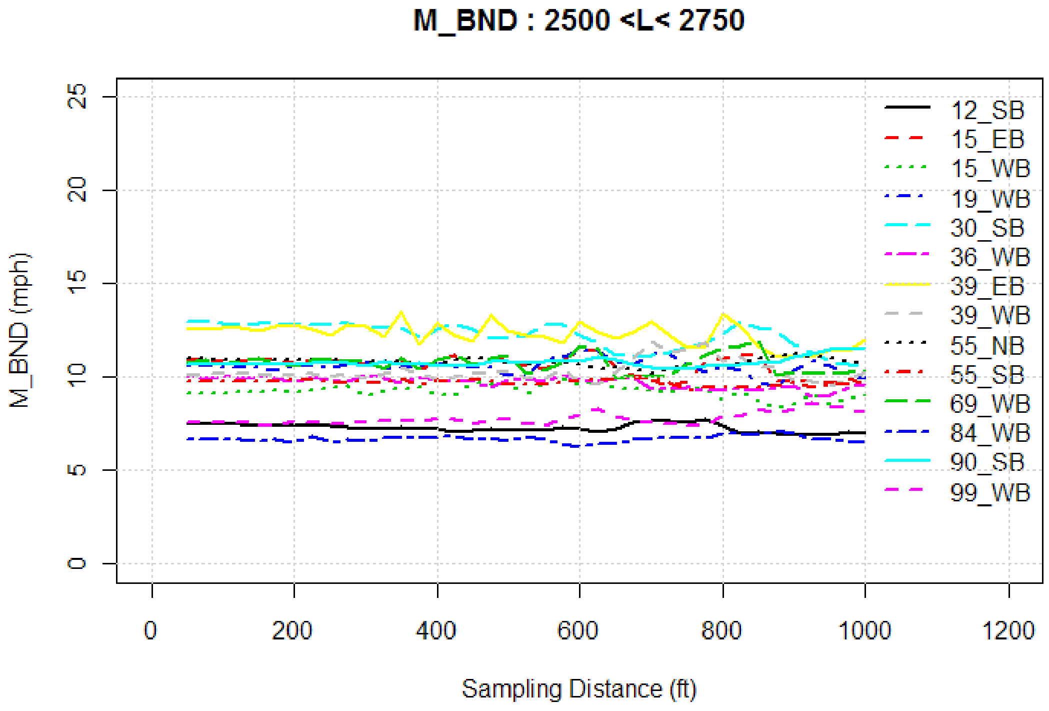
M_BND : 1500 <L< 2000

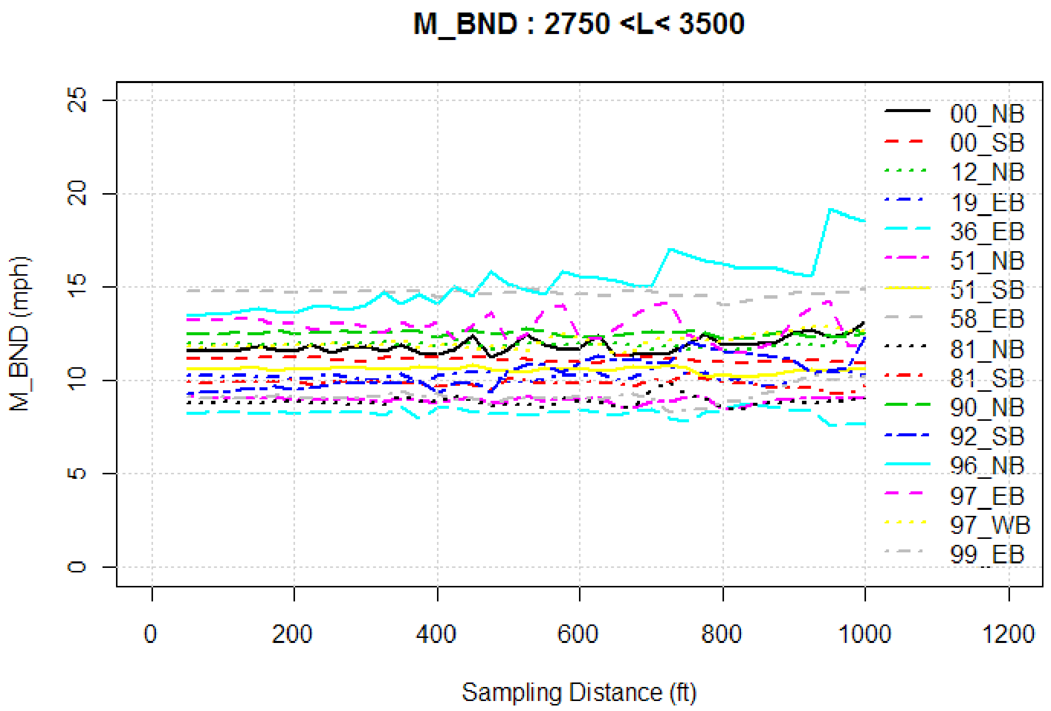


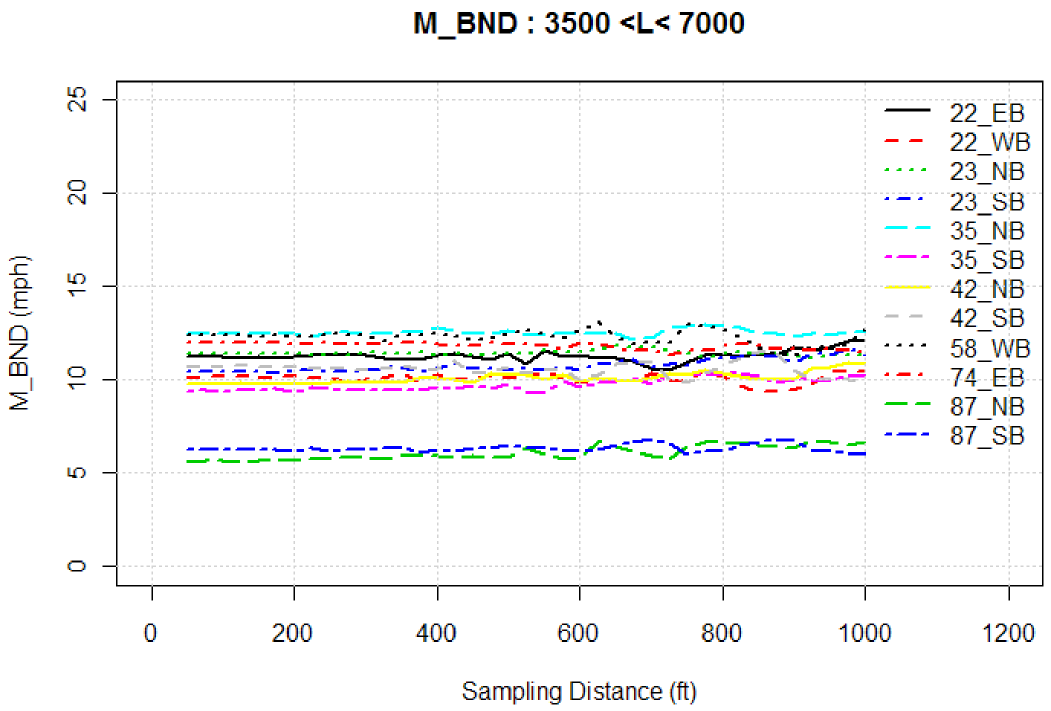
M_BND : 2000 <L< 2250



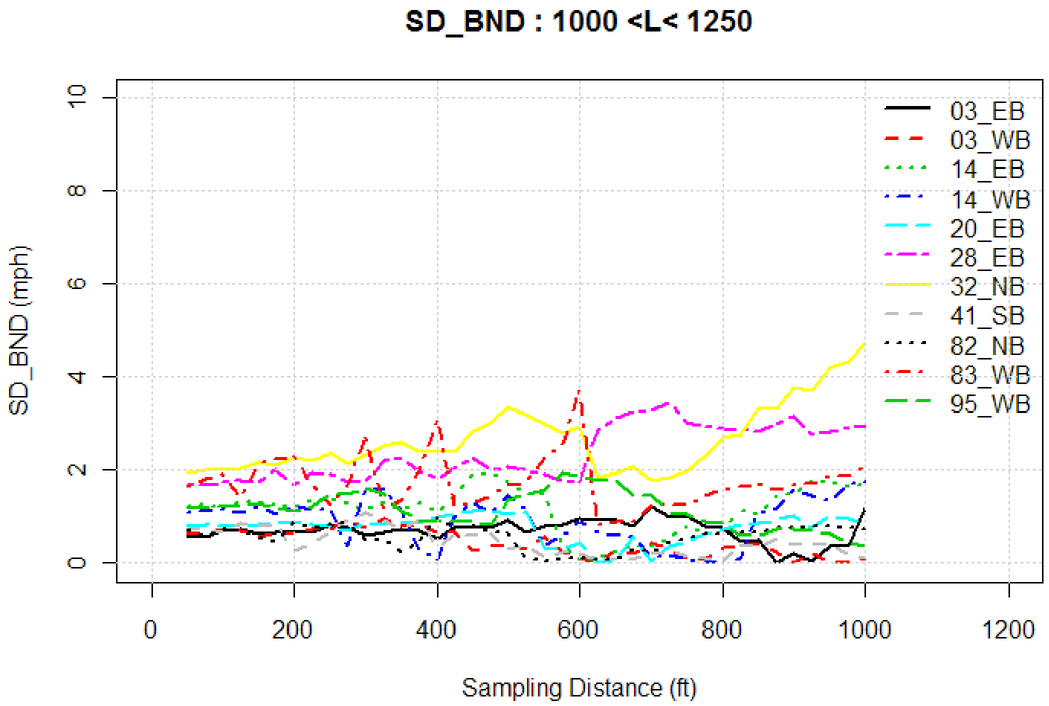

M_BND : 2250 <L< 2500
M_BND (mph)
0
5
10
15
20
25
0
200
400
600
800
1000
1200
Sampling Distance (ft)
21_WB
30_NB
38_NB
40_EB
84_EB
85_EB
86_NB
86_SB
89_NB
92_NB
93_SB

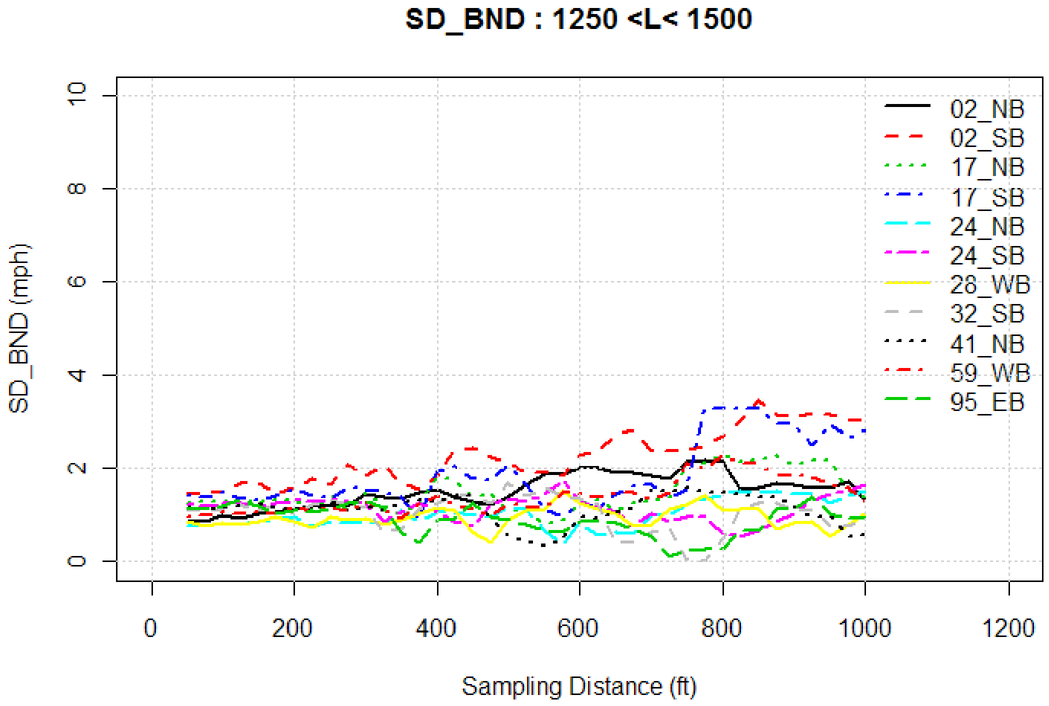

M_BND : 2500 <L< 2750
M_BND (mph)
0
5
10
15
20
25
0
200
400
600
800
1000
1200
Sampling Distance (ft)
12_SB
15_EB
15_WB
19_WB
30_SB
36_WB
39_EB
39_WB
55_NB
55_SB
69_WB
84_WB
90_SB
99_WB

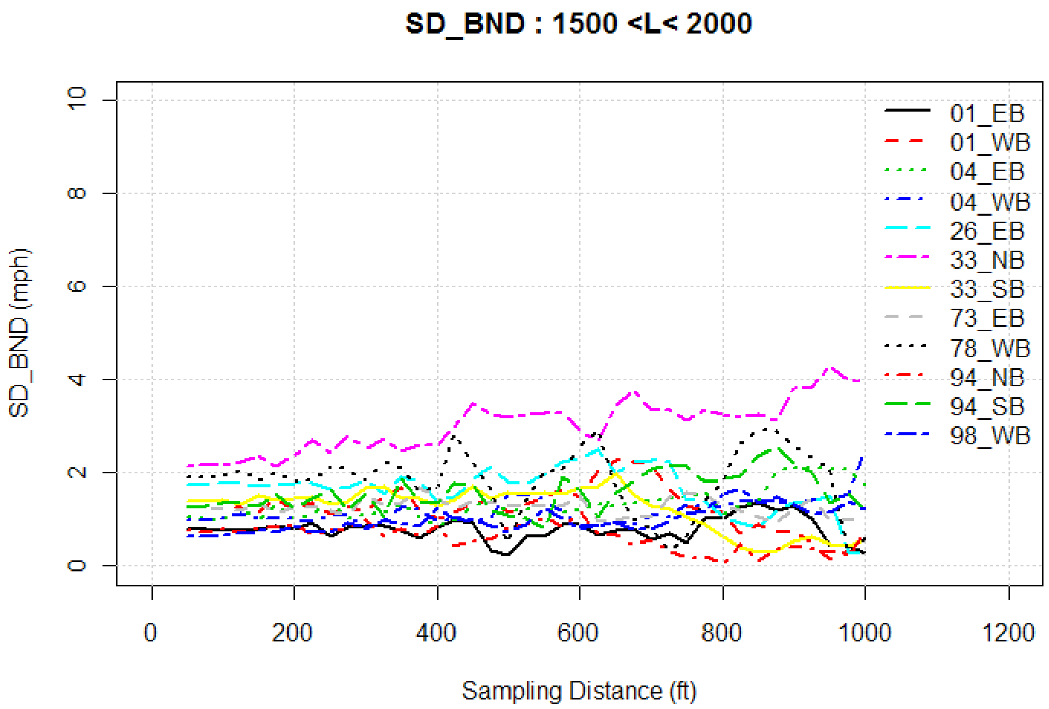

M_BND : 2750 <L< 3500
M_BND (mph)
0
5
10
15
20
25
0
200
400
600
800
1000
1200
Sampling Distance (ft)
00_NB
00_SB
12_NB
19_EB
36_EB
51_NB
51_SB
58_EB
81_NB
81_SB
90_NB
92_SB
96_NB
97_EB
97_WB
99_EB

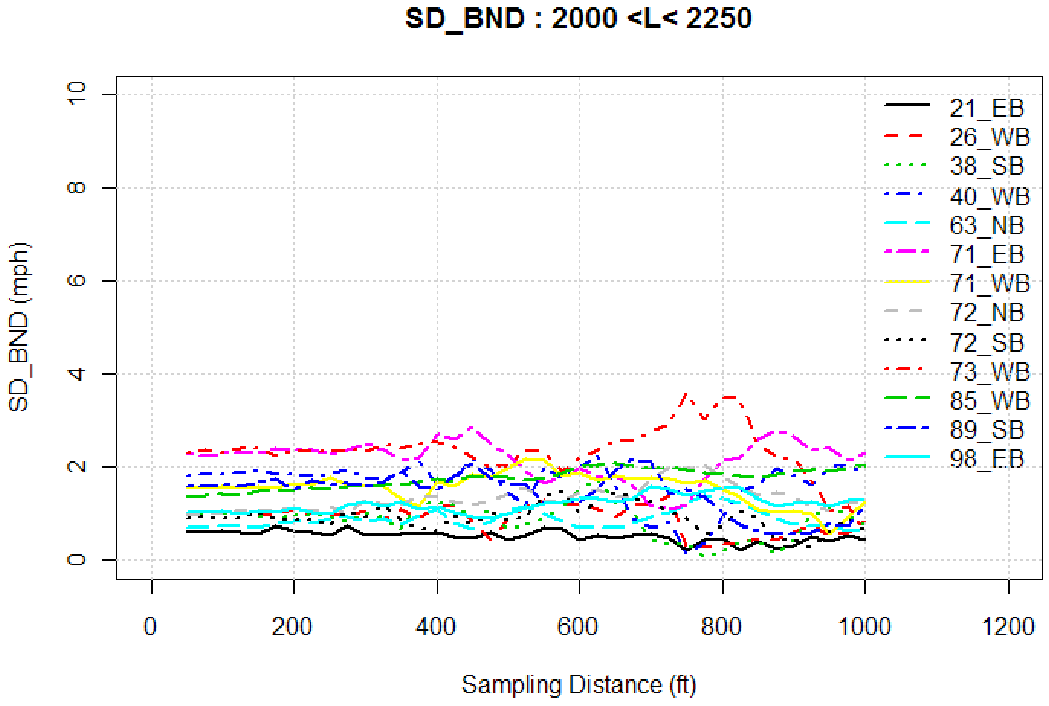

M_BND : 3500 <L< 7000
M_BND (mph)
0
5
10
15
20
25
0
200
400
600
800
1000
1200
Sampling Distance (ft)
22_EB
22_WB
23_NB
23_SB
35_NB
35_SB
42_NB
42_SB
58_WB
74_EB
87_NB
87_SB

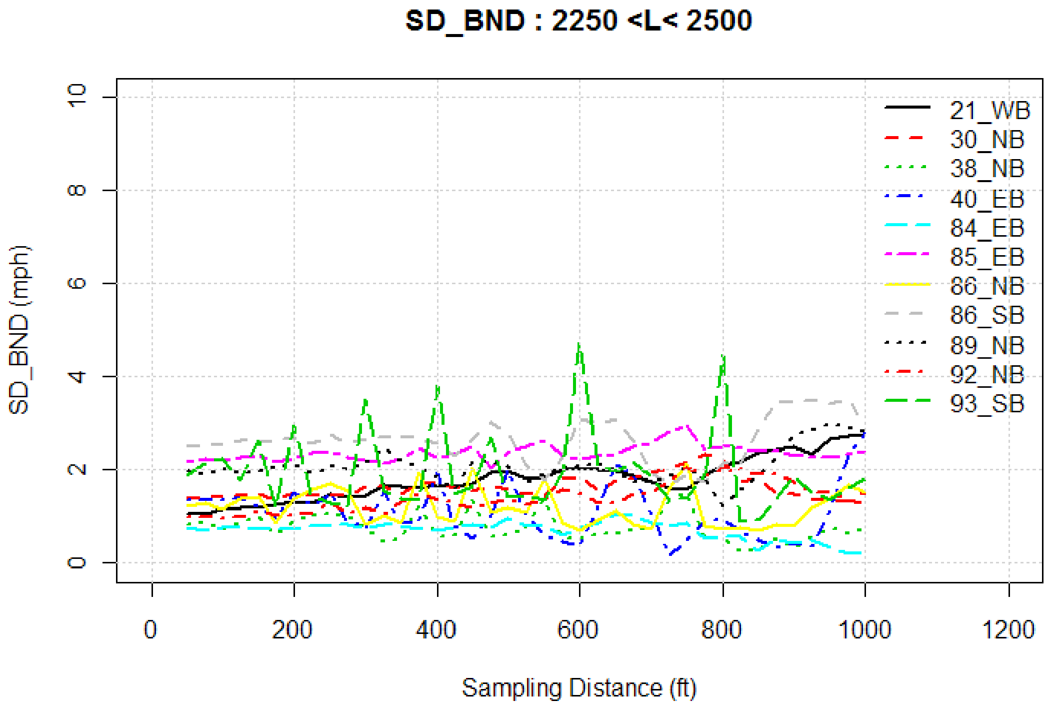
Sensitivity of Variation of Speed Bands (SD_BND) Plots

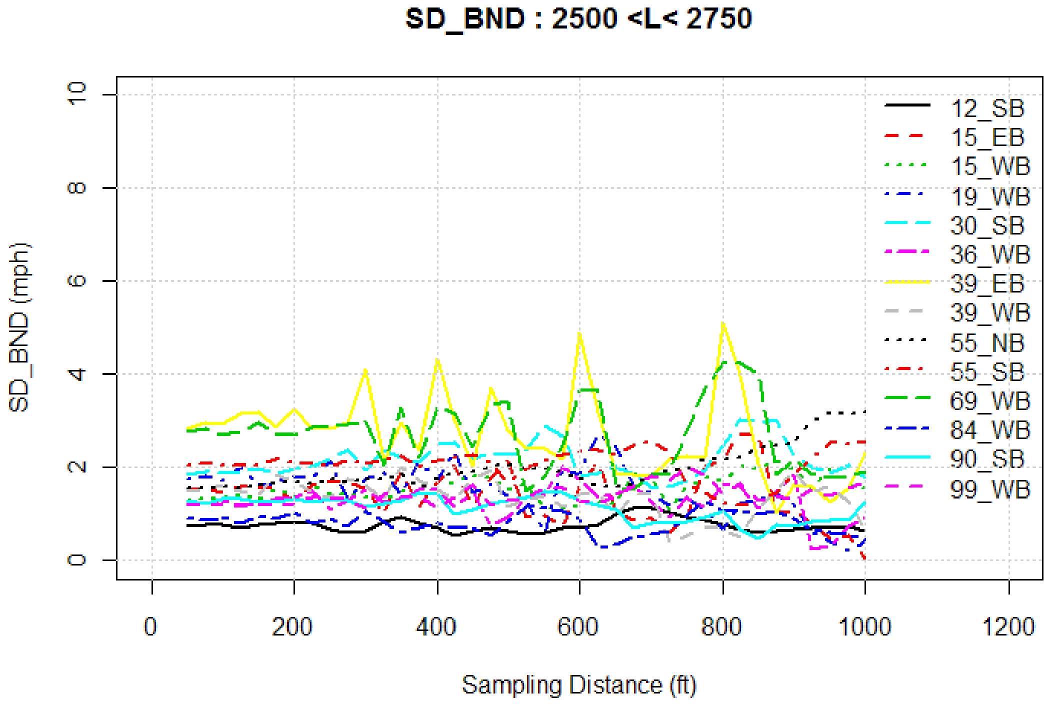


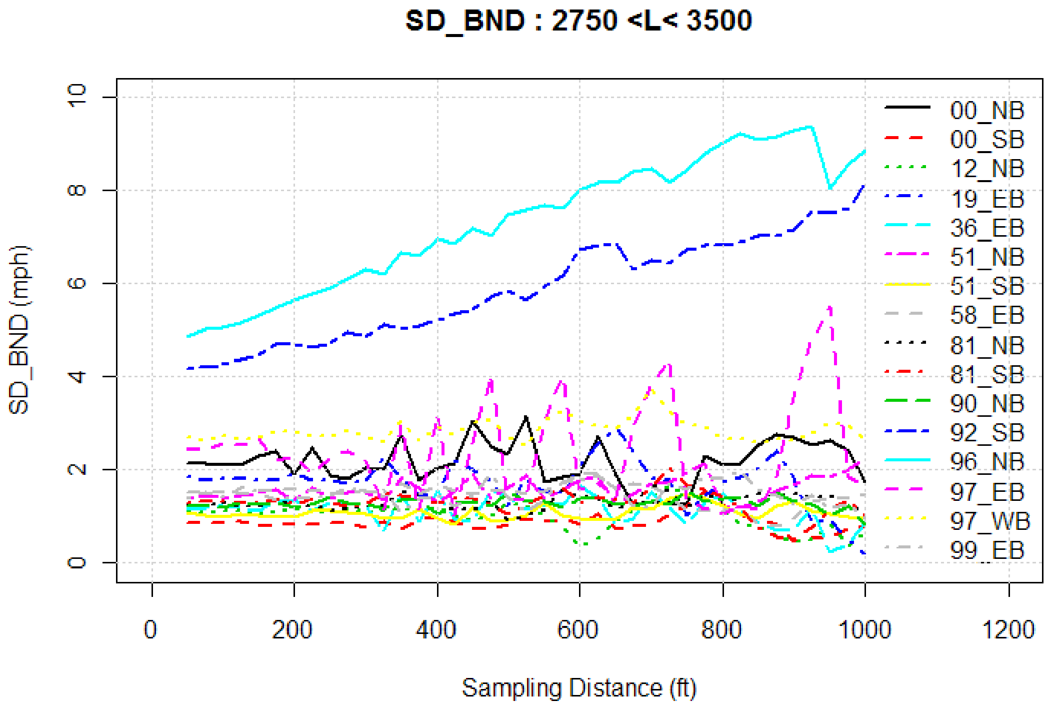


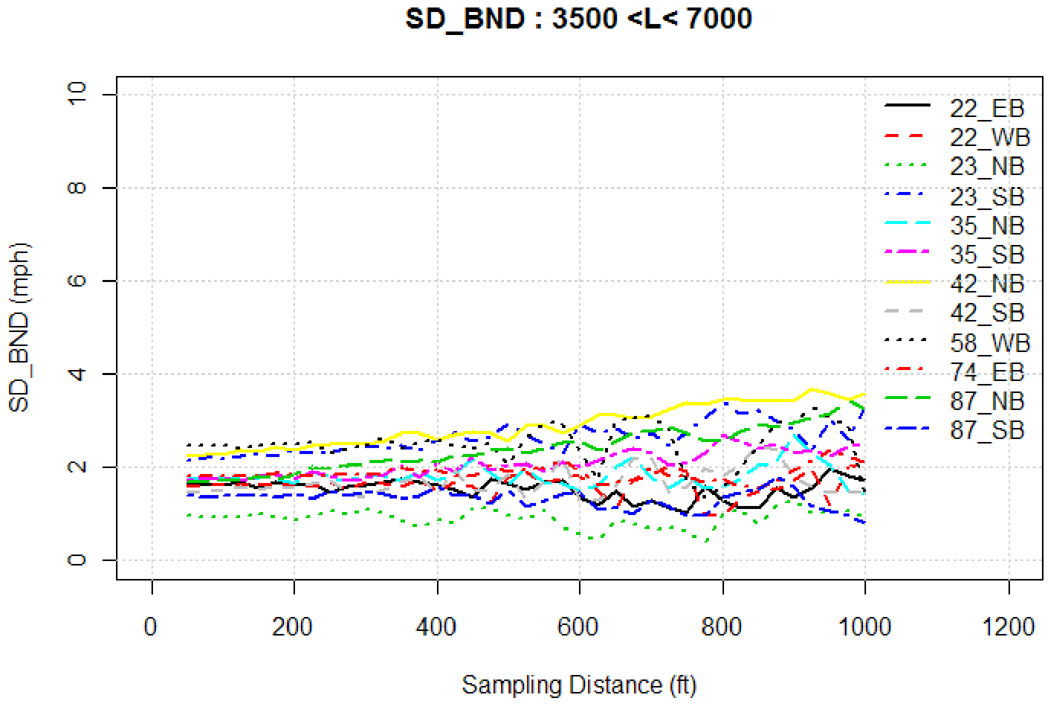

SD_BND : 1500 <L< 2000
SD_BND (mph)
0
2
4
6
8
10
0
200
400
600
800
1000
1200
Sampling Distance (ft)
01_EB
01_WB
04_EB
04_WB
26_EB
33_NB
33_SB
73_EB
78_WB
94_NB
94_SB
98_WB


SD_BND : 2000 <L< 2250
SD_BND (mph)
0
2
4
6
8
10
0
200
400
600
800
1000
1200
Sampling Distance (ft)
21_EB
26_WB
38_SB
40_WB
63_NB
71_EB
71_WB
72_NB
72_SB
73_WB
85_WB
89_SB
98_EB

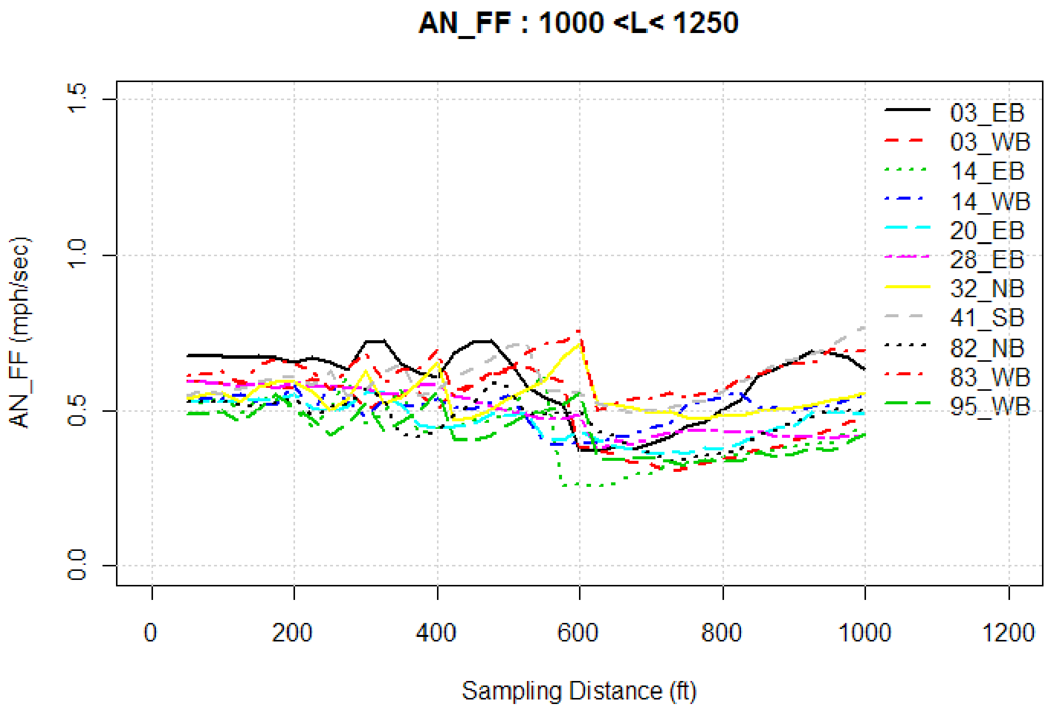

SD_BND : 2250 <L< 2500
SD_BND (mph)
0
2
4
6
8
10
0
200
400
600
800
1000
1200
Sampling Distance (ft)
21_WB
30_NB
38_NB
40_EB
84_EB
85_EB
86_NB
86_SB
89_NB
92_NB
93_SB

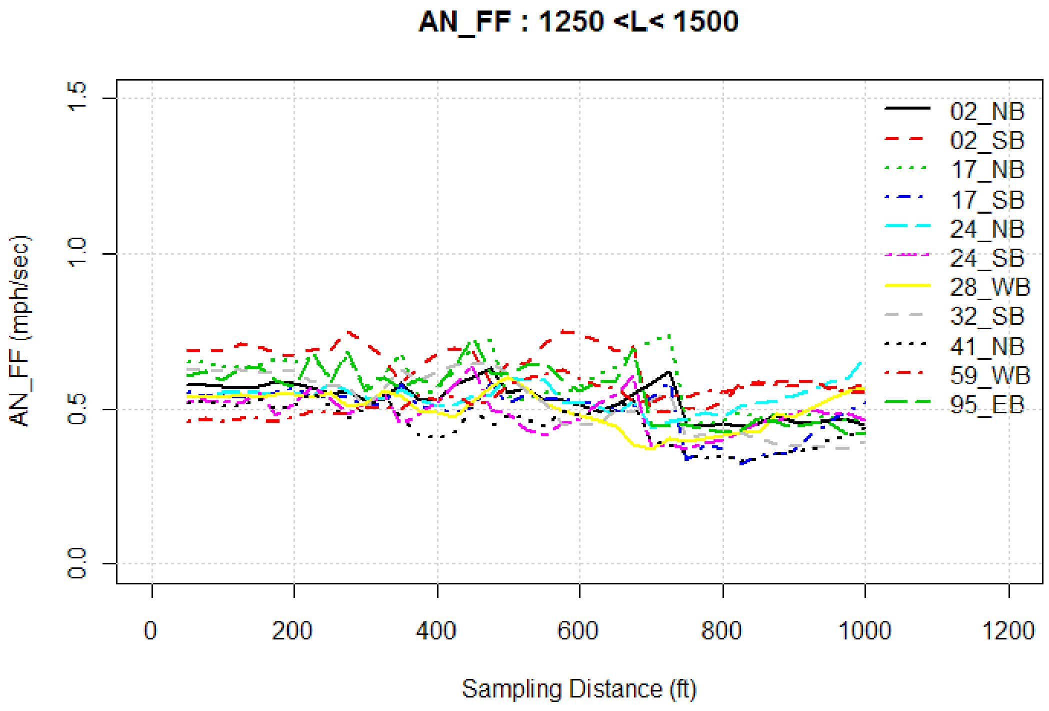

SD_BND : 2500 <L< 2750
SD_BND (mph)
0
2
4
6
8
10
0
200
400
600
800
1000
1200
Sampling Distance (ft)
12_SB
15_EB
15_WB
19_WB
30_SB
36_WB
39_EB
39_WB
55_NB
55_SB
69_WB
84_WB
90_SB
99_WB

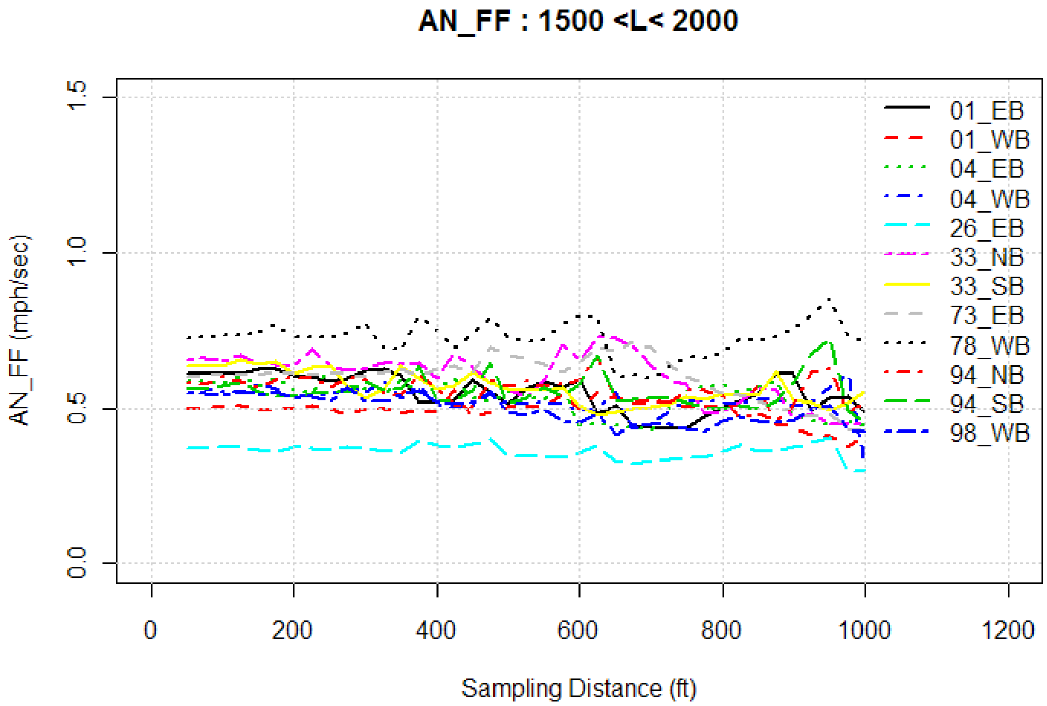

SD_BND : 2750 <L< 3500
SD_BND (mph)
0
2
4
6
8
10
0
200
400
600
800
1000
1200
Sampling Distance (ft)
00_NB
00_SB
12_NB
19_EB
36_EB
51_NB
51_SB
58_EB
81_NB
81_SB
90_NB
92_SB
96_NB
97_EB
97_WB
99_EB

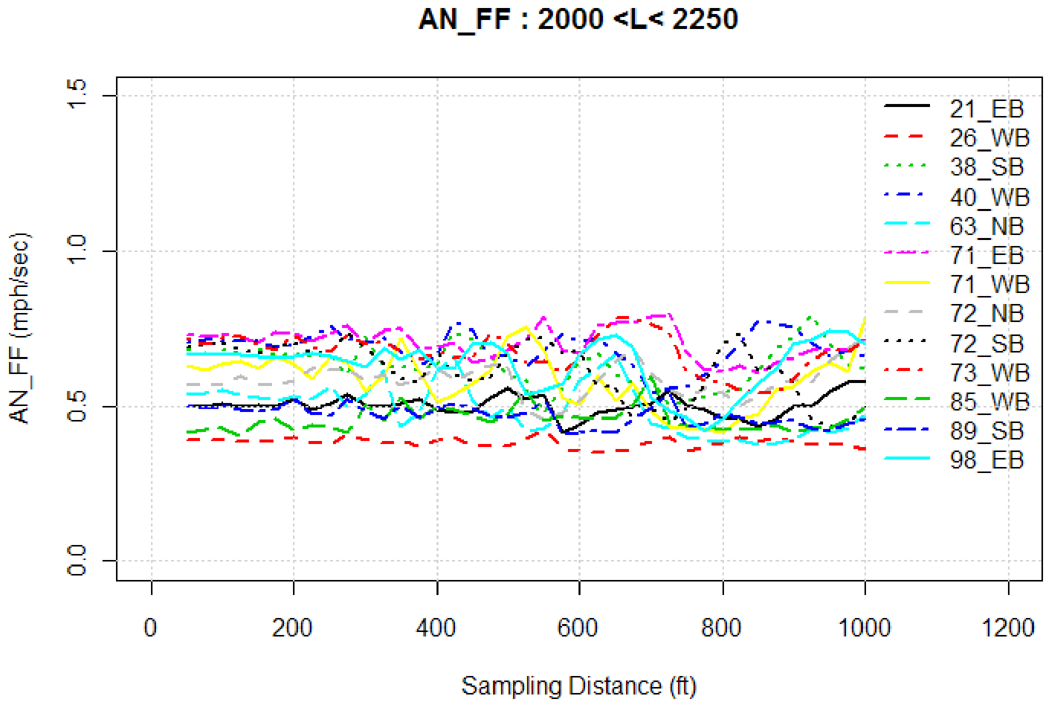

SD_BND : 3500 <L< 7000
SD_BND (mph)
0
2
4
6
8
10
0
200
400
600
800
1000
1200
Sampling Distance (ft)
22_EB
22_WB
23_NB
23_SB
35_NB
35_SB
42_NB
42_SB
58_WB
74_EB
87_NB
87_SB

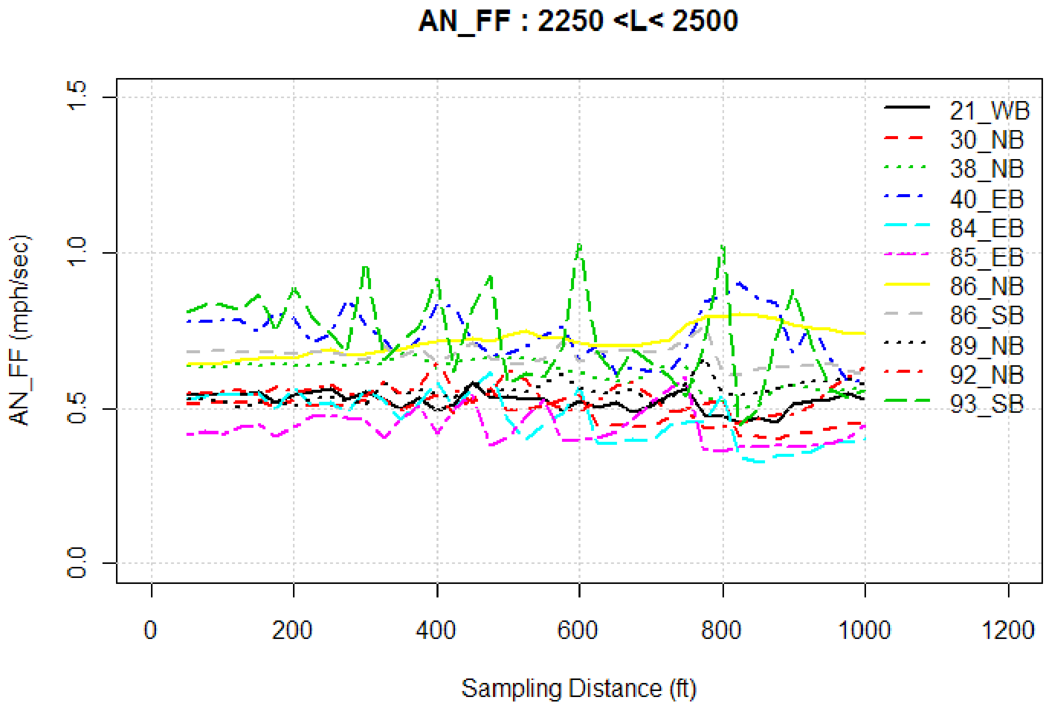
Sensitivity of Acceleration Noise under Free-Flow Condition (AN_FF) Plots

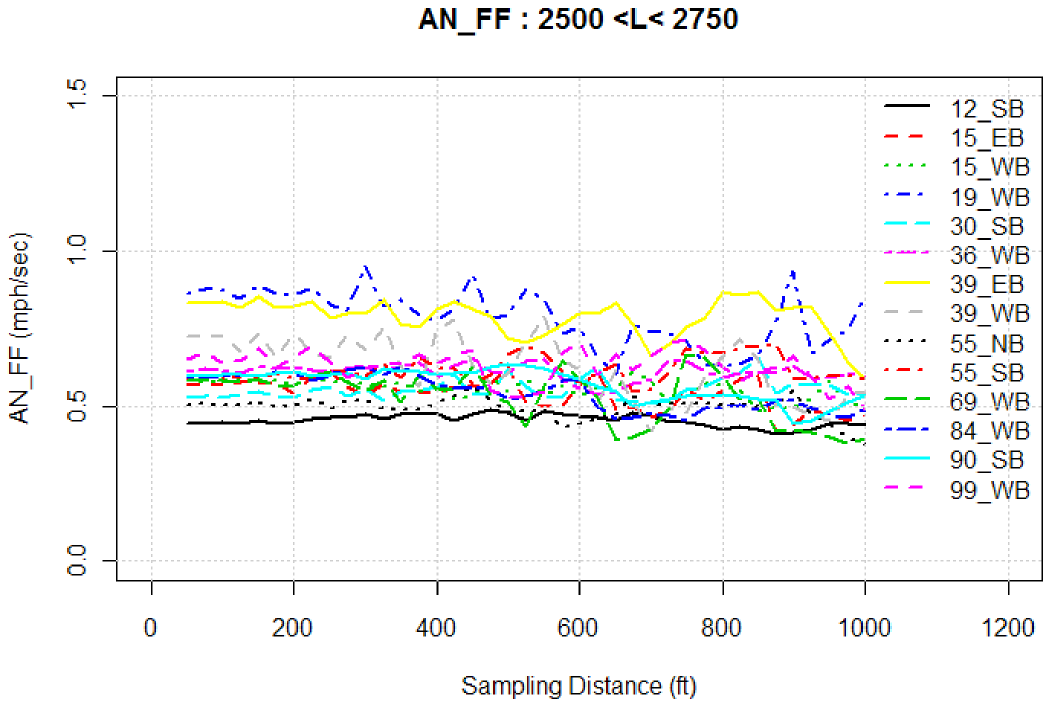


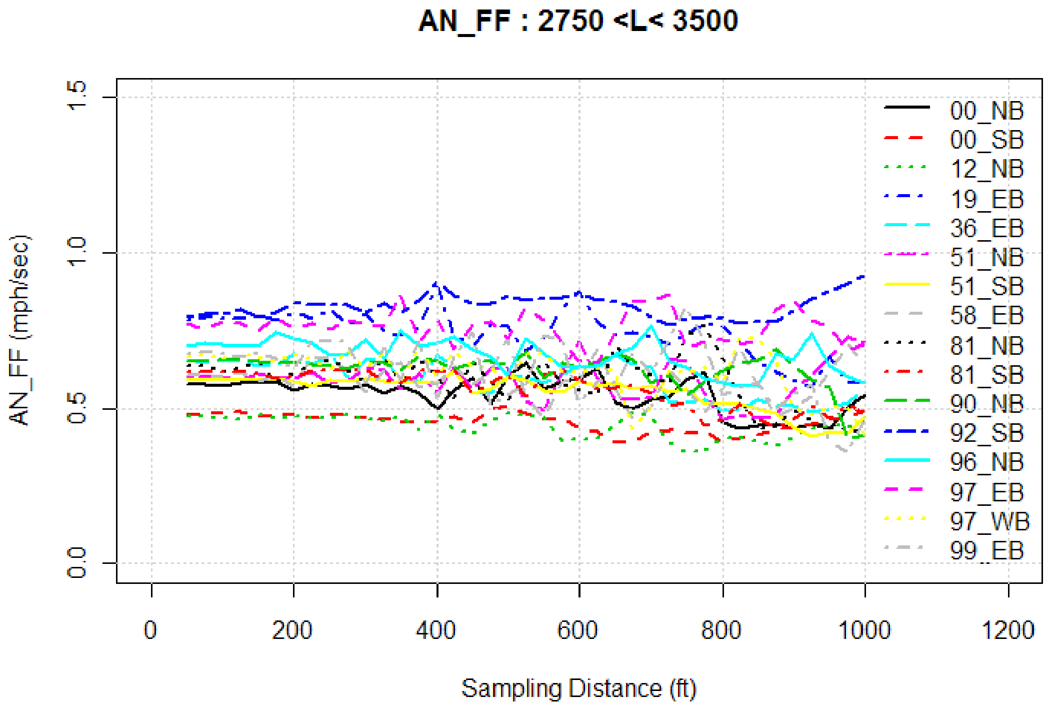


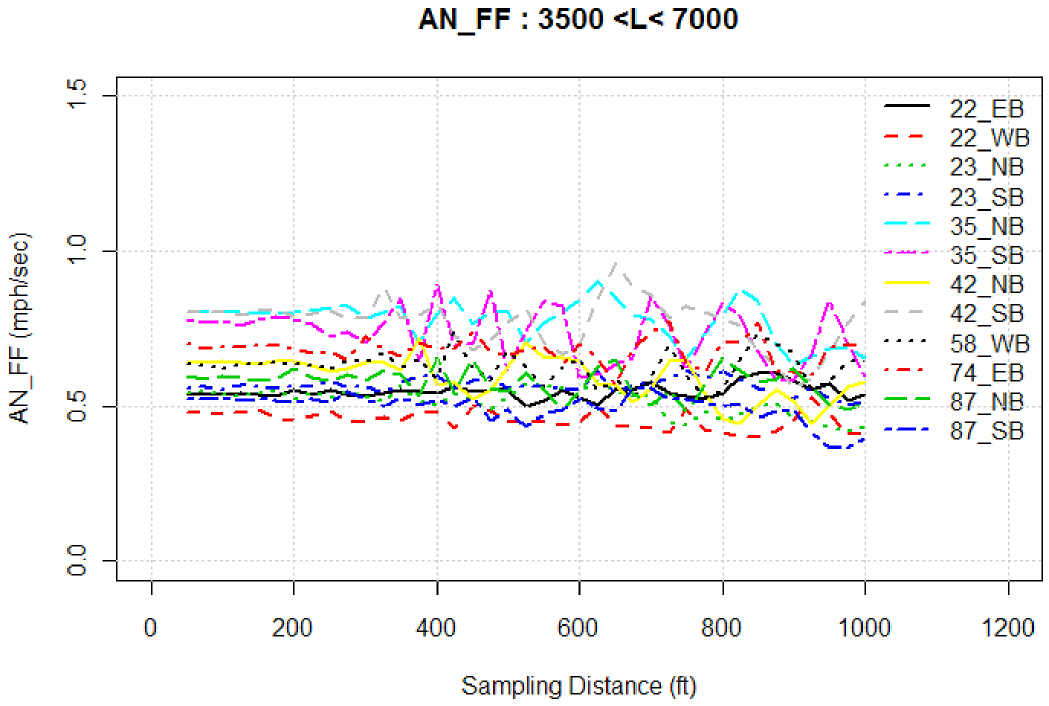

AN_FF : 1500 <L< 2000
AN_FF (mph/sec)
0.0
0.5
1.0
1.5
0
200
400
600
800
1000
1200
Sampling Distance (ft)
01_EB
01_WB
04_EB
04_WB
26_EB
33_NB
33_SB
73_EB
78_WB
94_NB
94_SB
98_WB


AN_FF : 2000 <L< 2250
AN_FF (mph/sec)
0.0
0.5
1.0
1.5
0
200
400
600
800
1000
1200
Sampling Distance (ft)
21_EB
26_WB
38_SB
40_WB
63_NB
71_EB
71_WB
72_NB
72_SB
73_WB
85_WB
89_SB
98_EB

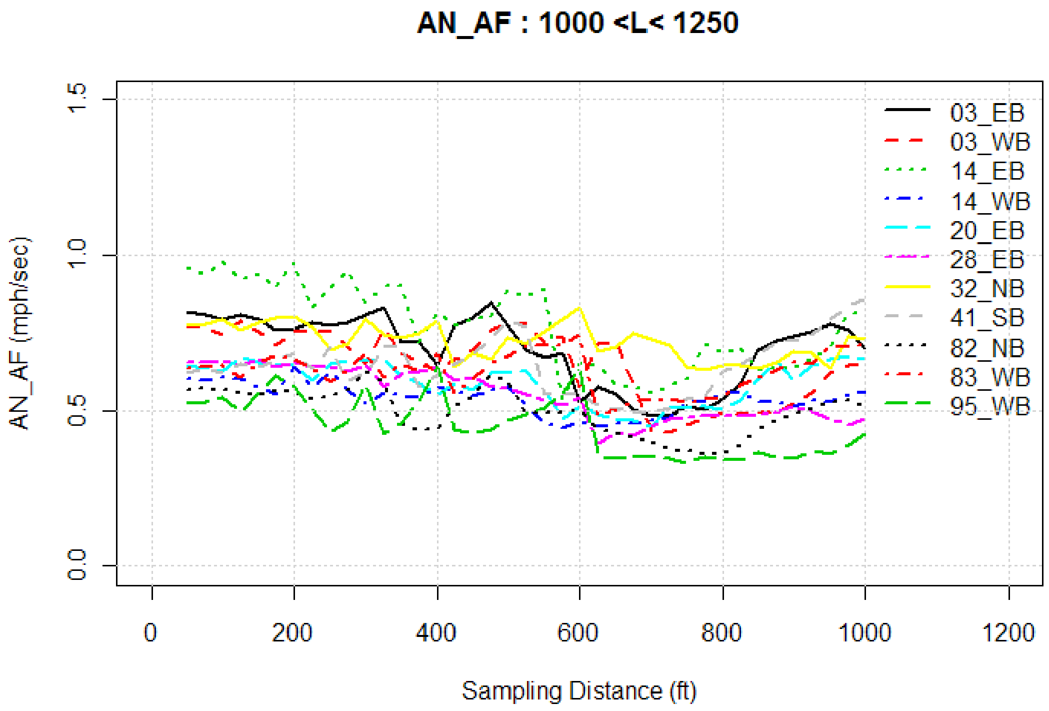

AN_FF : 2250 <L< 2500
AN_FF (mph/sec)
0.0
0.5
1.0
1.5
0
200
400
600
800
1000
1200
Sampling Distance (ft)
21_WB
30_NB
38_NB
40_EB
84_EB
85_EB
86_NB
86_SB
89_NB
92_NB
93_SB

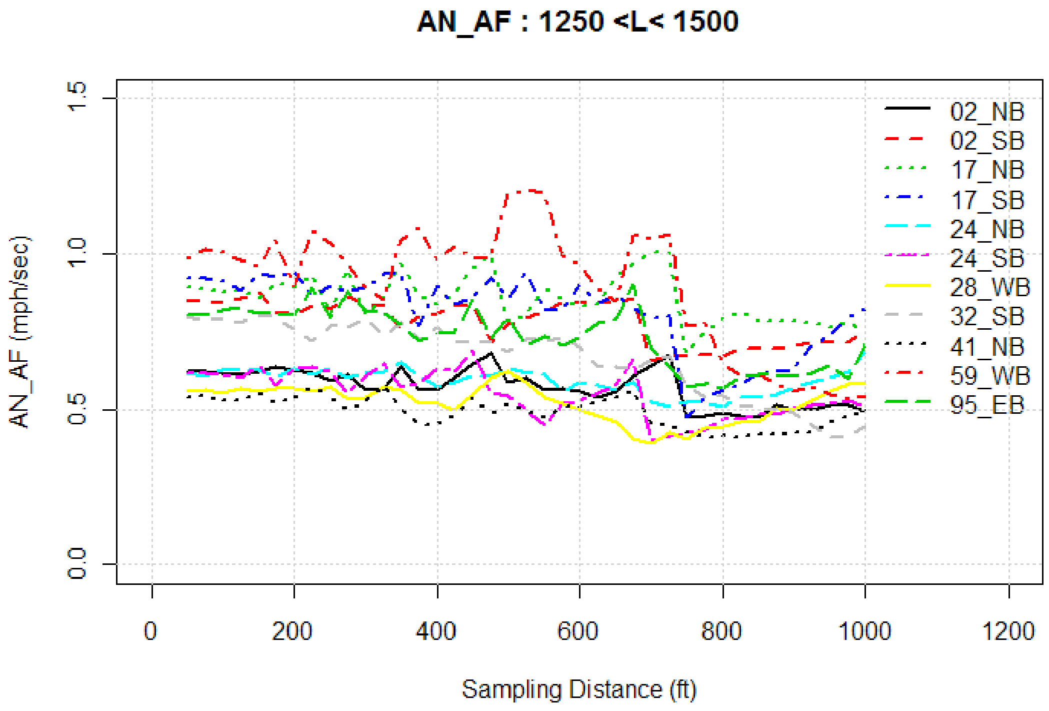

AN_FF : 2500 <L< 2750
AN_FF (mph/sec)
0.0
0.5
1.0
1.5
0
200
400
600
800
1000
1200
Sampling Distance (ft)
12_SB
15_EB
15_WB
19_WB
30_SB
36_WB
39_EB
39_WB
55_NB
55_SB
69_WB
84_WB
90_SB
99_WB

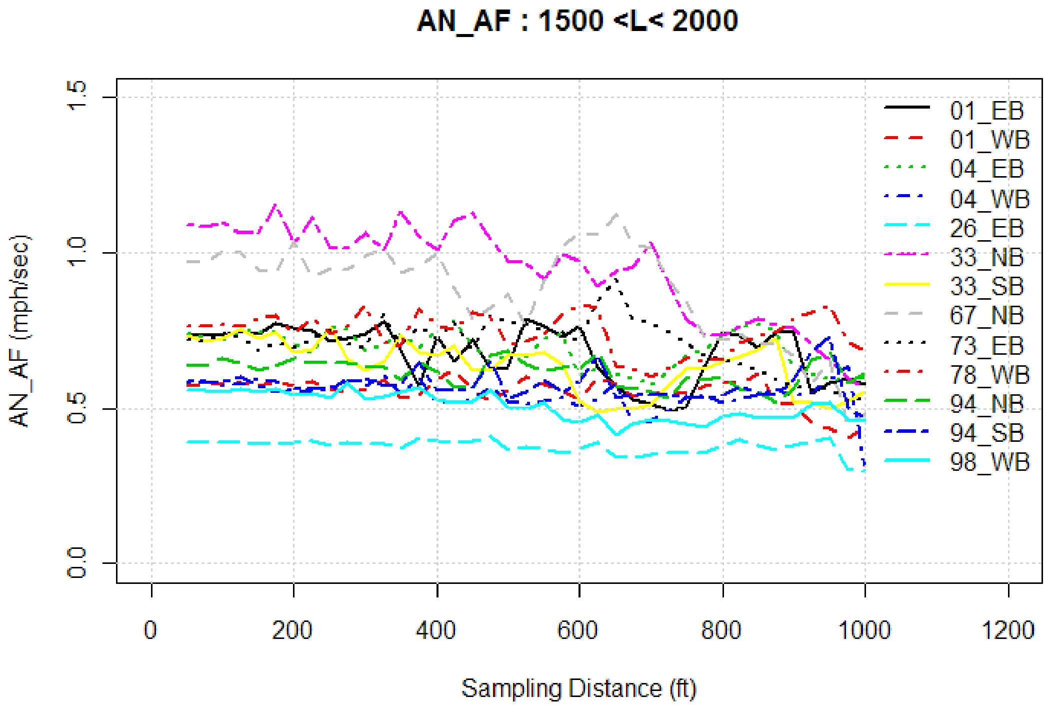

AN_FF : 2750 <L< 3500


AN_FF : 3500 <L< 7000

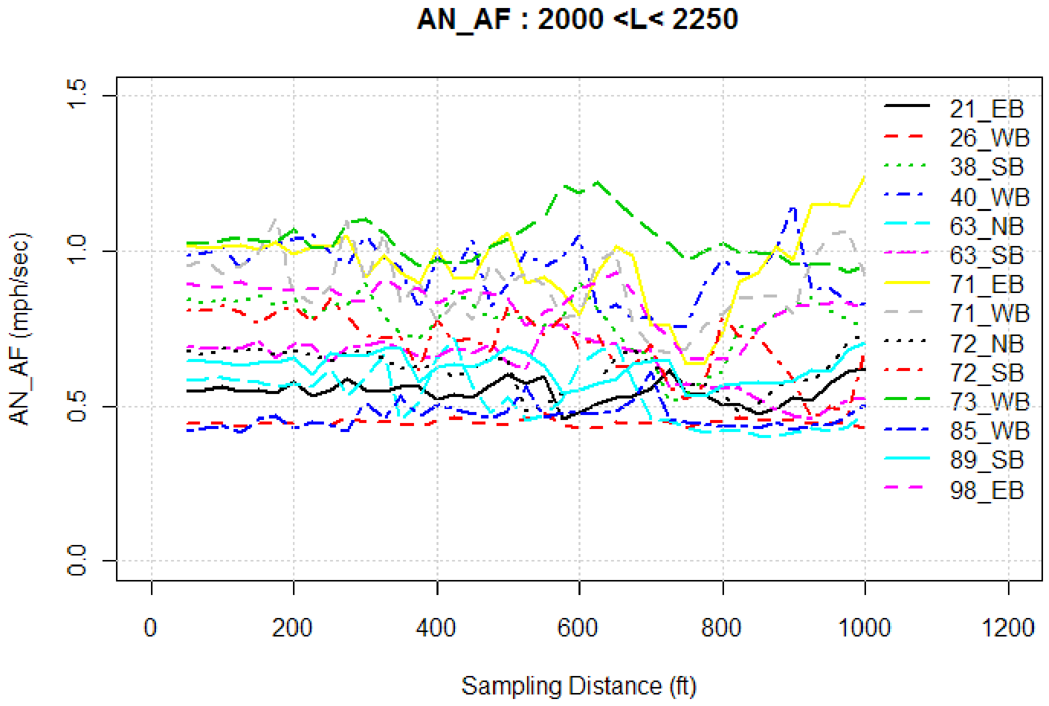
Sensitivity of Acceleration Noise under All-Flow Condition (AN_AF) Plots



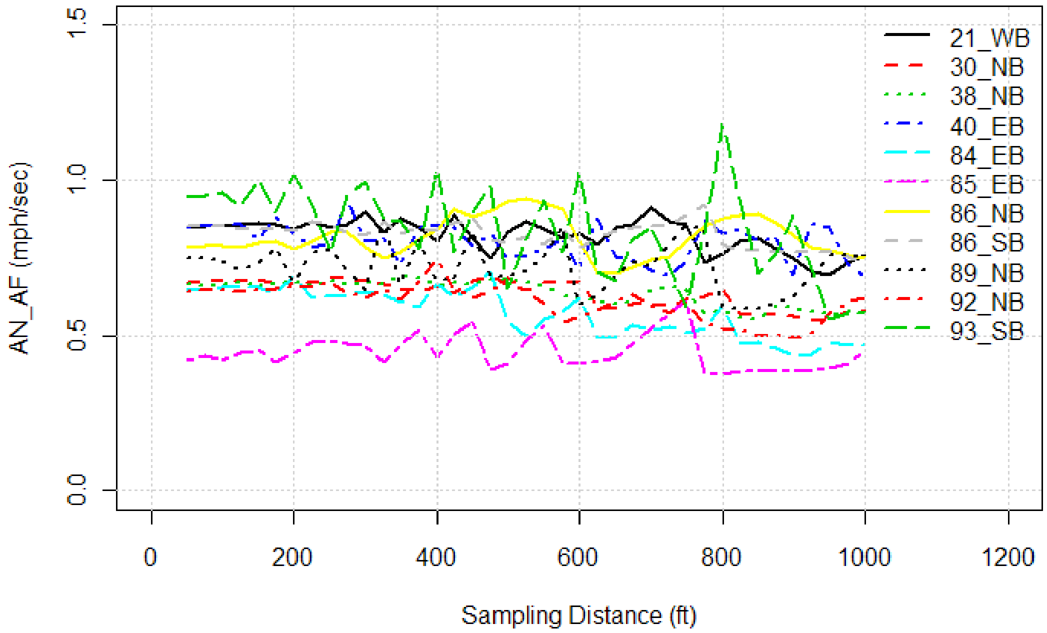




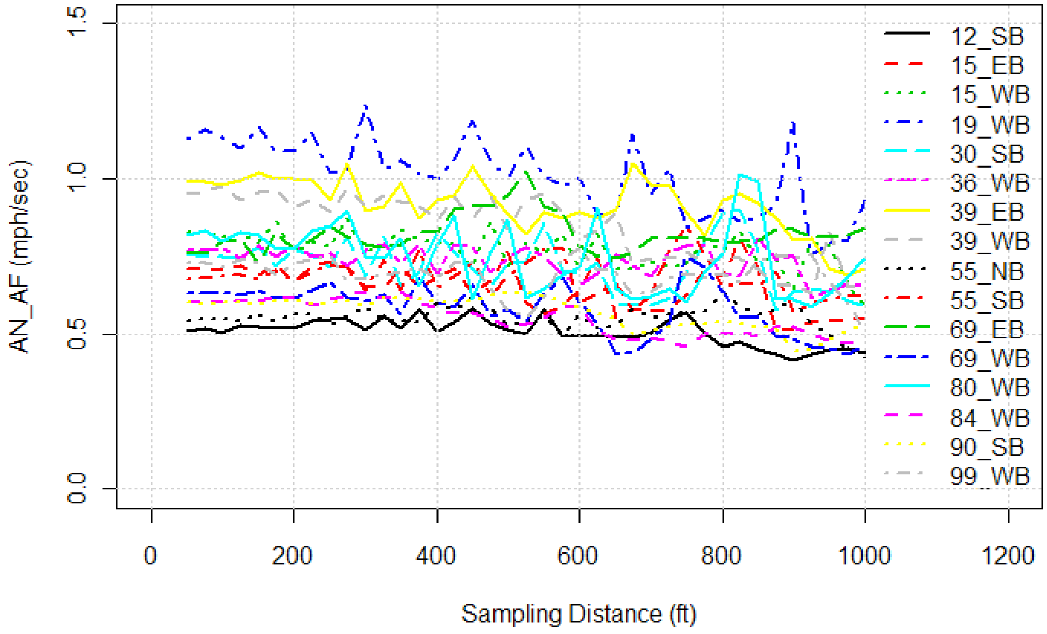
AN_AF : 1500 <L< 2000



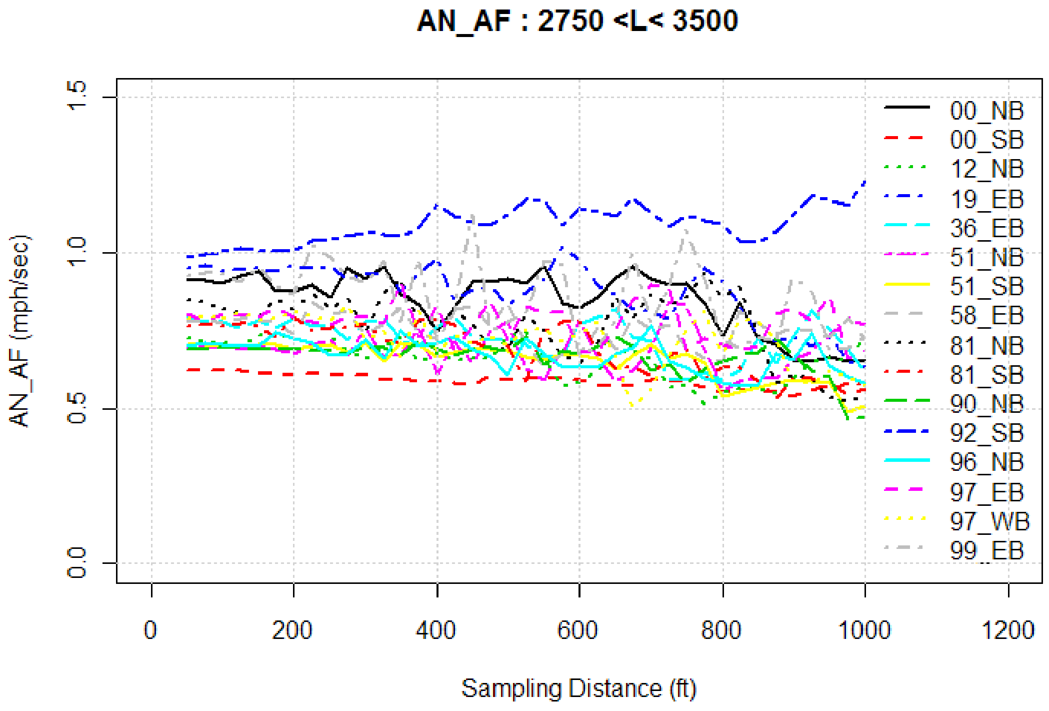
AN_AF : 2000 <L< 2250

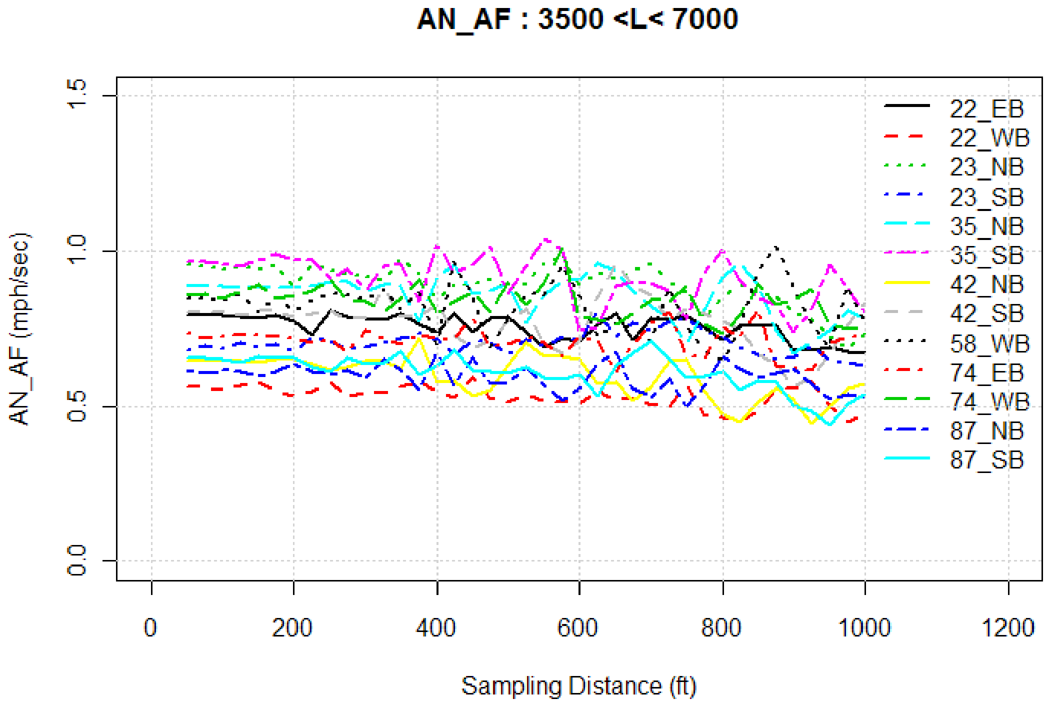


AN_AF : 2250 <L< 2500



AN_AF : 2500 <L< 2750


AN_AF : 2750 <L< 3500
AN_AF (mph/sec)
0.0
0.5
1.0
1.5
Sampling Distance (ft)
0
200
400
600
800
1000
1200
00_NB
00_SB
12_NB
19_EB
36_EB
51_NB
51_SB
58_EB
81_NB
81_SB
90_NB
92_SB
96_NB
97_EB
97_WB
99_EB


AN_AF : 3500 <L< 7000
AN_AF (mph/sec)
0.0
0.5
1.0
1.5
Sampling Distance (ft)
0
200
400
600
800
1000
1200
22_EB
22_WB
23_NB
23_SB
35_NB
35_SB
42_NB
42_SB
58_WB
74_EB
74_WB
87_NB
87_SB

REFERENCES

1. *Transportation Statistics Annual Report.* Publication U.S. Department of Transportation, Reserach and Innovative Technology Administration, Bureau of Transportation Statistics, 2006.

2. *Safe, Accountable, Flexible, Efficient Transportation Equity Act: A Legacy for Users (SAFETEA-LU) - A Summary of Highway Provisions.* Publication Office of Legislation and Intergovernmental Affairs, Federal Highway Administration, 2005.

3. Hauer, E., et al. Screening the Road Network for Sites with Promise. In *Transportation Research Record*. Vol.1784, 2002, pp. 27-32.

4. Gettman, D. and L. Head. Surrogate Safety Measures from Traffic Simulation Models. In *Transportation Research Record*. Vol.1840, 2003, pp. 104-115.

5. Aljanahi, A.A.M., A.H. Rhodes, and A.V. Metcalfe. Speed, speed limits and road traffic accidents under free flow conditions. In *Accident Analysis & Prevention*. Vol.31, 1999, pp. 161-168.

6. Dixon, K.K., et al. *Effects of Urban Street Environment on Operating Speeds.* Publication Federal Highway Administration, U.S. Department of Transportation, 2007.

7. Cooper, P.J. The relationship between speeding behaviour (as measured by violation convictions) and crash involvement. In *Journal of Safety Research*. Vol.28, 1997, pp. 83-95.

8. Fildes, B.N., G. Rumbold, and A. Leening. *Speed Behaviour and Drivers' Attitude to Speeding.* Publication Monash University, Accident Research Centre, 1991.

9. Garber, N.J. and R. Gadiraju. Factors affecting speed variance and its influence on accidents. In *Transportation Research Record*. Vol.1213, 1989, pp. 64-71.

10. Harrison, W.A., et al. *An investigation of characteristics associated with driving speed.* Publication 140,Monash University Accident Research Centre, 1998.

11. Kloeden, C.N., et al. *Traveling Speed and the Risk of Crash Involvement; Volume 1 - Findings.* Publication NHMRC Road Accident Research Unit, University of Adelaide, 1997.

12. Kockelman, K.M. and W.J. Murray. Freeway speeds and speed variations preceding crashes, within and across lanes. In *Journal of the Transportation Research Forum*. 2007, pp. 1-19.

13. Lave, C.A. Speeding, Coordination, and the 55 MPH Limit. In *The American Economic Review*. Vol.75, 1985, pp. 1159-1164.

14. Rogdriguez, R.J. Speed. Speed Dispersion, and the Highway Fatality Rate. In *Southern Economic Journal*. Vol.57, 1990, pp. 349.

15. Solomon, D. *Accidents on Main Rural Highways Related to Speed, Driver and Vehicle.* Publication U.S. Department of Transportation/ Federal Highway Administration, 1964.

16. West, R. and D. French. Direct observation of driving, self reports of driver behaviour, and accident involvement. In *Ergonomics*. Vol.36, 1993, pp. 11.

17. Anderson, I., et al. Relationship to Safety of Geometric Design Consistency Measures for Rural Two-Lane Highways. In *Transportation Research Record: Journal of the Transportation Research Board*. Vol.1658, 1999, pp. 43-51.

18. Ng, J.C.W. and T. Sayed. Effect of geometric design consistency on road safety. In *Canadian Journal of Civil Engineering*. Vol.31, 2004, pp. 218-227.

19. Misener, J.A. PATH investigations in vehicle-roadside cooperation and safety : a foundation for safety and vehicle-infrastructure integration research. In *2006 IEEE Intelligent Transportation Systems Conference : Toronto, Ontario, Canada, 17-20 September 2006. Vol. 1*. 2006, pp. 9-16.

20. Shladover, S.E., *Preparing the way for vehicle-infrastructure integration*. California PATH Program, Institute of Transportation Studies, University of California at Berkeley, Berkeley, Calif., 2005.

21. Hauer, E., *Speed and Safety*, in *Transportation Research Board 88th Annual Meeting*. 2009, Transportation Research Board: Washington DC.

22. Ogle, J. *Quantitative Assessment of Driver Speeding Behavior Using Instrumented Vehicles*.Doctoral Dissertation, Civil & Environmental Engineering, Georgia Institute of Technology, Atlanta, 2005.

23. Jun, J. *Potential Crash Exposure Measures Based on GPS-Observed Driving Behavior Activity Metrics*.Doctoral Dissertation, Civil and Environmental Engineering, Georgia Institute of Technology, Atlanta, 2006.

24. National Research Council . Transportation Research Board. Committee for Guidance on Setting and Enforcing Speed, L., *Managing speed : review of current practice for setting and enforcing speed limits*. Transportation Research Board, National Research Council, National Academy Press, Washington, D.C., 1998.

25. Lamm, R. and B. National Research Council . Transportation Research, *A Possible design procedure to promote design consistency in highway geometric design on two-lane rural roads*. National Research Council, Transportation Research Board, Washington, D.C., 1988.

26. Herman, R., et al. Traffic Dynamics: Analysis of Stability in Car Following. In *Operations Research*. Vol.7, 1959, pp. 86.

27. Jones, T.R. and R.B. Potts. The Measurement of Acceleration Noise - A Traffic Parameter. In *Operations Research*. Vol.10, 1962, pp. 745-763.

28. Garber, N.J., et al. *Safety Effects of Differential Speed Limits on Rural Interstate Highways.* Publication FHWA-HRT-05-042,Office of Safety, Federal Highway Administration, 2005.

29. Pande, A. and M. Abdel-Aty. Comprehensive analysis of the relationship between real-time traffic surveillance data and rear-end crashes on freeways. In *Transportation research record.* 2006, pp. 31-40.

30. Bonneson, J.A. and P.T. McCoy. Effect of median treatment on urban arterial safety : an accident prediction model. In *Transportation Research Record: Journal of the Transportation Research Board*. Vol.1581, 1997, pp. 27-36.

31. Hauer, E., *Observational Before-After Studies in Road Safety: Estimating the Effect of Highway and Traffic Engineering Measures on Road Safety*. Pergamon Press, Elsevier Science Ltd., Oxford, U.K., 1997.

32. Persaud, B.N. *Statistical Methods in Highway Safety Analysis.* Publication NCHRP Synthesis 295,National Cooperative Highway Research Program, 2001.

33. Persaud, B.N., C. Lyon, and T. Nguyen. Empirical Bayes procedure for ranking sites for safety investigation by potential for safety improvement. In *Transportation research record.* Vol.1665, 1999, pp. 7-12.

34. Kononov, J. and B. Allery. Level of service of safety conceptual blueprint and analytical framework. In *Transportation research record.* 2003, pp. 57-66.

35. Hauer, E. Statistical road safety modeling. In *Transportation Research Record: Journal of the Transportation Research Board*. Vol.1897, 2004, pp. 81-87.

36. Everitt, B. and T. Hothorn, *A handbook of statistical analyses using R*. Chapman & Hall/CRC, Boca Raton, 2006.

37. Persaud, B.N. Estimating accident potential of Ontario road sections. In *Transportation research record.* Vol.1327, 1991, pp. 47-53.

38. Fridstrom, L., et al. Measuring the contribution of randomness, exposure, weather, and daylight to the variation in road accident counts. In *Accident Analysis & Prevention*. Vol.27, 1995, pp. 1-20.

39. Washington, S., M.G. Karlaftis, and F.L. Mannering, *Statistical and econometric methods for transportation data analysis*. Chapman & Hall/CRC, Boca Raton, 2003.

40. Ogle, J., R. Guensler, and V. Elango. Georgia's commute Atlanta value pricing program : recruitment methods and travel diary response rates. In *Transportation research record.* 2005, pp. 28-37.

41. Jun, J., R. Guensler, and J. Ogle. Smoothing Methods to Minimize Impact of Global Positioning System Random Error on Travel Distance, Speed, and Acceleration Profile Estimates. In *Transportation Research Record: Journal of the Transportation Research Board*. Vol.1972, 2006, pp. 141-150.

42. Ko, J. *Measurement of freeway traffic flow quality using GPS-equipped vehicles*.Doctoral Dissertation, School of Civil and Environmental Engineering, Georgia Institute of Technology, Atlanta, 2006.

43. Quiroga, C.A. and D. Bullock. Measuring Control Delay at Signalized Intersections. In *Journal of Transportation Engineering*. Vol.125, 1999, pp. 271.

44. D'Este, G.M., R. Zito, and M.A.P. Taylor. Using GPS to Measure Traffic System Performance. In *Computer-Aided Civil and Infrastructure Engineering*. Vol.14, 1999, pp. 255-265.

45. Fitzpatrick, K., et al. *Design Speed, Operating Speed, and Posted Speed Practices.* Publication NCHRP Report 504,Transportation Research Board, 2003.

46. Wooldridge, M.D., P. National Cooperative Highway Research, and B. National Research Council . Transportation Research, *Geometric design consistency on high-speed rural two-lane roadways*. Transportation Research Board, National Research Council, Washington, D.C., 2003.

47. Harwood, D.W., et al., *Prediction of the Expected Safety Performance of Rural Two-Lane Highways*. U.S. Dept. of Transportation, Federal Highway Administration, Research Development, and Technology, Turner-Fairbank Highway Research Center, McLean, VA, 2000.

48. Hauer, E., F. Council, and Y. Mohammedshah. Safety Models for Urban Four-Lane Undivided Road Segments. In *Transportation Research Record: Journal of the Transportation Research Board*. Vol.1897, 2004, pp. 96-105.

49. Mitchell, A., *The ESRI guide to GIS analysis, Volume 2: Spatial Measurements & Statistics*. ESRI, Redlands, California, 2005.

50. Yang, J. *CP6521: Advanced Geographic Information Systems Lecture.* Unpublished Manuscript, Fall 2008.

51. FHWA. *FHWA Functional Classification Guidelines*. 1989 [Access March 1, 2009]; Available from: http://www.fhwa.dot.gov/planning/fcsec2_1.htm.

52. Kvam, P.H. and B. Vidakovic, *Nonparametric statistics with applications to science and engineering*. Wiley-Interscience, Hoboken, N.J., 2007.

53. Washington, S. Iteratively Specified Tree-Based Regression: Theory and Trip Generation Example. In *Journal of Transportation Engineering*. Vol.126, 2000, pp. 482.

54. Hastie, T., et al., *The Elements of Statistical Learning : Data Mining, Inference, and Prediction*. Springer, New York, 2001.

55. Faraway, J.J., *Extending the linear model with R : generalized linear, mixed effects and nonparametric regression models*. Chapman & Hall/CRC, Boca Raton, 2006.

56. Therneau, T.M., E.J. Atkinson, and M. Foundation. *An Introduction to Recursive Partitioning Using the RPART Routines*. Publication Stanford University, Department of Statistics, 1997.

57. Abdel-Aty, M.A., C.L. Chen, and J.R. Schott. An assessment of the effect of driver age on traffic accident involvement using log-linear models. In *Accident Analysis & Prevention*. Vol.30, 1998, pp. 851-861.

58. Miaou, S.-P. Relationship between truck accidents and highway geometric design : a poisson regression approach. In *Transportation Research Record: Journal of the Transportation Research Board*. 1992, pp. 10-18.

59. Mitra, S. and S. Washington. On the nature of over-dispersion in motor vehicle crash prediction models. In *Accident Analysis & Prevention*. Vol.39, 2007, pp. 459-468.

60. Shankar, V., F. Mannering, and W. Barfield. Effect of roadway geometrics and environmental factors on rural freeway accident frequencies. In *Accident Analysis & Prevention*. Vol.27, 1995, pp. 371-389.

61. Vogt, A. and C. Turner-Fairbank Highway Research, *Crash models for rural intersections four-lane by two-lane stop-controlled and two-lane by two-lane signalized*. U.S. Dept. of Transportation, Federal Highway Administration, Research, Development, and Technology, Turner-Fairbank Highway Research Center ; Available to the public through the National Technical Information Service, McLean, VA; Springfield, Va., 1999.

62. Venables, W.N., B.D. Ripley, and W.N. Venables, *Modern applied statistics with S*. Springer, New York, 2002.